Letter

On the origin of new genes in *Drosophila*

Qi Zhou,[1,2,4] Guojie Zhang,[1,2,3,4] Yue Zhang,[1,4] Shiyu Xu,[1] Ruoping Zhao,[1] Zubing Zhan,[1,2] Xin Li,[1,2] Yun Ding,[1,2] Shuang Yang,[1,3] and Wen Wang[1,5]

[1]*CAS-Max Planck Junior Research Group, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China;* [2]*Graduate School of Chinese Academy Sciences, Beijing 100086, China;* [3]*Beijing Genomics Institute-Shenzhen, Shenzhen 518083, China*

Several mechanisms have been proposed to account for the origination of new genes. Despite extensive case studies, the general principles governing this fundamental process are still unclear at the whole-genome level. Here, we unveil genome-wide patterns for the mutational mechanisms leading to new genes and their subsequent lineage-specific evolution at different time nodes in the *Drosophila melanogaster* species subgroup. We find that (1) tandem gene duplication has generated ~80% of the nascent duplicates that are limited to single species (*D. melanogaster* or *Drosophila yakuba*); (2) the most abundant new genes shared by multiple species (44.1%) are dispersed duplicates, and are more likely to be retained and be functional; (3) de novo gene origination from noncoding sequences plays an unexpectedly important role during the origin of new genes, and is responsible for 11.9% of the new genes; (4) retroposition is also an important mechanism, and had generated ~10% of the new genes; (5) ~30% of the new genes in the *D. melanogaster* species complex recruited various genomic sequences and formed chimeric gene structures, suggesting structure innovation as an important way to help fixation of new genes; and (6) the rate of the origin of new functional genes is estimated to be five to 11 genes per million years in the *D. melanogaster* subgroup. Finally, we survey gene frequencies among 19 globally derived strains for *D. melanogaster*-specific new genes and reveal that 44.4% of them show copy number polymorphisms within a population. In conclusion, we provide a panoramic picture for the origin of new genes in *Drosophila* species.

[Supplemental material is available online at www.genome.org.]

The origin of new genes is of inherent interest to evolutionary biologists. But it was not until recently that the molecular details of the origin of new genes have been rigorously examined (Long et al. 2003). Multiple mechanisms including gene duplication, retroposition, horizontal gene transfer, and de novo origination from noncoding sequences have been proposed for the birth of a new gene (Long et al. 2003). Nonetheless, a genome-wide comparison of all of these mechanisms remains desirable due to the light it could shed on their relative contributions to the origin of new genes.

Among the individual mechanisms, gene duplication has been considered as the singular most important one in creating new genes (Ohno 1970, 1973; Kimura and Ota 1974; Hughes 1994). Its contribution to the evolution of new functional genes has been widely demonstrated in various organisms (Zhang et al. 2002; Katju and Lynch 2003; Arguello et al. 2006; Yang et al. 2008), and the gene duplication rate has been extensively studied in different species (Lynch and Conery 2000; Gu et al. 2002; Gao and Innan 2004; Hahn et al. 2007). Duplicate genes usually can be classified into tandem and dispersed duplicates; however, the relative mechanistic contributions to their origins is largely unknown. In addition, the classic gene duplication model predicted that the most common fate for a newly duplicated copy is to become nonfunctional (Haldane 1933; Fisher 1935; Ohno 1970; Nei and Roychoudhury 1973). However, subsequent discoveries of numerous functional duplicated genes have inspired theoretical discussions on the evolution of redundancy (Clark 1994; Nowak et al. 1997), subfunctionalization (Force et al. 1999; Lynch and Force 2000), and neofunctionalization (Walsh 1995, 2003). All of these theoretical treatments overwhelmingly assume that nascent duplicates are functionally and structurally redundant to their parental genes. Empirical evidence indicates, however, that pervasive structural heterogeneity can emerge between even young duplicates through the recruitment of new regulatory/coding sequences (Brosius and Gould 1992; Katju and Lynch 2003, 2006; Yang et al. 2008). Such chimeric structures, if non-deleterious, are expected to immediately confer novel functions that the parental genes do not have and thus may also well explain the existence of functional duplicates (Patthy 1999). Besides the reported cases of chimeric new genes found within different species (Patthy 1999; Long et al. 2003; Arguello et al. 2006; Cordaux et al. 2006; Xue et al. 2007), their generality, the mutational mechanisms, and selective forces responsible for their formation remains an intriguing question to be addressed.

Unlike regular gene duplication occurring at the DNA level, retroposition events generate intronless new copies by reverse transcription of a parental gene's mRNA (Brosius 1991). Numerous retrocopies have been identified in mammals (Marques et al. 2005; Vinckenbosch et al. 2006), plants (Zhang et al. 2005; Wang et al. 2006), and *Drosophila* (Betran and Long 2003; Bai et al. 2007). But the proportion of new genes derived from retroposition is still unknown except in *Caenorhabditis elegans* (Katju and Lynch 2006). Horizontal gene transfer has commonly occurred among prokaryotic species, and thus led to the introduction of new genes (Koonin et al. 2001). It is also possible that prokaryotic/eukaryotic parasites can occasionally transfer their genetic materials to their eukaryotic hosts (Bergthorsson et al. 2003; Hotopp et al. 2007). For example, Hotopp et al. (2007) recently reported

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-mail wwang@mail.kiz.ac.cn; fax 86-871-5193137.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.076588.108.

18:1446–1455 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

that the endosymbiont *Wolbachia* has transferred nearly its entire genome to its host, *Drosophila ananassae*, but the functionality of these *Wolbachia* genes in the host is uncertain. Each of the above three mechanisms are based on Ohno's notion that "each new gene must have arisen from an already existing gene," (Ohno 1970) and de novo gene origination from noncoding sequences should be rare, if not absent. In an influential essay, François Jacob (1977) also stated: "The probability that a functional protein would appear de novo by random association of amino acids is practically zero." However, a pioneer study reported five de novo genes in *Drosophila melanogaster* (Levine et al. 2006). This discovery urged a systematic evaluation regarding the extent that de novo formation has contributed to the evolution of protein-coding genes.

The most efficient approach to answer all of the above questions is to systematically study young genes that are in their early evolutionary stages (Long et al. 2003). Overall, we seek to tackle five central questions regarding both the origination and evolution of new genes: What is the general picture for the relative contribution of each mutational mechanism in creating new genes? How and how often would a new gene form a novel chimeric gene structure? What are the relative contributions of tandem and dispersed duplication to the creation of new duplicates? Is de novo origination an infrequent phenomenon during the origin of new genes? And, finally, beyond the apparent gene duplication rate, what is the rate of origin of functional new genes?

The *D. melanogaster* species subgroup is an excellent model for answering these questions due to their short divergence times (<12.8 million years [Myr]) (Tamura et al. 2004) and the availability of abundant genetic data and techniques. Our previous studies characterizing the complete origination process of several new genes within this clade have demonstrated that such a system allows scrutinization of the original process and the structural features after the birth of new genes (Wang et al. 2000, 2002, 2004; Long et al. 2003; Arguello et al. 2006; Yang et al. 2008). Now the availability of genome sequences for five species within this clade (Clark et al. 2007) offers an unprecedented opportunity to address the above questions at the whole-genome level and thereby reveal the general principles governing the origin of new genes.

