

# Repetitive Element-Mediated Recombination as a Mechanism for New Gene Origination in *Drosophila*

Shuang Yang<sup>1,2</sup>, J. Roman Arguello<sup>3</sup>, Xin Li<sup>1,2</sup>, Yun Ding<sup>1,2</sup>, Qi Zhou<sup>1,2</sup>, Ying Chen<sup>4</sup>, Yue Zhang<sup>1</sup>, Ruoping Zhao<sup>1</sup>, Frédéric Brunet<sup>3</sup>, Lixin Peng<sup>1</sup>, Manyuan Long<sup>3,4\*</sup>, Wen Wang<sup>1\*</sup>

**1** Chinese Academy of Sciences (CAS)—Max Planck Junior Research Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China **2** Graduate School of Chinese Academy Sciences, Beijing, China, **3** Committee on Evolutionary Biology, The University of Chicago, Chicago, Illinois, United States of America, **4** Department of Ecology and Evolution, The University of Chicago, Chicago, Illinois, United States of America

**Previous studies of repetitive elements (REs) have implicated a mechanistic role in generating new chimerical genes. Such examples are consistent with the classic model for exon shuffling, which relies on non-homologous recombination. However, recent data for chromosomal aberrations in model organisms suggest that ectopic homology-dependent recombination may also be important. Lack of a dataset comprising experimentally verified young duplicates has hampered an effective examination of these models as well as an investigation of sequence features that mediate the rearrangements. Here we use ~7,000 cDNA probes (~112,000 primary images) to screen eight species within the *Drosophila melanogaster* subgroup and identify 17 duplicates that were generated through ectopic recombination within the last 12 mys. Most of these are functional and have evolved divergent expression patterns and novel chimeric structures. Examination of their flanking sequences revealed an excess of repetitive sequences, with the majority belonging to the transposable element DNAREP1 family, associated with the new genes. Our dataset strongly suggests an important role for REs in the generation of chimeric genes within these species.**

Citation: Yang S, Arguello JR, Li X, Ding Y, Zhou Q, et al. (2008) Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. PLoS Genet 4(1): e3. doi:10.1371/journal.pgen.0040003

## Introduction

Gene duplication followed by the acquisition of novel molecular function is a fundamental process underlying biological diversity. It has been theoretically and empirically demonstrated that functionally distinct duplicates are capable of evolving through a neofunctionalization process in which there is an accumulation of mutations in a redundant copy of a preexisting gene [1–3]. In addition, there is mounting evidence for the rapid generation of new genes through the recombination of preexisting exons and functional domains. This latter process does not exclude, and in fact often relies on, the duplication of the loci involved [4,5]. Excluding chimeric genes formed through retroposition [6–8], more than three hundred gene families are believed to have originated through exon shuffling [9]. Most of these gene families have introns, suggesting that DNA level recombination was involved (DLR; DLR as opposed to a retroposition event involving an RNA intermediate).

Since its initial proposal [10], the genetic mechanisms involved in the formation of chimeric genes through exon shuffling have largely remained a mystery. The classic model states that nonhomologous recombination (NHR) brings together exons or domains from ectopic positions [10]. Experimental evidence for the role of NHR has been gained through transfection experiments [11,12] and through surveys of rearrangement hotspots which are often disease-associated [13–15]. Breakpoint analyses on these datasets revealed little or no sequence identity between the loci recombined, supporting a NHR model. While these experiments show such a model is possible for exon shuffling, it

remains an open question how frequently such processes in non-artificial systems, and over evolutionary time, will contribute to the formation of fixed chimeric genes.

Another potential NHR mechanism that can mediate nonhomologous recombination is through the activity of transposable elements (TEs). If a TE is capable of mobilizing adjacent sequence, novel junctions that share no sequence identity could be generated [16]. The capacity for such events has been documented with the imprecise excision of well studied TEs such as P elements [17] as well as in plant pack-MULE and Helitron TEs [18–20]. These investigations implicate a role for TEs in the generation of chimeric genes. Whether these shuffled products are under functional constraint remains an interesting question.


Alternatively, non-allelic homologous recombination (NAHR) between ectopic sequences can lead to the formation


**Editor:** R. Scott Hawley, Stowers Institute for Medical Research, United States of America

**Received:** August 22, 2007; **Accepted:** November 27, 2007; **Published:** January 18, 2008

**Copyright:** © 2008 Yang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

\* To whom correspondence should be addressed. E-mail: mlong@uchicago.edu (ML); wwang@mail.kiz.ac.cn (WW)

 These authors contributed equally to this work.