Results

Identification and classification of new genes

To insure high-quality results, we first focused on *D. melanogaster* and *D. yakuba*, which have deep-coverage (more than 8×) genome sequences of high quality and a relatively short divergence time. We inferred the paralog number in *D. melanogaster* and *D. yakuba* after aligning 12,017 *D. melanogaster* cDNAs to their genome sequences. Taking both species as reciprocal outgoups, focal genes with extra copies in certain species were taken as candidates for new genes. New gene candidates in *D. melanogaster* were further divided into those specific to *D. melanogaster* and the *D. melanogaster* species complex by searching the genomes of *Drosophila simulans* and *Drosophila sechellia*. The latter group of new genes represents those originated in the common ancestors of *D. melanogaster*, *D. simulans*, and *D. sechellia*. Most likely, they should also exist in *Drosophila mauritiana*, the single species in the *D. melanogaster* species complex whose genome sequence is not yet available. After the initial screen, we acquired 169 new gene candidates in *D. melanogaster*, 253 in *D. yakuba*, and 191 shared by the *D. melanogaster* species complex.

We further performed extensive manual checks on the UCSC Genome Browser (http://genome.ucsc.edu/; Karolchik et al. 2003) and excluded spurious results caused by exon duplication/origination, repetitive elements, or sequencing gaps in the outgroup. Given the recently reported gene loss phenomenon in the 12 *Drosophila* genomes (Clark et al. 2007; Hahn et al. 2007), we also excluded candidates that resulted from loss events in the outgroup by comparison with the genomes of *Dropsophila erecta* and *Drosophila ananassae* as additional outgroups (see Methods). We retained only those well-annotated new genes with empirical support (i.e., expressed sequence tags or cDNAs) and intact open reading frames (ORFs) in *D. melanogaster*, each of which has an annotation ID and is likely to be functional. For *D. yakuba*, we retained new gene copies with intact ORFs according to the ab initio prediction. With this stringent criteria, we finally acquired three new gene datasets: 72 new genes specific to *D. melanogaster*, 177 genes specific to *D. yakuba*, and 59 new genes shared by the *D. melanogaster* species complex (Fig. 1; Supplemental Tables S1–S3). We have compared our results in detail with other work characterizing orthology/paralogy in *Drosophila* species (Supplemental Data S1; Bai et al. 2007; Bhutkar et al. 2007; Clark et al. 2007; Heger and Ponting 2007). The new gene numbers are very close to two other recent estimates of lineage-specific gene duplications in *D. melanogaster* (77 duplications) or *D. yakuba* (200 duplications) (Hahn et al. 2007; Heger and Ponting 2007). These identified genes include the well-characterized new genes *sphinx* (Wang et al. 2002), *Sdic* (Nurminsky et al. 1998), *Ntf-2r* (Betran and Long 2003), *Acp29AB* (Clark et al. 1995; Aguade 1999), and five recently identified de novo genes in *D. melanogaster* (Levine et al. 2006). We didn't include recently reported new genes *Hun* (Arguello et al. 2006), *monkey king protein* (Wang et al. 2004), *hydra* (Chen et al. 2007), and several de novo new genes specific to *D. yakuba*/*D. erecta* (Begun et al. 2007) because they originated in lineages not focused on by this study, or their identifications depend on annotations other than those of *D. melanogaster*. Inclusion/exclusion of these experimentally verified new genes indicates the robustness of our in silico screen. We subsequently calculated the ratio of nonsynonymous substitution rate over

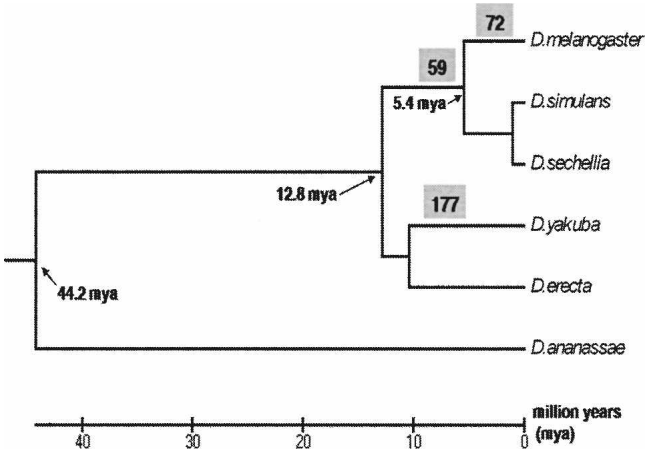

Figure 1. Phylogenetic distribution of new genes. We designated numbers of new genes restricted to different lineages. Phylogenetic tree of investigated *Drosophila* species and their divergence times in each node are indicated (Tamura et al. 2004).

synonymous substitution rate (K_a/K_s) as an indicator of functionality between all the retained new genes and their parental genes. The average between-paralog K_a/K_s ratios are less than 0.5 for *D. melanogaster*- and *D. yakuba*-specific new genes, and the average between-ortholog K_a/K_s ratio of the *D. melanogaster* species complex-specific new genes is less than 1. In particular, 47.1% of *D. melanogaster*-specific, 55.7% of *D. yakuba*-specific, and 71.4% of the *D. melanogaster* species complex-specific new genes show a K_a/K_s ratio significantly less than 1 (Fisher's test, $P < 0.05$, Supplemental Tables S1–S4). These data likely indicate that a majority of the new genes are under functional constraints.

A majority of the constrained functional new genes are dispersed duplicates

We can estimate the ages of the identified new genes based on their phylogenetic distributions (Fig. 1): The *D. melanogaster*-specific new genes are younger than 5.4 Myr, the *D. yakuba*-specific new genes are younger than 12.8 Myr, and the *D. melanogaster* species complex-specific new genes have originated within 5.4–12.8 Myr (Tamura et al. 2004). To quantify the contribution of each mechanism creating new genes, we classified all the new genes into four types: tandem duplicates, dispersed duplicates, retrogenes, and de novo originated genes (Table 1). Comparison among these different types of new genes should reveal their relative contributions not only at a genome-wide scale but also in a dynamic chronological order (Fig. 1).

We found that the proportion of gene duplication (both tandem and dispersed) consistently outnumbers other mechanisms in generating new genes in all the three datasets (Fig. 2). Specifically, tandem gene duplication predominantly accounts for the emergence of nascent gene copies specific to *D. melanogaster* (59/72, 81.9%) and *D. yakuba* (138/177, 77.9%). Nevertheless, this proportion decreases to only 33.9% (20/59) for those constrained new genes shared by the whole *D. melanogaster* species complex. Accordingly, the proportion of new dispersed duplicates rises from only 8.3% (6/72) to 44.1% (26/59) (Fig. 2). Since more new genes shared by the *D. melanogaster* species complex show selective constraints indicated by K_a/K_s ratios (see above), they are more likely to be functional than those younger ones specific only to *D. melanogaster* or *D. yakuba*. Thus, we can conclude that new constrained functional genes are mainly dispersed duplicates in *Drosophila* species. Regarding the remaining mechanisms, retropostion generated ~10.2% (6/59) of the new genes, and de novo genes, which have originated from noncoding sequences, unexpectedly comprise 11.9% (7/59) of the constrained new genes shared by the *D. melanogaster* species complex.

Table 1. New gene number categorized by different mechanisms

	Gene duplication				
	Tandem	Dispersed	Retroposition	de novo	Total
D. mel	59	6	5	2	72
D. mel–D. sch–D. sim	20	26	6	7	59
D. yak	138	39	NA	NA	177

(*D. mel*) *D. melanogaster*, (*D. sch*) *D. sechellia*, (*D. sim*) *D. simulans*. These three species together stand for the *D. melanogaster* species complex. Due to the lack of reliable annotation information of gene structure in *D. yakuba* (*D. yak*), we didn't identify new genes originated from retroposition and de novo origination.

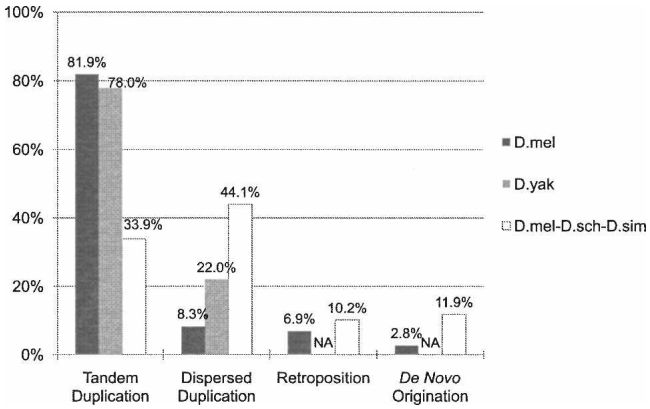

Figure 2. Quantification of contributions to new gene origination from different mechanisms. (*D. mel*) *D. melanogaster*, (*D. sch*) *D. sechellia*, (*D. sim*) *D. simulans*. These three species together stand for the *D. melanogaster* species complex. We didn't identify new genes originating from retroposition or de novo processes in the *D. yakuba* (*D. yak*) lineage due to the lack of reliable annotations in this species.

Tandem duplicates have younger ages but lower survivorships

We can also separately date each new gene using a molecular clock-based method. Based on the synonymous substitution rate (K_s) between ancestral and new genes (Table 2), and assuming the neutral substitution rate is ~5.8×10^{-9} per generation for the *D. melanogaster* genome (Haag-Liautard et al. 2007), we estimated ages for *D. melanogaster*-specific, *D. yakuba*-specific, and the *D. melanogaster* species complex-specific new genes. The average ages for the three groups of genes are 0.56 ± 0.12 Myr, 1.27 ± 0.22 Myr, and 7.01 ± 1.35 Myr, respectively, indicating that most of the *D. melanogaster*- and *D. yakuba*-specific copies are much younger than the speciation time, but the ages of *D. melanogaster* species complex-specific genes are more or less consistent with the age of this lineage. In particular, the new tandem duplicates we identified have an average K_s that is lower than dispersed duplicates (Table 2). The average ages of *D. melanogaster*-specific, *D. yakuba*-specific, and the *D. melanogaster* species complex-specific new tandem duplicates are estimated to be only 0.33–0.59 Myr, 0.79–1.22 Myr, and 3.88–7.10 Myr, respectively, in contrast to 0.62–0.95 Myr, 1.59–2.89 Myr, and 7.50–12.38 Myr for dispersed duplicates in these lineages. This result suggests that tandem duplicates are on average younger than dispersed ones, and that many new tandem duplicates identified in *D. yakuba* may have actually originated after the split of the ancestral lineage that led to *D. yakuba*, *Drosophila santomea*, and *Drosophila teissirri* (6.8 Myr) (Tamura et al. 2004). We cannot completely exclude the possibility that gene conversion may homogenize some parental and new gene pairs (Gao and Innan 2004), but a recent study strongly suggests it might play a minor role in genome-wide scale in *Drosophila* species (Hahn et al. 2007).

We further investigated whether *D. melanogaster*-specific new genes have been fixed in the population. We conducted PCR experiments to test copy number variations (CNVs) according to the presence/absence of bands within different lines for 45 out of 72 (62.5%) *D. melanogaster*-specific genes. Most of the PCR products are designed to span the boundary sites of two duplicated segments where parental and new genes reside, and we used multiple primer sets (three to four sets) to minimize the effect of mispairing for most of the PCR assays (see below, and Methods). In total, we inspected 10 South American iso-female strains as a

Table 2. Synonymous substitution rate (K_s) calculated between paralogs in different categories of new genes

K_s	Tandem duplication	Dispersed duplication	Retroposition
D. mel	0.053 ± 0.015	0.091 ± 0.019	0.195 ± 0.054
D. mel–D. sch–D. sim	0.637 ± 0.187	1.153 ± 0.283	0.186 ± 0.032
D. yak	0.117 ± 0.025	0.260 ± 0.076	NA

(*D. mel*) *D. melanogaster*, (*D. sch*) *D. sechellia*, (*D. sim*) *D. simulans*. These three species together stand for the *D. melanogaster* species complex. (*D. yak*) *D. yakuba*. We calculated substitution rates between paralogs using the method described by Yang and Nielsen (2000). The standard error is shown. The detail ratio for each gene is also provided in Supplemental Tables S1–S3.

local population sample, together with nine other strains from different continents as a world-wide sample. 44.4% (20 out of 45) of the nascent genes show CNVs among the 19 *D. melanogaster* strains (Table 3, Supplemental Table S5). A majority (75%) of these investigated nascent genes are tandem duplicates, and 58.8% of them have been fixed (Table 3). This is consistent with the proportion of tandem duplicates fixed in the *D. melanogaster* complex, and suggests tandem new genes have a lower survivorship compared with other types of new genes.

Most new tandem duplicates reside in segmental duplications

All identified tandem duplicates except for one *D. melanogaster*-specific copy (*CG18816–CG30160*) are direct repeats. This differs from the pattern of *C. elegans*, where most of the young duplicated genes (45%) are tandem duplicates with inverted directions (Katju and Lynch 2003). Further characterizations indicated that 61.0% (36/59) of *D. melanogaster*-specific, 77.5% (107/138) of *D. yakuba*-specific, and 65% (13/20) of the *D. melanogaster* species complex-specific tandem new genes reside in tandem segmental duplications (Supplemental Tables S3, S6). All of the *D. melanogaster*-specific, 84% of the *D. yakuba*-specific, but only 20% of the *D. melanogaster* species complex-specific tandem segmental duplications are in immediate conjunction with each other with no or few intervening base pairs (<50 bp). A total of 22 such tandem segmental duplications whose lengths range from 580 bp to >33 kb generated 36 *D. melanogaster*-specific tandem duplicate genes (Supplemental Table S6). Among them, eight transcribed new duplicates emerged with chimeric gene structures at the boundaries of the duplicated segments, through shuffling exons/introns of two genes located at the 5′ and 3′ end of the segments (Supplemental Fig. S1).

Additionally, we found that 30.5% (18/59) of *D. melanogaster*-specific, 11.8% (16/138) of *D. yakuba*-specific, and 25% (5/20) of *D. melanogaster* species complex-specific tandem duplicates or their ancestral genes have at least one repetitive element at the duplication breakpoints. Interestingly, this proportion for dispersed duplicates rises to 33.3% (2/6), 28.2% (11/39), and 46.2% (12/26) for each lineage, respectively, showing a higher correlation between dispersed new duplicates and repetitive elements, which conforms to our previous report (Yang et al. 2008).

Remarkable role of de novo new gene origination

We observed an unexpectedly high proportion of de novo new genes (11.9%, Fig. 2) in the *D. melanogaster* species complex. Besides the previously reported cases (Levine et al. 2006), we identified a total of nine de novo genes comprised of seven in the *D. melanogaster* species complex and two specific to *D. melanogaster*. These genes are likely functional protein-coding genes, based on two pieces of evidence: First, seven of them are shared by the *D. melanogaster* species complex and have been retained in the genome for more than 5.4 Myr. Second, all of them have transcriptional evidence provided by ESTs or full-length cDNAs. Because the de novo gene *CG33666* was annotated without a start codon by FlyBase (www.flybase.org), we performed a 5′ RACE (rapid amplification of cDNA ends; see Methods) experiment and re-annotated its intact ORF (Supplemental Table S1, Note). Thus, all the de novo genes are predicted to have intact ORFs. Notably, the lengths of all the newly identified de novo genes' protein products are predicted to be longer than 100 amino acids. It is also noteworthy that both EST (http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene) and microarray (www.flyatlas.org) data have confirmed that the gene *CG33235* has evolved a testis-specific expression pattern, which parallels previously reported de novo genes (Levine et al. 2006; Chen et al. 2007).