 Current address: Ingénieur de Recherche en Bioinformatique Equipe Génomique Evolutive des Vertébrés, Institut de Génomique Fonctionnelle de Lyon (IGFL)—Ecole Normale Supérieure de Lyon (ENSL) 46, Lyon, France

## Author Summary

In numerous organisms, many new genes have been found to originate through dispersed gene duplication and exon/domain shuffling. What recombination mechanisms were involved in the duplication and the shuffling processes? Lack of the intermediate products of recombination that share adequate sequence identity between homologous sequences, or the parental sequences from which the new genes were derived, often makes answering these questions difficult. We identified a number of young genes that originated in recently diverged branches in the evolutionary tree of the eight *Drosophila melanogaster* subgroup species, by using fluorescence *in situ* hybridization with polytene chromosomes. We analyzed the genomic regions surrounding 17 new dispersed duplicate genes and observed that most of these genes are flanked by repetitive elements (REs), including a large and diverged transposable element family, DNAREP1. Several copies of these REs are kept in both new and parental gene regions, and their degeneration is correlated with the increasing ages of the identified new genes. These data suggest that REs mediate the recombination responsible for the new gene origination.

of chimeric genes. Recently, a surge of evidence has begun to demonstrate the importance of NAHR to genomic architecture, especially in primates [21–26]. Intriguingly, several studies have reported on a limited number of chimeric gene structures, some of which appear functional and nondeleterious, but most remain putative [24,27]. Focus has primarily been placed on NAHR's role in human disease [26]. However, given that NAHR appears to be a common mutational mechanism, a new hypothesis for exon shuffling has been motivated: Despite the frequently deleterious effects, NAHR is capable of making a contribution to the origin of new chimeric genes as an exon shuffling mechanism [24,25,28,29].

A difficulty in investigating the relative contributions of these mechanisms to the formation of chimeric genes is that most of the available examples are evolutionarily ancient [9]. These genes provide few clues for understanding the recombination mechanisms that generated their initial structures because the sequence features, especially those non-constrained sequence traits, that may have fostered their formations have likely been lost (the half life is 120 mys for mammals and 10 mys in *Drosophila* [30]). While sequence analyses of ancient chimeric genes provide little mechanistic insight, a sample of young chimeric genes that potentially retain these sequence features may. A second difficulty arises from the limited number of young chimeric genes that are thought to have arisen by DLR. While several case studies exist, evolutionary analyses demonstrating that the new chimeras are functional are largely lacking [24,27,31].

Here we report on a large-scale experimental genomic screen for young chimeric genes generated by DLR within the *D. melanogaster* subgroup. We utilized an integrated approach based on fluorescent *in situ* hybridizations (FISH), Southern hybridizations, expression and transcript experiments, BLAST queries, and evolutionary analyses. This approach allowed us to focus on dispersed duplication events, ignoring tandem duplications. Consequently, the total number of chimeric formations are likely larger than the total we report on here. Nonetheless, our results show that, rather than providing redundant copies, dispersed duplication events via DLR have generated new chimeric structures at a high

frequency. Interestingly, none of these chimeric structures involved two or more genic sequences; all chimeric regions were formed from the fusion of the duplicated loci and intergenic sequences. Furthermore, we provide strong evidence that REs, in particular the TE family DNAREP1, are a major mediator of these events. Finally, using multiple well-established methods [6,7,32–34], we demonstrate that most of these new chimeric genes are functional.