We searched the entire GenBank database, and none of these de novo genes can retrieve any significant BLAST hits from organisms other than insects. All but one (*CG31909*) of the de novo genes can be mapped to their homologous noncoding counterparts in *D. yakuba* (see Methods). Gene *CG31909* cannot find any homologous sequences in all the 12 *Drosophila* genomes. Interestingly, it is homologous to a region of noncoding sequence in the honeybee, which raises the possibility that this gene may have originated though lateral gene transfer. The other eight de novo genes appear to have evolved from ancestral noncoding sequences.

Based on the characterization of their homologous sequences, we found heterogeneity during the origin of these new genes (see Supplemental Data S2 for the alignments). More specifically, the flanking gene orders of four genes (*CG33666*, *CG32690*, *CG33235*, and *CG2042*) are conserved between *D. melanogaster* and *D. yakuba* or *D. simulans*, suggesting they should have derived from orthologous ancestral sequences. Interestingly, *CG33235* originated from lineage-specific expansion of simple tandem repeats, forming a gene putatively encoding a protein with a length of 1584 amino acids. As mentioned above, it probably has evolved a testis-specific expression pattern. It might be the first identified young gene with >75% of its coding region comprised of repetitive sequences. *CG2042* is *D. melanogaster*-specific and emerged through recruiting a long interspersed element BS2 as its first exon (see Supplemental Table S1 for annotation updates to this gene). The formation process of four de novo genes (*CG40384*, *CG15323*, *CG32582*, and *CG32824*) appears to have involved lin-

Table 3. Fixation pattern in different kinds of new genes specific to *D. melanogaster*

	Total tested	Fixed genes[a]	Fixed proportion
Tandem duplication	34	20	58.8%
Dispersed duplication	3	0	0
Retroposition	6	3	50%
De novo	2	2	100%
Total	45	25	55.5%

[a]A gene is considered "fixed" if all the investigated strains can detect positive results with PCR assays.

eage-specific genomic rearrangements. Their flanking gene order is not conserved between species, but they can be mapped to another homologous genomic region in *D. yakuba*. For example, the gene *CG32824* on chromosome X in *D. melanogaster* can be mapped to a region on chromosome 2 in *D. yakuba*. A DNA transposable element PROTO_B is adjacent to this region in *D. yakuba*, and this gene's second exon in *D. melanogaster*. This suggests that ancient genomic rearrangements such as DNA transposition or chromosomal translocation events might lead to the emergence of these genes. In summary, these data shed additional light on how noncoding sequences can be recruited in various ways to give rise to an entirely new gene.

Substantial proportion of new genes formed chimeric structures

The extensive annotations available for both parental and new genes in *D. melanogaster* gave us the unique opportunity to trace their gene structure changes. We focused on the comparison of the protein-coding (CDS) regions between ancestral and new genes in *D. melanogaster* and the *D. melanogaster* species complex due to the possible inadequate annotation in UTR regions (see Methods). Interestingly, a high proportion (29.2% and 32.2%, respectively) of new genes from these two levels of taxa formed chimeric gene structures by recruiting flanking sequences into their protein-coding regions (Table 4; Supplemental Table S7). The recruited regions come from a variety of genomic sources (Supplemental Fig. S2), including coding regions of other genes (i.e., exon shuffling), intronic or intergenic sequences (i.e., exonization), and repetitive elements (long interspersed repeats or simple tandem repeats). The unique amino acids in the new genes are very likely to contribute to the process of neofunctionalization. We found that the proportion of new genes completely duplicating their ancestors' CDS decreases with an increase of the new genes' ages (Fig. 3). In fact, there are only ~16.3% of such new genes in the *D. melanogaster* species complex, in contrast to 40.9% in *D. melanogaster* (Fig. 3). This indicates that functional redundancy resulting from complete duplication may have less of a chance to be fixed during new gene origination.

Chromosomal distribution of new genes

We found a significant (Fisher's exact test, $P < 0.01$) excess of new genes located on the X chromosome rather than chromosome 2 or 3 (Fig. 4) in both lineages of *D. melanogaster* and the *D. melanogaster* species complex. De novo genes are overrepresented (66.6%, one gene *CG40384* doesn't have chromosome information) on the X chromosome, consistent with the previously reported pattern (Levine et al. 2006). Also, retroposed new genes show a "gene traffic" pattern on the X chromosome (Betran et al. 2002; Emerson et al. 2004): Four out of five (80%) new retrogenes specific to *D. melanogaster* and five out of six (83.3%) retrogenes specific to the *D. melanogaster* species complex derived from an X-to-autosome or autosome-to-X retroposition event. In contrast, the new genes specific to *D. yakuba* are overrepresented on the third autosome but are significantly underrepresented on the X chromosome (Fig. 4; Fisher's exact test, $P < 0.01$). The chromosomal distribution patterns of new genes in these species are in agreement with a recent analysis on lineage-specific duplications in *Drosophila* species (Heger and Ponting 2007), which found that lineage-specific duplications have almost always been enriched within the X chromosome in *D. melanogaster*/*D. simulans*. Further studies are required to answer whether such dramatically different patterns of chromosomal distribution of new genes in *D. melanogaster* and *D. yakuba* is due to a lack of reliable gene annotation in *D. yakuba* or different genomic evolution history of these species.

Table 4. Proportion of chimeric new genes generated by different mechanisms

	D. mel	*D. mel–D. sch–D. sim*
Tandem duplication	17/72 (23.6%)	8/59 (13.6%)
Dispersed duplication	3/72 (4.2%)	8/59 (13.6%)
Retroposition	1/72 (1.4%)	3/59 (5%)
Total	29.2%	32.2%

(*D. mel*) *D. melanogaster*, (*D. sch*) *D. sechellia*, (*D. sim*) *D. simulans*. These three species together stand for the *D. melanogaster* species complex.

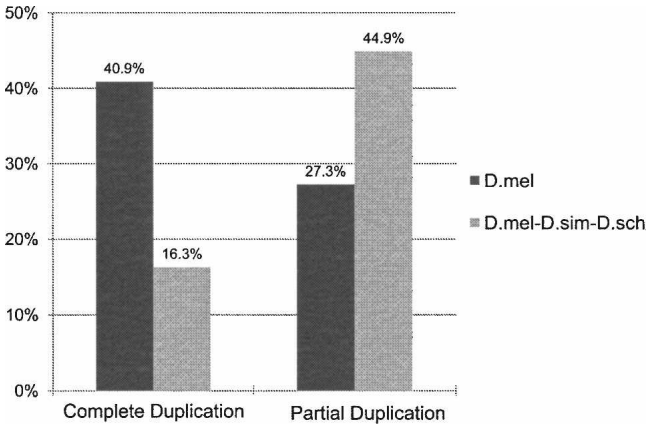

Figure 3. Decrease of proportion of complete duplicated new genes with time. (*D. mel*) *D. melanogaster*, (*D. sch*) *D. sechellia*, (*D. sim*) *D. simulans*. These three species together stand for the *D. melanogaster* species complex. We compared structures and lengths of parental and new genes in *D. melanogaster* and the *D. melanogaster* species complex. New genes completely duplicating their ancestors' coding regions seem more vulnerable to subsequent loss or structure changes during evolution.

Discussion

A general picture for the origin of new genes in *Drosophila*

The large disparities in gene numbers among organisms imply the fundamental process of new gene origination during evolution (Tatusov et al. 1997; Rubin et al. 2000). To date, various mechanisms have been demonstrated to be capable of generating new genes (Long et al. 2003). In this work, we selected the *D. melanogaster* species subgroup with its high-quality genome data and its short divergence times to provide an integrative insight into these mechanisms. We compared the different mechanisms of the origin of new genes both within genomes and over evolutionary time. We uncovered that gene duplication (both tandem and dispersed) generated the most abundant new genes (Table 1). However, different types of gene duplication may be generated by different mechanisms and have different contributions to the emergence of new genes.