## Results/Discussion

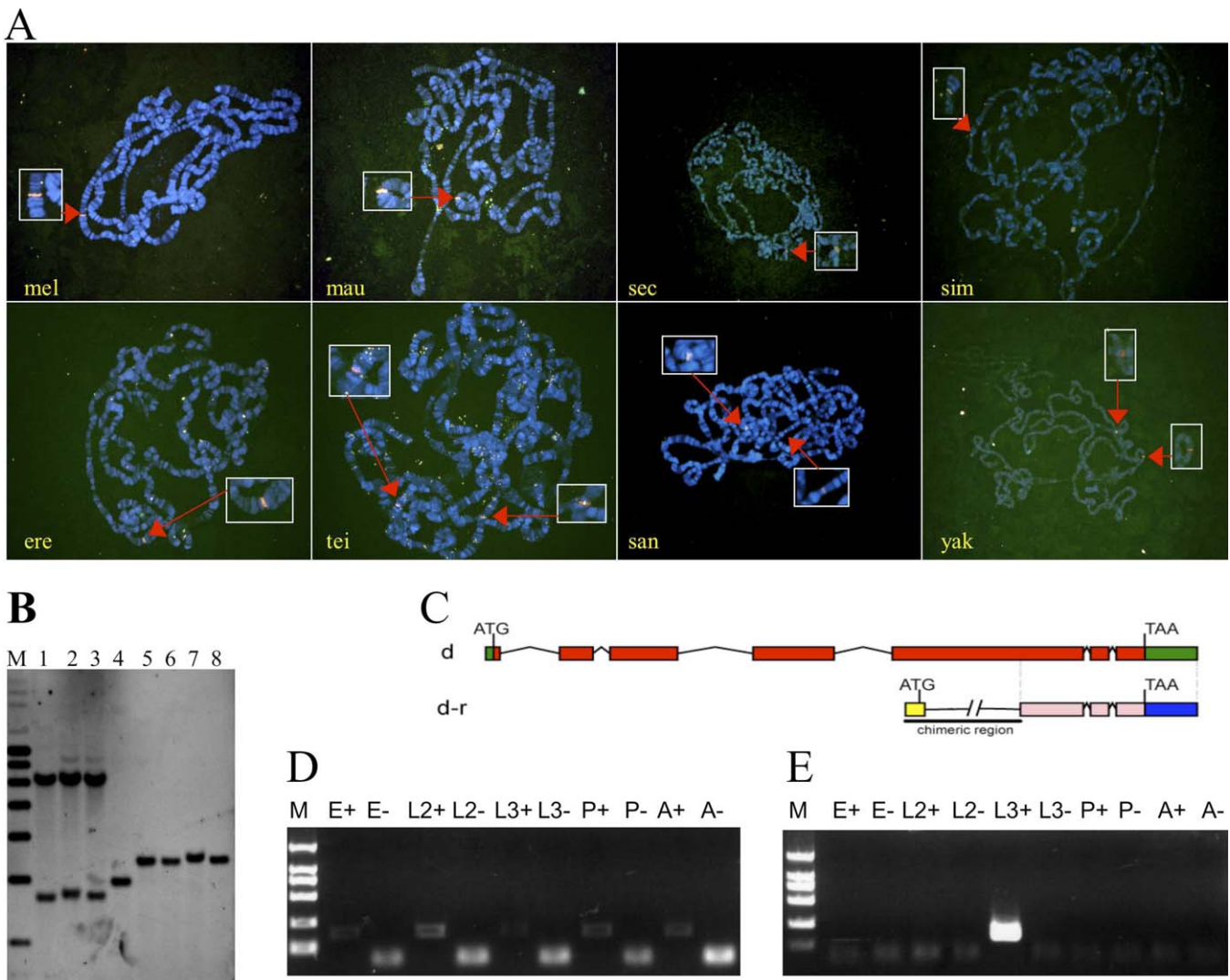
Two cDNA unigene libraries from *D. melanogaster* comprised of ~7,000 cDNA probes were used for cFISH experiments over all tested species. Each hybridization generated at least two images for each species. In total, our experiment produced ~112,000 primary images. Including those probes that gave weak or paradoxical signals, the *Drosophila* Gene Collection (DGC) library version 1.0 set resulted in 266 candidates. The unigene library included 1,000 cDNA probes, most of which were included in the DGC 1.0 library. From this set, 5 new genes, *jingwei* [33], *Hun* [32], *sphinx* [34], *monkey-king* [35], and *Dntf-2r* [36] have previously been described.

To exclude false positives from the 266 candidates, we carried out Southern hybridizations and conducted BLAST searches against the available genome sequences of *D. simulans* (droSim1), *D. yakuba* (droYak1), *D. sechellia* (droSec1), *D. melanogaster* (dm2) and *D. erecta* (droEre1) (<http://genome.ucsc.edu>) (Figure 1). The Southern and BLAST analyses confirmed 17 young duplicates generated through DLR (Table 1; Figure 2). The genomic sequences of all 17 dispersed duplicates contain the intron(s) and/or non-coding flanking sequences that exist in their parental copies, suggesting that the new genes originated through DLR. In addition, we also identified ten new copies of retrogenes and 53 young copies of REs including retroelements and other repetitive sequences. In this report, we have focused on the 17 dispersed duplicates and investigate possible DLR mechanisms that generate dispersed duplications.

Interestingly, the *kep1* gene family has six new duplicates that have been dispersed to different chromosomal locations, while the other 11 gene families have only a single new duplicate (Table 1). Thirteen of these duplications are intrachromosomal, and 4 are interchromosomal (Table 1). Two putative pseudogenes exist in this list: CR33318 and CR9337. CR33318 is found only in *D. melanogaster*, however CR9337 has a disrupted reading frame in *D. melanogaster* but is intact in *D. sechellia* and *D. simulans*. Mapping these results onto the species tree reveal an age <8 mys for almost all these origination events except the 12-my-old CG5372 (Figure 2).

Excluding the two putative pseudogenes (CR33318 and CR9337) paralog-specific reverse transcriptase (RT-PCR) experiments detected transcripts for all paralogs. Twelve out of these 15 duplicates display differential expression patterns from their parental copies in development and/or sex (Table S1). These observations indicate that most of the new genes have evolved divergent expression patterns, and that generally the patterns are more restricted.

To examine whether the new duplicates have evolved chimeric gene structures, we utilized previously reported cDNA sequences, RACE, or RT-PCR based on computationally predicted structures (Materials and Methods). Among the 17 new genes, 13 were found to have evolved chimeric gene



**Figure 1.** An Example Illustrating the Detection of New Genes

(A) The probe LD47348 (CG10595) detected two signals in the clade of *D. yakuba-santomea-teissieri* while only detecting one signal in other species. The new additional signal suggests a new gene candidate.

(B) Southern hybridization results further confirm the extra copy in the *D. yakuba-santomea-teissieri* clade (M is 1-kb extension marker [Invitrogen]). Lanes 1–8 correspond to Xho I digested DNAs of *D. yakuba*, *D. teissieri*, *D. santomea*, *D. erecta*, *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia*, respectively).

(C) Cartoon figure displaying the gene structures of the parental gene (d, or CG10595) and the new duplicate (d-r). The duplicated region is indicated by vertical dash lines. d-r recruited one upstream exon as indicated by yellow box.

(D)

Expression patterns of the parental gene.

(E) Expression patterns of the new gene d-r revealed by one round of RT-PCR and a second round of nested PCR (M indicates DL2000 DNA molecular marker (Takara); E+, E–, L2+, L2–, L3+, L3–, P+, P–, A+, and A– correspond to positive and negative reactions for embryos, second instar larvae, third instar larvae, pupae, and adults, respectively). From these gels, it is clear that d-r is only expressed in the third instar larvae while the parental copy is expressed ubiquitously. All the bands in the negative control lanes are primer dimer bands. E+ and L3+ are weak but clearly visible.

doi:10.1371/journal.pgen.0040003.g001

sequences through the recruitment of flanking sequence near the insertion site or as the result of extensive deletions (CG5372, CG9902, CG4021, CG3875, CG3927, CR9337, CG7635-r, CG3101-r, CG3071-r, *d-r*, *Dox-A3-r*, *Hun*, and *klg-r*; Figure 3). Among these chimeric genes, 11 can encode chimeric proteins (CG5372, CG9902, CG4021, CG3875, CG3927, CR9337, CG7635-r, CG3101-r, CG3071-r, *Hun* and *d-r*). For example, *d-r* and CG9902 have both recruited novel coding regions following their duplications, and possibly in conjunction with their deletions events that followed (Figure 3). These observations reveal that the majority of young duplicated genes have evolved chimeric gene structures. In

addition, it is notable that the chimeric genes that we have detected involve only the duplicated loci and intergenic sequences. This suggests that for dispersed duplication events, the formation of chimeric genes by recombining two or more genic sequences may be relatively rare.

To test for functional constraint, we conducted substitution analyses by estimating the Ka/Ks ratio for both paralogous and orthologous comparisons. For the paralogous comparisons, our conservative null hypothesis was that the parental genes are under strong functional constraint with the new copy subject to no constraint (a pseudogene). These estimates suggest that most of the genes are under functional



**Table 1.** List of the Young Duplicates Identified and their Parental Loci

Probes Used			Parental Gene Name	Parental Gene Location	Species with New Copy	New Copy Name	New Copy Location
DGC1.0 ID	Clone ID	Gene ID					
42E2	GH08776	CG7163	mkg-r	X	mau	mkg-r2	X
16E1	LD46502	CG3584	kep1 (CG3584)	2R	mel, mau, sim, sch	CG3875	2R
			kep1 (CG3584)	2R	mel, mau, sim, sch	CG4021	2R
			kep1 (CG3584)	2R	mel, mau, sim, sch	CR9337	2L
			CG3875	2R	mel, mau, sim, sch	CG3927	2R
			CR9337	2L	mel	CR33318	2L
			CR9337	2L	mau, sim, sch	CR9337-r	2L
55F8	GM14421	CG9902	CG7692	3L	mel, mau, sim, sch	CG9902	X
56D9	LP12257	CG8095	scb (CG8095)	2R	mel, mau, sim, sch, yak,tei, san	CG5372	2R
10E3	LD09009	CG6386	Bällchen (CG6386)	3R	mau, sim, sch	Hun	X
46F7	SD05291	CG7635	Mec2 (CG7635)	X	mau, sim, sch	CG7635-r	X
47H7	LD43561	CG3071	CG3071	X	mau, sim, sch	CG3071-r	X
56A2	LD27988	CG3101	CG3101	X	mau, sim, sch	CG3101-r	X
62E3	GM04983	CG8490	CG8490	2R	mau, sim, sch	CG8490-r	X
38D7	LD47348	CG10595	d (CG10595)	2L	yak, tei, san	d-r	2L
46G11	SD09866	CG2952	Dox-A3 (CG2952)	2R	yak	Dox-A3-r	3R
75A6	LD10776	CG6669	Klg (CG6669)	3R	sim	klg-r	3R