Regarding other mechanisms, we detected eight putatively functional new retrogenes with transcriptional evidence, and shared by the *D. melanogaster* species complex (Table 1). These retrogenes account for ~10% of the new genes, which translates

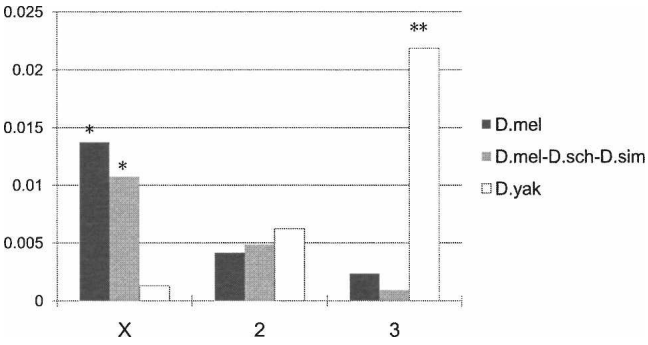

Figure 4. Chromosomal distribution of new genes in three datasets. New gene numbers were divided by total gene numbers on the specific chromosome as a normalization. We marked significant (*$P < 0.01$, Fisher's exact test) and highly significant (**$P < 0.001$) overrepresentation with of new genes on certain chromosomes. We didn't take chromosome 4 into account given the extremely low gene number on this chromosome.

to an average origination rate of 0.6–1.5 new retrogenes per million years. It supports previous reports that retroposition is an important mechanism for creating new genes (Brosius 1991; Marques et al. 2005; Bai et al. 2007). However, given the compact genomes of *Drosophila* species (Petrov et al. 1996; Clark et al. 2007), their limited sources of retroposons may render the role of retropostion in the origin of new genes not as significant as that in human and rice (>36% of human and ~38% of rice duplicates are generated from retroposition) (Marques et al. 2005; Wang et al. 2006; Pan and Zhang 2007). Most strikingly, we found that new genes originating de novo from noncoding sequences have contributed as much as 11.9% of the putatively functional new genes. This further suggests its important yet underappreciated role during the origin of new genes. By summarizing the contribution of each mechanism, we are able to for the first time provide an integrative picture about all new genes on the genomic level.

We are also able to provide a rough estimate for the origination rate of functional new genes, which is an important parameter for understanding the tempo of new gene origination. Much attention has previously been paid to gene duplication rate (Lynch and Conery 2000; Gu et al. 2002; Hahn et al. 2007; Pan and Zhang 2007), and it has been estimated to range from 0.0010 (Hahn et al. 2007) to 0.0023 (Lynch and Conery 2000) per gene per million years in *Drosophila* species. Compared with the gene duplication rate, the rate of functional new gene origination makes more sense because only such new genes authentically contribute to the evolution of organismal diversities. These new genes fixed in the *D. melanogaster* species complex provide a valuable dataset to estimate this parameter. The genes that are more likely to be functional are those that have been kept intact in the genomes for a long time (>5.4 Myr). Therefore, we estimate that the rate of new functional gene origination is five to 11 per million years in the *Drosophila* genome (60 new genes divided by 5.4 and 12.8 Myr), which translates to 0.000391 to 0.000925 per gene per millions years, assuming 12,017 genes in *D. melanogaster* genome (the number used in this study). As expected, this rate is much lower than the estimated gene duplication rate. In fact, if we use all *D. melanogaster*-specific new genes, including both fixed and unfixed ones, the estimated new gene origination rate (0.0011 per gene per million years, 72 genes/5.4 Myr, assuming 12,017 genes) is very close to the recent estimate of gene duplication rate in 12 *Drosophila* species (0.0010) (Hahn et al. 2007). However, our population data (Table 3) and a recent study on paralogs of 12 *Drosophila* species (Heger and Ponting 2007) suggest that substantial gene duplicates used in previous studies perhaps are polymorphic within populations, and thus many of them could disappear with time. In this sense, the rate of new gene origination is a more realistic indicator than gene duplication rate to assess speed of generation of genetic novelties in evolutionary biology.

A dynamic view on new duplicates

From Table 1, we can see that tandem gene duplication comprises most of the nascent duplicates (~80%, or >11 copies/Myr) in both *D. melanogaster* and *D. yakuba*. However, only 58.8% of them are fixed throughout the populations of *D. melanogaster* (Table 3), and only 33.9% of the constrained new genes shared in the *D. melanogaster* species complex are tandem duplicates (Table 1; Fig. 2). This is probably due to a deletion event, or it can be explained by a rapid birth but infrequent fixation model (Heger and Ponting 2007). In contrast, although dispersed new duplicates seem to have a lower birth rate (Table 1, one copy per million years for *D. melanogaster*, three copies per million years for *D. yakuba*), they comprise the highest proportion (44.1%) among all the constrained new genes shared in the *D. melanogaster* species complex. These results suggest that dispersed duplicates might have a higher survivorship than nascent tandem ones.

However, while many tandem duplicates may subsequently get lost, some may actually become dispersed. This expectation is based on the observation that there is a significantly (Table 1; Fisher's exact test, $P = 0.000129$) higher number of dispersed duplicates in the *D. melanogaster* species complex compared with *D. melanogaster,* and this cannot simply be explained from higher survivorship of original dispersed duplicates. Even if we assume all of the dispersed new duplicates are able to survive after their emergence, we would expect only 8%–21% of dispersed new genes to emerge in the 5.4–12.8 Myr shared by the *D. melanogaster* species complex (one copy per million years for dispersed gene duplication in *D. melanogaster*, see above). Such a discrepancy from the observed 44.1% thus strongly suggests that some of the tandem duplicates might have become dispersed through subsequent mutation events (Clark et al. 2007; Heger and Ponting 2007). This scenario is consistent with our result showing that most nascent (all *D. melanogaster*-specific and 84% of *D. yakuba*-specific) tandem segmental duplications are in immediate conjunction with each other, while only 20% of them shared by the *D. melanogaster* species complex show this pattern. It is also supported by a recent study on gene families in human and mouse, which found that tandem duplicated members are of more recent origin than those interchromosomal ones (Friedman and Hughes 2003).

In addition to the evolutionary dynamics of tandem and dispersed duplicates, we can also infer that the underlying mechanisms of the origin of these two types of duplicates might be different. Several genomic processes have been proposed to account for duplication events: non-allelic homologous recombination (NAHR) (Stankiewicz and Lupski 2002; Bailey et al. 2003; Fiston-Lavier et al. 2007), transposon-mediated transposition (Lal et al. 2003; Jiang et al. 2004; Yang et al. 2008), and illegitimate recombination (IR) (Roth et al. 1985; Roth and Wilson 1988; Slack et al. 2006). Duplicates generated by NAHR or transposon-mediated transposition usually have certain homologous sequences or repetitive elements at the breakpoint sites of the duplicated blocks.

Alternatively, duplication could happen via IR (also called nonhomologous end joining [NHEJ]), which depends on little or no sequence homology. Our recent experimental efforts characterizing new genes in eight *D. melanogaster* subgroup species with comparative fluorescence in situ hybridization (cFISH) has found that a majority of new dispersed duplicate genes are associated with repetitive elements at the breakpoints (Yang et al. 2008). In this work, we interestingly found that dispersed and tandem duplicated new genes show different degrees of associations with repetitive elements at their breakpoints. It suggests that genomic processes depending on repetitive elements or homologous sequences such as NAHR or transposon-mediated DNA fragment transposition play an important role in creating both tandem and dispersed new genes. However, a lower association in tandem duplicates (30.5% of *D. melanogaster*-specific tandem new genes, 11.8% of *D. yakuba*-specific tandem new genes) compared with dispersed duplicates (33.3% and 26.8%, respectively) suggests that IR is more important generating tandem duplicated genes.