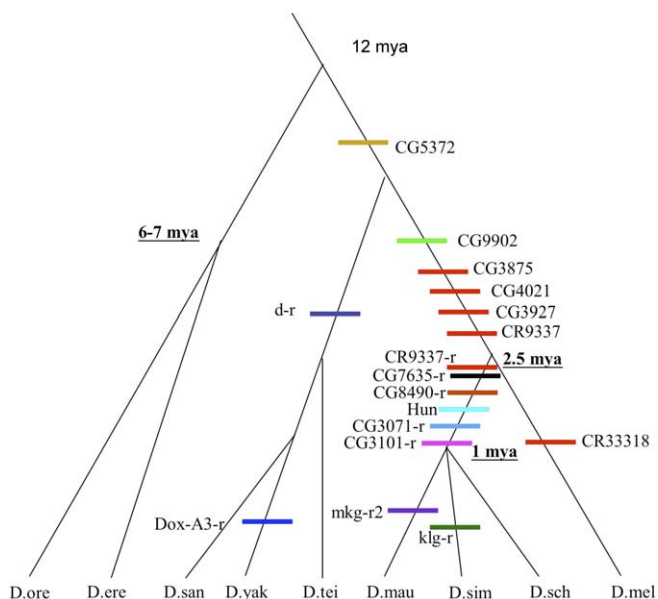
doi:10.1371/journal.pgen.0040003.t001

constraint: Ka/Ks values are lower than 0.5 for 8 genes, lower than 1 but higher than 0.5 for 5 genes, and  $\sim 1$  for 2 genes (Table S2). Furthermore, analyses of the functional domains for these genes (Materials and Methods), revealed that almost all genes have Ka/Ks ratios lower than or close to 0.5 (Table S2). For orthologous comparisons, the null hypothesis was that the new copies are pseudogenes (Ka/Ks = 1). The results were similar, showing that Ka/Ks ratios are significantly less than 1 for most genes except CG3071-r (Ka/Ks = 2.3091) and CG8490-r (Ka/Ks = 1.2230), indicating the possibility that positive selection may be acting on these two (Table S3). The

statistical tests of the null hypothesis of neutrality [1] in the paralogous and orthologous comparisons reveal that most of these new genes are under significant functional constraint over the tested coding sequences. These complementary analyses of expression, gene structure, and nucleotide substitution suggest that all 15 new genes are functional and that many of these have undergone neofunctionalization by evolving new gene structures with new expression patterns.

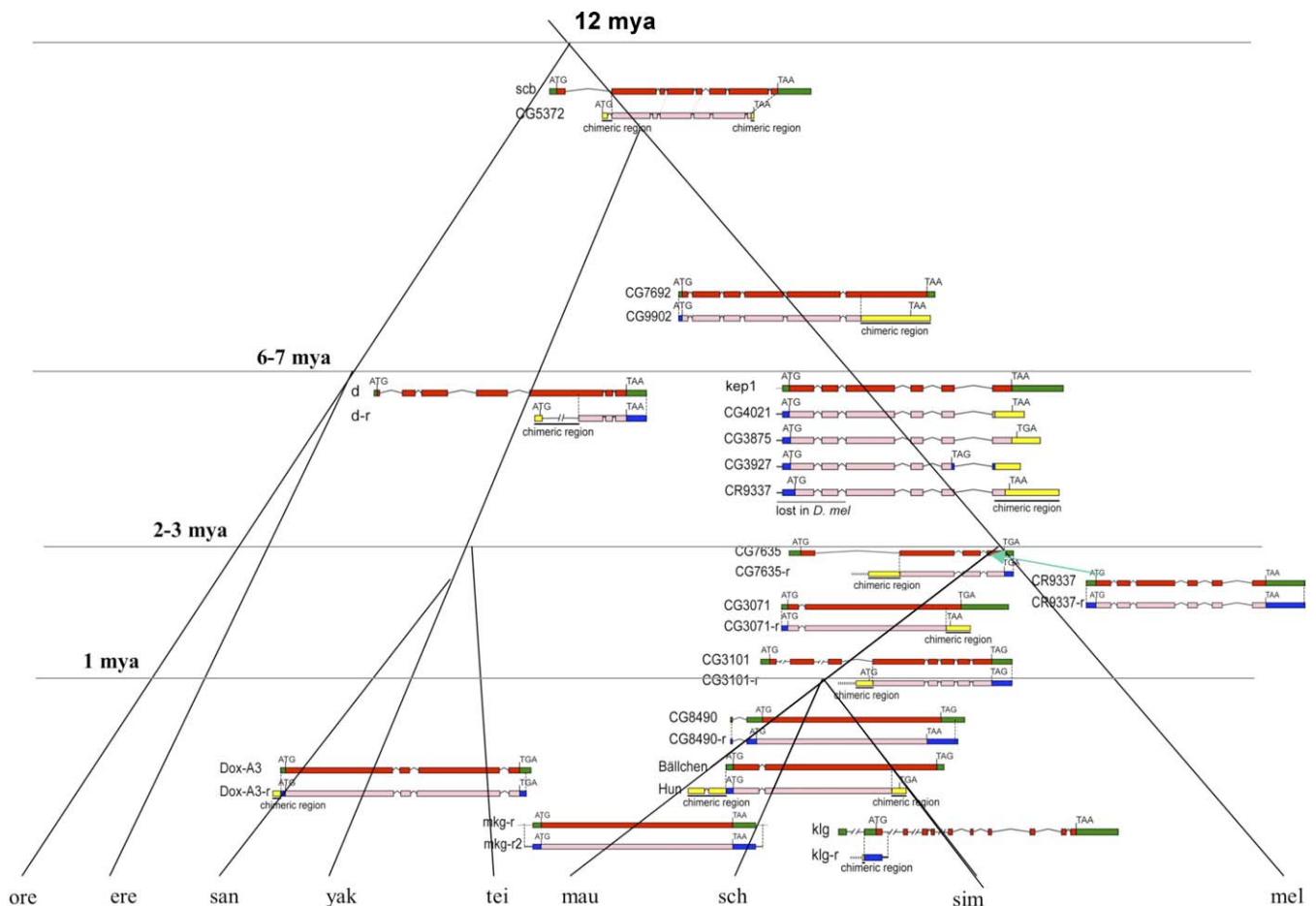
The classical models of gene duplication assume a completely redundant (in sequence and function) duplicate copy [1,2]. In these models the most likely outcome is that one copy will become non-functionalized, with a low probability that one or the other becomes neofunctionalized or subfunctionalized through subsequent mutations [37,38]. However, our results show that the majority of new duplicates generated through DLR in *Drosophila* are not structurally, and are thus unlikely to be functionally, identical to their parental copies. It is also a general result that DLR is an important mechanism for the generation of dispersed genes with novel functions, adding to other potential mechanisms [39]. Interestingly, Katju and Lynch [40] have recently found that many new duplicates in *C. elegans* have unique exons in one or both members of a duplicate pair. Consistent with our observations, these latter cases are also likely DLR-derived duplicates that have recruited new gene fragments and have evolved stable chimeric structures.

Having established that 15 of these new duplicates are likely functional, with many having chimeric structures, we then investigated the mutational mechanisms that generated them. Data, largely originating from detailed sequence analyses of human disease-related loci, have shown correlations between structural variation and REs, most notably *Alu* elements in primate genomes [13,22,23,25,28,41]. Though a causal relationship between the repetitive elements and segmental duplications is difficult to establish, several studies have argued for their causative role in genomic rearrangements through NAHR. Based in part on these findings, we

**Figure 2.** The Phylogenetic Distribution of the 17 New DLR Duplicates Identified in This Study

The species phylogeny and time scale are from [58]. Different color bars show different gene families. The *kep1* gene family has six new duplicates (indicated by red bars).

doi:10.1371/journal.pgen.0040003.g002



**Figure 3.** The Gene Structures of 16 New Duplicates Mapped on the Species Phylogeny

CR33318 is not shown because it is a truncated copy without detectable expression and has frame shift mutations. Duplicated regions are indicated with vertical dash lines. Horizontal dash lines in CG7635-r, CG3101-r, d-r, and klg-r indicate that we only obtained partial coding regions with RT-PCR and longer coding regions may exist outward. Boxes are exon regions and lines indicate introns. Yellow boxes indicate recruited chimeric regions, green boxes indicate parental loci UTRs, and blue boxes indicate duplicate loci UTRs. Positions of start and stop codons are marked. doi:10.1371/journal.pgen.0040003.g003

were interested in whether there was evidence for repetitive sequence surrounding these duplicated regions.

We identified both 5' and 3' breakpoints for each young duplicate by comparing genomic sequences of each of these new gene duplicates with its parental copy (Table 2). Interestingly, we observed REs at or near the breakpoints for 10 out of the 17 duplicates (including the 2 duplicates that are likely pseudogenes) (Table 2; Figure S1). These REs consist of 7 TEs, 2 satellite sequences, and 1 simple repeat. They are associated with the new genes that are in different genomic locations, suggesting independent events. Furthermore, all TEs belong to the DNAREP1 family, the largest TE family in *Drosophila* which has very diverged members [42,43].

Among these 10 pairs associated with REs, 5 have shared repeats at or near the breakpoints of both the parental and the new duplicate copies (Table 2). For these 5 paralog pairs, 4 (CG3875-CG3927, mkg-r-mkg-r2, CG3101-CG3101-r and CR9337-CR9337-r) maintain very high sequence identity over the flanking elements; the remaining CR9337-CR33318 pair, though both harboring DNAREP1 sequence at their 5' ends, provides a weak alignment. The other five paralog pairs contain a repetitive element at the breakpoint of one copy (Table 2; 2 examples with highly similar TEs shown in Figure

4). In addition, *klg-r*, CG7635-r and CG8490-r (not included in the ten above) were found next to sequencing gaps in the genomic databases (Table 2), and resequencing these regions resulted in sequence profiles characteristic of repetitive sequences (data not shown). If these are included, the majority of new duplicates (13/17, 76.5%) are associated with repetitive elements.

Four lines of evidence indicate that this association has not been observed by chance. The first is based on orthology assignments available from current genome databases, indicating that all ten in our set are euchromatic and not on the 4<sup>th</sup> chromosome. High-resolution analyses of *D. melanogaster* TEs have verified that the paracentromeric regions of the major chromosome arms and chromosome 4 harbor the highest densities of TEs [44]. Second, simulations show that the probability that the number of genes flanked by TEs  $\geq 7$  given the sample size of seven genes (with 14 breakpoints) is low ( $p < 0.05$ ) given a TE-free region (TFR) of  $\sim 15$  kb or larger (Figure S2; Materials and Methods). Despite TE differences between species, 15 kb is less than half the mean TFR found in *D. melanogaster* [44]. Given that the TEs in our dataset are comprised primarily of DNAREP1 family members, the distance is even greater. Furthermore, the proba-