A new perspective on de novo origination

Ohno stated: "In a strict sense, nothing in evolution is generated de novo," (Ohno 1970). New genes with novel functions were believed to be derived mainly from preexisting genes (Ohno 1970; Long et al. 2003; Long 2007). A lack of comparison of de novo origination with other mechanisms of new gene origination has hampered a proper evaluation for this mechanism. In this study, we unexpectedly found 11.9% of the constrained new genes emerge from noncoding sequences, suggesting that de novo origination is not rare and its role during the origin of new genes is important (Table 1; Fig. 2). In fact, our current estimation for the contribution of this process is conservative for two reasons. First, we excluded several cases (e.g., *CG15930*) based on sequencing gaps located at their syntenic regions in the outgroup species. Second, our stringent criteria (see Methods) cannot find those de novo genes with alignable orthologous sequences that don't have expression in the outgroup species. Given these two reasons, the proportion of de novo new genes could even be higher than 11.9%. It would be very intriguing to investigate whether de novo origination also plays a prominent role in the origin of new genes in species other than *Drosophila*.

In addition, for the first time, we have analyzed the orthologous noncoding regions of these de novo genes and reconstructed their origination processes to show how a region of noncoding sequences can give rise to an entirely new gene. We revealed that multiple noncoding genomic sources, including intergenic sequences, simple tandem repeats, or long interspersed repetitive elements, have the potential to become a new gene. All of these de novo genes have intact ORFs, and some of them might have acquired novel functions, which is supported by the fact that some of these genes (*CG33235*, and five previously reported cases) have evolved a testis-specific expression pattern (Levine et al. 2006). It is also noteworthy that we found 13 *D. yakuba*-specific ESTs (data not shown) mapped to the de novo gene *CG32690*'s orthologous region in *D. yakuba*. However, there is no evidence showing that this region contains an intact ORF that is capable of encoding a protein longer than 50 amino acids. Consistent with our finding on a de novo gene in budding yeast (Cai et al. 2008), these data suggest that some de novo protein-coding genes may acquire the ability of transcription before being capable of encoding proteins (Casci 2008).

Chimeric structures significantly contribute to the evolution of new genes

Population genetic models have been developed to account for the fixation of new genes within a population (Ohta 1988; Clark 1994; Lynch and Force 2000; Lynch et al. 2001). Most of them assume that gene duplication generates a new gene copy that is functionally and structurally redundant at birth. Under this assumption, the redundant new copy is most likely subject to nonfunctionalization, with a low probability of undergoing subfunctionalization or neofunctionalization. Consistent with such predictions, our structure analysis results for new genes show that structurally redundant copies (complete duplicates) are vulnerable to losses (Fig. 3).

More importantly, we found that 29.2% of new genes specific to *D. melanogaster* formed chimeric structures. Given their extremely young ages, it suggests new genes are not necessarily structurally identical at birth. In particular, eight of them (Supplemental Table S6; Supplemental Fig. S1) arose from fusion events of partial coding regions/introns of two separate genes at the boundary position of two duplicated segments in less than 5.4 Myr in *D. melanogaster*. This vividly demonstrates the process of exon shuffling (Gilbert 1978). Since the involved duplicated segments are in immediate conjunction with each other, and five of such new genes exist without any synonymous substitutions (data not shown), it is likely that they originated as chimeric genes. We also found that 32.2% of the new genes specific to the *D. melanogaster* species complex formed chimeric structures. These constrained new genes are more likely to be functional, and thus it suggests that chimeric structure formation is an important solution to the preservation of new copies within the population. Although previous theoretical models predict a low fixation probability of a new copy, they consider only structurally redundant cases (Lynch and Force 2000; Lynch et al. 2001). If a new copy has formed a chimeric structure, it has the potential to immediately confer a novel function that the parental gene does not bear (Patthy 1999).

The generality of chimeric new gene formation can be reflected by the fact that it is not restricted to certain species or a certain type of new genes. Recent characterization of 37 young duplicates in *C. elegans* found that 38% of them formed chimeric gene structures (Katju and Lynch 2003, 2006). And our parallel work in rice has found that 42% of transcribed retrogenes have formed chimeric structures (Wang et al. 2006). It is noteworthy that chimeric structures formed at the untranslated regions of the new genes could also contribute to the acquisition of novel functions by introducing new regulatory regions (Begun 1997; Arguello et al. 2006). Here, however, we investigated only protein-coding regions, given their more reliable annotation. All these results indicate that forming chimeric gene structures is probably a common path for new genes to acquire novel functions and thus be preserved within the population.

Methods

Identification of new genes

To identify the species-specific candidate new genes, we downloaded the genome sequences of *D. melanogaster* (dm2, Apr.2004), *D. simulans* (droSim1, Apr.2005), and *D. yakuba* (droYak2, Nov.2005) from UCSC (http://hgdownload.cse.ucsc.edu/downloads.html), as well as 19,235 *D. melanogaster* cDNAs and their mapping information from FlyBase (http://www.flybase.org/). We removed redundancies resulting from alterna-

tive splicing and obtained 12,570 cDNAs, each representing the longest transcript from a unique genome locus. We used both sequence similarity and syntenic gene order to infer the bona fide orthologous relationship (Gao and Innan 2004). We first aligned unique cDNAs to each other using BLASTN (Altschul et al. 1990) with 10^{-5} as the *E*-value cutoff. To define paralogous genes, we grouped those cDNA pairs with an aligned length >30% and sequence identity >80% into 541 gene families. We picked up the longest cDNAs to represent each family. Together with 11,476 single-copy cDNAs, these 12,017 genes comprise our query sequences. We aligned the nonredundant query cDNAs against the genomes of *D. melanogaster*, *D. simulans*, and *D. yakuba* using BLASTN (Altschul et al. 1990) with 10^{-5} as the *E*-value cutoff. We retained those hits with aligned lengths >200 bp or 30% of the query cDNA length, taking 80% as the sequence identity cutoff for the aligned region. If a cDNA had two or more hits in the *D. melanogaster* genome, we then investigated the conservation of syntenic gene order defined by two flanking genes of the focal cDNA/gene between *D. melanogaster* and its outgroup species to discriminate the parental gene (conserved) and daughter genes (not conserved). In a few cases, the corresponding region to the flanking gene contained sequence gaps on one side in the outgroup species. In these cases, we used one available flanking gene to judge the synteny. For tandem gene duplications with the same flanking orthologous genes, we picked one at random as the new gene. It is reported that lineage-specific chromosomal rearrangements could also break the conservation of synteny (Bhutkar et al. 2007). However, such events are rare for most investigated lineages (fewer than five events except for *D. ananassae*; also see Supplemental Data S1) (Bhutkar et al. 2007) in this work and the microsynteny, i.e., the order of three adjacent genes, should be conserved (Heger and Ponting 2007). Based on the syntenic gene order and paralog number in three species, we acquired 169 candidate new genes in *D. melanogaster*, 253 genes in *D. yakuba*, and 191 genes shared only by *D. melanogaster* and *D. simulans*.