**Table 2.** Repetitive Elements at the Breakpoints of Duplicate Pairs

Pair of Duplication	Species Analyzed	5' Breakpoint	Length of Repeats	Repeat Class/Family	3' Breakpoint	Length of Repeats	Repeat Class/Family
mkg-r	<i>mau</i>	SAR_DM	355 bp	Satellite			
mkg-r2		SAR_DM	343 bp	Satellite			
CG3101	<i>sim</i>	SAR_DM	755 bp	Satellite	SAR_DM	105 bp	Satellite
CG3101-r		SAR_DM	602 bp	Satellite	SAR_DM	433 bp	Satellite
kep1	<i>mel</i>						
CG3875		DNAREP1_DM	230 bp	Transposon			
CG3875	<i>sim</i>	DNAREP1_DM	58 bp	Transposon	DNAREP1_DM	54 bp	Transposon
CG3927		DNAREP1_DM	51 bp	Transposon			
kep1	<i>mel</i>						
CG4021				DNAREP1_DM	179 bp		Transposon
kep1	<i>sch</i>						
CR9337		DNAREP1_DM	338 bp	Transposon	DNAREP1_DM	367 bp	Transposon
CR9337	<i>sch</i>	DNAREP1_DM	338 bp	Transposon	DNAREP1_DM	367 bp	Transposon
CR33318	<i>mel</i>	DNAREP1_DM	206 bp	Transposon			
CR9337	<i>sch</i>	DNAREP1_DM	339 bp	Transposon	DNAREP1_DM	368 bp	Transposon
CR9337-r		DNAREP1_DM	357 bp	Transposon	DNAREP1_DM	408 bp	Transposon
CG7692	<i>mel</i>						
CG9902		DNAREP1_DM	398 bp	Transposon			
CG2952	<i>yak</i>						
CG2952-r					(TATATG) <sub>n</sub>	44 bp	Simple-repeat
Mec2	<i>sch</i>						
CG7635-r				Gap			
CG8490	<i>sim</i>						
CG8490-r		Gap		Gap			
klg	<i>sim</i>						
klg-r				Gap			

doi:10.1371/journal.pgen.0040003.t002

bility that both paralogs contain the same TE sequence in their flanking regions, as three (and possibly four) do in our dataset, is much lower (Table 2; Figure S1). Finally, our data reveal a gradation of degeneration in the TEs and other REs with the ages of the gene duplicates that the repeats flank (Figure 5). This gradation is consistent with observed degeneration rate of functionless elements in *Drosophila* [30], as well as any potential internal deletions that could be part of a self-regulation system as seen in *D. melanogaster* TEs [45].

The striking association with REs provides evidence for the relationship between RE sequences and genomic rearrangements leading to novel functions. This relationship differs from previous reports of TE themselves becoming part of a novel transcript in *D. melanogaster* [46,47]. Instead, our dataset supports a model whereby REs are mediating the recombination of flanking sequences to form chimeric products that do not include RE sequence. The precise mechanism defining “RE-mediation” would likely be NAHR or the mobilization of flanking sequence through the activity of the DNAREP1 transposons. Recent studies of DNAREP1 elements suggest a burst of activity occurred just prior to or during the formation of the *D. melanogaster* subgroup, followed by nearly complete inactivation ~5–10 mya [42]. Interestingly, there is evidence of a very recent revival of activity in the *D. yakuba* lineage [43]. If these estimates on inactivity are correct, NAHR would be the most likely mechanism generating the rearrangement in our dataset. This possibility is also supported by the identified non-mobile repeat sequences that are associated with the new chimeric genes (Table 2). However, if DNAREP1 has been active in the *D. melanogaster*

subgroup for a longer period than reported, as implicated by the observation in *D. yakuba* [43], and if this class of TEs does in fact mobilize flanking DNA, a combination of mechanisms is possible.

Alternatively, the REs flanking the new duplicates could be the result of larger duplications that included the REs (segmental duplication), rather than the REs mobilizing the region. However, we would expect that under this hypothesis we would see longer stretches of identity outside REs. Inspecting the flanking regions of our dataset indicate that identity is lost in close proximity with the repetitive sequences. A second alternative hypothesis is that the repetitive sequence presents a preferential site for strand breakage. Similar suggestions have been made for *Alu*, satellite repeats, and other sequence demonstrating fragility [23,31,48]. If imperfect repair were to follow strand breakage, this too would be akin to a nonhomologous end-joining event and would support the classical view of exon-shuffling. Further experimental work is needed to address this possibility.

Our observation that there is an excess of repetitive elements around dispersed functional duplicates is of general importance in light of advancements in identifying copy number variation in other model organisms, and the increased recognition for the role of repetitive sequences in shaping chromosomal architecture [14,22–26,31,49,50]. Despite these advancements, little is known about the potential non-deleterious outcomes that such rearrangements may present. Our work helps fill this void by providing an extensive chimeric gene dataset that is supported by experiments that test for functionality.



duplicated 3' regions have resulted in varying peptide sequences in the C terminal (Figure 3). In CG8490-r, both the start and the stop codons have been shifted, resulting in different peptides at both N and C termini. Finally, CG3101-r has recruited part of its previous intron 3, which becomes the 5' UTR and a short stretch of protein-coding sequence in the new duplicate gene.