Manual check and classification of new genes

First, we checked the homolog numbers with each candidate new gene's genomic sequences in *D. sechellia* (droSec1, Oct.2005), *D. erecta* (droEre2, Feb.2006), and *D. ananassae* (droAna3, Feb.2006) with BLAT to exclude the possibility of gene loss in *D. simulans*, *D. yakuba*, or *D. erecta* (Kent 2002; Clark et al. 2007). We counted homologous hits with BLAT scores of higher than 100 in each species. If one of the outgroup species showed both the same homolog number and conservation of synteny compared with *D. melanogaster* for a certain candidate gene, it was removed from the dataset. Second, based on the annotation information, we excluded candidates that resulted from internal exon duplications or a sequencing gap in the outgroup. Only those new genes with annotation IDs (e.g., *CG10102*, Release 4.2 annotations from FlyBase) in *D. melanogaster* or an intact ORF predicted by Genscan (Burge and Karlin 1997) in *D. yakuba* were retained as authentic new genes. We have also included six noncoding genes termed with *CR*-ids as new genes, because they have shown differential expression patterns during early developmental stages or in specific tissues (Manak et al. 2006; Chintapalli et al. 2007; Supplemental Tables S1–S2, Notes). We initially defined de novo genes as those without significant sequence homology (aligned length longer than 200 bp or 30% of the query sequence length) detected by BLASTN in all the outgroup species. We have 22 such candidates. To exclude possible double gene loss events in the two outgroup species (*D. erecta* and *D. ananassae*), genomic sequences of these genes were further subjected to BLASTN using a discontinuous megablast algorism and then a TBLASTN search with their protein sequences against the whole GenBank database. We excluded a candidate, *CG15882*, that has homologous sequences detected by TBLASTN in *D. yakuba* and *D. erecta* and may be a de novo gene originated in the ancestor of the *D. melanogaster* subgroup species. None of remaining candidates retrieved significant BLAST hits (score higher than 50) in organisms other than species within the *D. melanogaster* species complex. Using the more sensitive BLASTZ program and based on the "Alignment Net" information on the UCSC Genome Browser, we further excluded 12 cases aligned to an annotated gene (mapped by BLASTZ) in *D. melanogaster* and reconstructed the origins of these genes (Schwartz et al. 2003). We also excluded a previously reported case, *CG32712* (Levine et al. 2006), which maintained its open reading frames in other insects' genomes except *D. yakuba* and *D. erecta*. We verified its expression with RT-PCR in *D. pseudoobscura* (data not shown). We performed a 5′RACE (FirstChoice RLM-RACE Kit, Ambion, Inc.) experiment using testis-derived RNA of *D. melanogaster* and reannotated the ORF structure of *CG33666* (Supplemental Table S2, Note).

We defined new genes without any introns, compared with their intron-containing parents, as retroposed new genes. Coding regions of these candidates were manually aligned back to the genome to assure that no hits in the introns of parental genes were detected. Finally, we classified newly duplicated genes as tandem when they are adjacent to each other or reside in duplicated segments without any intervening genes. Those with intervening genes, or those located on different chromosomes, are classified as dispersed gene duplications. For all the duplicated new genes, we extended the flanking regions of gene pairs until they cannot align with each other by BLAT. Thus, we determined the region of duplicated segment encompassing the focal genes in the genome. The segment with higher BLAT scores with the outgroup species is considered as the segment containing the parental genes. Thereby, we defined the parental–daughter relationship for tandem gene duplications. To avoid assembly error, we further performed MEGABLAST against the NCBI Trace Archive database (http://www.ncbi.nlm.nih.gov/Traces/) using sequence regions expanding the boundaries of the tandem duplicated segments as query and ensuring there is at least one sequence read covering the boundary sites. We surveyed repetitive elements beyond the breakpoints (50 bp around) of the duplicated segments or duplicated gene regions using RepeatMasker (Jurka 2000) and Tandem Repeat Finder (Benson 1999) information provided on the UCSC Genome Browser (Karolchik et al. 2003). The boundary sites were also surveyed for the orthologous regions in the outgroup species based on syntenic information.

Gene structure characterization and K_s calculation

We aligned the protein-coding regions of both parental and daughter genes with bl2seq. We considered it as a chimeric gene if the new gene recruited a stretch of unalignable sequence longer than 50 bp. The other new duplicates were assumed to generate from complete gene duplication if their length differences with parental genes are less than 50 bp. If the differences are longer than 50 bp and the new genes are shorter, they were characterized as partial duplications. We predicted gene regions and coding frames of parental and new genes in *D. yakuba* using GeneWise (Birney et al. 2004). Coding sequences were further aligned according to the protein sequences of *D. melanogaster* using Perl scripts. We calculated the synonymous (K_s) and nonsynonymous substitution rate (K_a) with KaKs_Calculator_1.2 (http://evolution.genomics.org.cn/software.htm) using methods described by Yang and Nielsen (2000).

Population study

We extracted genomic DNA of a single male fly from each of 19 iso-female *D. melanogaster* strains using the PUREGENE DNA Isolation Kit (Gentra). These strains included 10 Ecuadorian lines representing a South American local population, two North American lines, three African lines, one European line, and three Chinese lines. The Ecuadorian, North American, and African lines were obtained from M. Long's lab at the University of Chicago. The European line was obtained from Aike Guo's lab at the Institute of Neuroscience at Shanghai, and the Chinese lines were a gift from Wenxia Zhang of Peking University. Using these DNA samples, we used PCR to test if a new gene has been fixed in the population of *D. melanogaster* or not. We designed PCR primers either surrounding the boundary sites of duplicated segments or at sites that were capable of discriminating parental and daughter genes. To exclude possible false negatives due to sequence mismatch in some strains, three to four nonredundant primer sets were used for each gene. We performed PCR amplification with rTaq (Takara Bio Inc.) for these 19 strains together with a negative control without adding templates. Presence/absence data were recorded for each line based on the PCR results.

Acknowledgments

We thank all members of the CAS-Max Planck Junior Research Group in the Kunming Institute of Zoology, and J. Roman Arguello, Margarida Moreira, and Rebekah Rogers for their comments and help during preparing the manuscript. We also thank three anonymous reviewers for their constructive comments. This work was supported by a CAS-Max Planck Society Fellowship, an award (no. 30325016) of the National Science Foundation of China (NSFC), two NSFC key grants (nos. 30430400 and 30623007), and a 973 Program (no. 2007CB815703-5) to W.W., and a NSFC grant for junior researchers to S.Y. (no. 30500283).

References

Aguade, M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. *Genetics* **152:** 543–551.
Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.
Arguello, J.R., Chen, Y., Yang, S., Wang, W., and Long, M. 2006. Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. *PLoS Genet.* **2:** e77. doi: 10.1371/journal.pgen.0020077.
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. doi: 10.1186/gb-2007-8-1-r11.
Bailey, J.A., Liu, G., and Eichler, E.E. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73:** 823–834.
Begun, D.J. 1997. Origin and evolution of a new gene descended from alcohol dehydrogenase in *Drosophila*. *Genetics* **145:** 375–382.
Begun, D.J., Lindfors, H.A., Kern, A.D., and Jones, C.D. 2007. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. *Genetics* **176:** 1131–1137.
Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573–580.
Bergthorsson, U., Adams, K.L., Thomason, B., and Palmer, J.D. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* **424:** 197–201.
Betran, E. and Long, M. 2003. Dntf-2r, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. *Genetics* **164:** 977–988.
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.
Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res.* **14:** 988–995.
Brosius, J. 1991. Retroposons—Seeds of evolution. *Science* **251:** 753.
Brosius, J. and Gould, S.J. 1992. On "genomenclature": A comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc. Natl. Acad. Sci.* **89:** 10706–10710.
Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.
Cai, J., Zhao, R., Jiang, H., and Wang, W. 2008. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179:** 487–496.
Casci, T. 2008. Evolution: A gene is born. *Nat. Rev. Genet.* **9:** 415.
Chen, S.T., Cheng, H.C., Barbash, D.A., and Yang, H.P. 2007. Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in *Drosophila melanogaster*. *PLoS Genet.* **3:** e107. doi: 10.1371/journal.pgen.0030107.
Chintapalli, V.R., Wang, J., and Dow, J.A. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat. Genet.* **39:** 715–720.
Clark, A.G. 1994. Invasion and maintenance of a gene duplication. *Proc. Natl. Acad. Sci.* **91:** 2950–2954.
Clark, A.G., Aguade, M., Prout, T., Harshman, L.G., and Langley, C.H. 1995. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. *Genetics* **139:** 189–201.
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.
Cordaux, R., Udit, S., Batzer, M.A., and Feschotte, C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci.* **103:** 8101–8106.
Emerson, J.J., Kaessmann, H., Betran, E., and Long, M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* **303:** 537–540.
Fisher, R.A. 1935. The sheltering of lethals. *Am. Nat.* **69:** 446–455.
Fiston-Lavier, A.S., Anxolabehere, D., and Quesneville, H. 2007. A model of segmental duplication formation in *Drosophila melanogaster*. *Genome Res.* **17:** 1458–1470.
Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.
Friedman, R. and Hughes, A.L. 2003. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.* **20:** 154–161.
Gao, L. and Innan, H. 2004. Very low gene duplication rate in the yeast genome. *Science* **306:** 1367–1370.
Gilbert, W. 1978. Why genes in pieces? *Nature* **271:** 501.
Gu, Z., Cavalcanti, A., Chen, F.-C., Bouman, P., and Li, W.-H. 2002. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* **19:** 256–262.
Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D.L., Charlesworth, B., and Keightley, P.D. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445:** 82–85.
Hahn, M.W., Han, M.V., and Han, S.-G. 2007. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet.* **3:** e197. doi: 10.1371/journal.pgen.0030197.
Haldane, J.B.S. 1933. The part played by recurrent mutation in evolution. *Am. Nat.* **67:** 5–19.
Heger, A. and Ponting, C.P. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. *Genome Res.* **17:** 1837–1849.
Hotopp, J.C., Clark, M.E., Oliveira, D.C., Foster, J.M., Fischer, P., Torres, M.C., Giebel, J.D., Kumar, N., Ishmael, N., Wang, S., et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317:** 1753–1756.
Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.* **256:** 119–124.
Jacob, F. 1977. Evolution and tinkering. *Science* **196:** 1161–1166.
Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569–573.
Jurka, J. 2000. Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16:** 418–420.
Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31:** 51–54.
Katju, V. and Lynch, M. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome.

Genetics **165:** 1793–1803.
Katju, V. and Lynch, M. 2006. On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Mol. Biol. Evol.* **23:** 1056–1067.
Kent, W.J. 2002. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12:** 656–664.
Kimura, M. and Ota, T. 1974. Probability of gene fixation in an expanding finite population. *Proc. Natl. Acad. Sci.* **71:** 3377–3379.
Koonin, E.V., Makarova, K.S., and Aravind, L. 2001. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* **55:** 709–742.
Lal, S.K., Giroux, M.J., Brendel, V., Vallejos, C.E., and Hannah, L.C. 2003. The maize genome contains a helitron insertion. *Plant Cell* **15:** 381–391.
Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A., and Begun, D.J. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci.* **103:** 9935–9939.
Long, M. 2007. Journal club. *Nature* **449:** 511.
Long, M., Betran, E., Thornton, K., and Wang, W. 2003. The origin of new genes: Glimpses from the young and old. *Nat. Rev. Genet.* **4:** 865–875.
Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.
Lynch, M. and Force, A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459–473.
Lynch, M., O'Hely, M., Walsh, B., and Force, A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* **159:** 1789–1804.
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.
Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* **3:** e357. doi: 10.1371/journal.pbio.003035.
Nei, M. and Roychoudhury, A.K. 1973. Probability of fixation of nonfunctional genes at duplicate loci. *Am. Nat.* **107:** 362–372.
Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. 1997. Evolution of genetic redundancy. *Nature* **388:** 167–171.
Nurminsky, D.I., Nurminskaya, M.V., De Aguiar, D., and Hartl, D.L. 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* **396:** 572–575.
Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York.
Ohno, S. 1973. Ancient linkage groups and frozen accidents. *Nature* **244:** 259–262.
Ohta, T. 1988. Time for acquiring a new gene by duplication. *Proc. Natl. Acad. Sci.* **85:** 3509–3512.
Pan, D. and Zhang, L. 2007. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: A novel strategy to estimate gene duplication rates. *Genome Biol.* **8:** R158. doi: 10.1186/gb-2007-8-8-r158.
Patthy, L. 1999. Genome evolution and the evolution of exon-shuffling—A review. *Gene* **238:** 103–114.
Petrov, D.A., Lozovskaya, E.R., and Hartl, D.L. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384:** 346–349.
Roth, D. and Wilson, J. 1988. Illegitimate recombination in mammalian cells. In *Genetic recombination*, pp. 621–653. American Society for Microbiology, Washington, DC.
Roth, D.B., Porter, T.N., and Wilson, J.H. 1985. Mechanisms of non-homologous recombination in mammalian cells. *Mol. Cell. Biol.* **5:** 2599–2607.
Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204–2215.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Slack, A., Thornton, P.C., Magner, D.B., Rosenberg, S.M., and Hastings, P.J. 2006. On the mechanism of gene amplification induced under stress in *Escherichia coli*. *PLoS Genet.* **2:** e48. doi: 10.1371/journal.pgen.0020048.
Stankiewicz, P. and Lupski, J.R. 2002. Molecular-evolutionary mechanisms for genomic disorders. *Curr. Opin. Genet. Dev.* **12:** 312–319.
Tamura, K., Subramanian, S., and Kumar, S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21:** 36–44.
Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.
Vinckenbosch, N., Dupanloup, I., and Kaessmann, H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl. Acad. Sci.* **103:** 3220–3225.
Walsh, J.B. 1995. How often do duplicated genes evolve new functions? *Genetics* **139:** 421–428.
Walsh, B. 2003. Population-genetic models of the fates of duplicate genes. *Genetica* **118:** 279–294.
Wang, W., Zhang, J., Alvarez, C., Llopart, A., and Long, M. 2000. The origin of the *Jingwei* gene and the complex modular structure of its parental gene, *yellow emperor*, in *Drosophila melanogaster*. *Mol. Biol. Evol.* **17:** 1294–1301.
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.* **99:** 4448–4453.
Wang, W., Yu, H., and Long, M. 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. *Nat. Genet.* **36:** 523–527.
Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., Zhang, G., Liu, D., Zhang, J., Vang, S., et al. 2006. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **18:** 1791–1802.
Xue, B., Rooney, A.P., Kajikawa, M., Okada, N., and Roelofs, W.L. 2007. Novel sex pheromone desaturases in the genomes of corn borers generated through gene duplication and retroposon fusion. *Proc. Natl. Acad. Sci.* **104:** 4467–4472.
Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17:** 32–43.
Yang, S., Arguello, J.R., Li, X., Ding, Y., Zhou, Q., Chen, Y., Zhang, Y., Zhao, R., Brunet, F., Peng, L., et al. 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. *PLoS Genet.* **4:** e3. doi: 10.1371/journal.pgen.0040003.
Zhang, J., Zhang, Y., and Rosenberg, H.F. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* **30:** 411–415.
Zhang, Y., Wu, Y., Liu, Y., and Han, B. 2005. Computational identification of 69 retroposons in *Arabidopsis*. *Plant Physiol.* **138:** 935–948.

Received January 25, 2008; accepted in revised form May 28, 2008.

On the origin of new genes in *Drosophila*

Qi Zhou, Guojie Zhang, Yue Zhang, et al.

Genome Res. 2008 18: 1446-1455 originally published online June 11, 2008
Access the most recent version at doi:10.1101/gr.076588.108