We have used ~7,000 cDNA probes to screen new gene duplicate copies. The estimated number of genes in the genome is ~14,000. The total number of new gene duplicates can be estimated as  $177,000 \times 14,000 = 34$ , over an evolutionary time equal to ~20 mys (the sum of the branch lengths of the *D. melanogaster* subgroup). Thus, on average, the origination rate is  $34/20 = 1.7$  per mys per genome, or  $0.121 \times 10^{-9}$  per year per gene. We note that, because our method ignores tandem duplicates, and because our FISH probes were all based on *D. melanogaster* sequence, this is an underestimate. However, this rate is an order of magnitude higher than the gene duplication rate estimated in yeast [52] but still 30 times lower than a previous estimate that were based on the assumption of a molecular clock [53]. Our estimate may not be inconsistent with previous estimates [53] because our focus was much narrower, investigating DLR events only.

Only two new duplicates (*d-r* and *Dox-A3-r*) in the *yakuba-santomea-teissieri* lineage (*yakuba* lineage) were observed, while 5 new duplicates were detected between the common ancestor of *melanogaster* and *yakuba* and the common ancestor of the *melanogaster* complex (Figure 2). In addition, we did not detect any new duplicate in *D. erecta*. This may be a technical result attributable to the difficulty of hybridization with *D. erecta* polytene chromosomes, or sequence divergence relative to our probes. Alternatively, the putative inconsistent duplication rate may be associated with episodic activities of transposons or repetitive sequences. For example, the transposon DNAREP1 members were associated with 5 new duplicates in the *kep1* gene family and CG9902. As noted above, it has been suggested that there was an active episode of DNAREP1 before the *D. melanogaster* lineage separated from the *D. yakuba* lineage and then again within the *D. yakuba* lineage [42,43].

Previous investigations have revealed several important roles for REs in the generation of evolutionary novelties including the donation of their own sequences into protein coding regions [46,47,54,55], retrotransposing and recruiting novel gene sequence [5], increasing genic diversity in the maize genome by the helitron-like transposons [56], potentially providing greater overall genome plasticity [16,57], and elevating expression of a nearby insecticide resistant gene [58,59]. The observation reported here further demonstrates a mechanistic role for REs in mediating the origins of new genes by facilitating gene recombination. The precise mechanism for this recombination is unclear, but likely include NAHR, as implicated by both TEs and non-TE repetitive sequences being detected, and NHR as a consequence of transposon enzymatic activities [43]. However, the conventional NAHR model is much more likely between the homologous repeats that are located on the same chromosome [22]. Four of the 17 new genes identified are on different chromosomes from their parental genes. These four new genes may have been generated by a different homology-dependent recombination model that assumes a replication-

dependent mechanism involving no crossover [22], the explicit model depicted in Figure 8.

## Materials and Methods

**Materials.** In order to screen for young chimeric genes systematically, we designed an experimental genomics approach using the *D. melanogaster* species subgroup as a comparative model system. This subgroup includes *D. melanogaster* (hereafter abbreviated as *mel* in presented tables and figures), *D. simulans* (*sim*), *D. mauritiana* (*mau*), *D. sechellia* (*sec*), *D. yakuba* (*yak*), *D. teissieri* (*tei*), *D. santomea* (*san*), and *D. erecta* (*ere*). *D. orena* was excluded from analyses because of its unclear placement in the phylogeny. The phylogeny of this subgroup is well resolved [60,61] and the divergence times among these species provide a considerable range over which to detect the presence of young genes. The polytene chromosomes of the salivary gland of *Drosophila* allow detection of gene copy number using a fluorescent *in situ* hybridization (FISH) approach. Therefore, we can use cDNA probes to visualize FISH signals that are about 100 kb away from each other in the species of *D. melanogaster* subgroup, and count the signal number in each species [34].

**FISH and Southern hybridizations.** We carried out dual-color FISH on the polytene chromosome preparations of the aforementioned 8 species. Our probe sets comprised 5,928 full-length *D. melanogaster* cDNA clones from the Berkeley Drosophila Gene Collection (DGC) version 1.0 (<http://www.fruitfly.org/DGC/index.html>) and about 1,000 cDNA clones from an early *Drosophila* Unigene Library (Research Genetics).

Probes were labeled with digoxigenin (DIG) or biotin using PCR [34,62]. Polytene chromosomes from four species were simultaneously squashed on a slide and then hybridized with a pair of DIG and biotin labeled probes [34]. For a given probe, FISH is capable of resolving two signals across two adjacent polytene bands, which is equivalent to ~100 kb in linear DNA sequence. As a result, all duplicates we report in this study have been involved in translocations; they are not tandem duplications. The probes that revealed extra signals in a particular lineage were subject to further confirmation using southern hybridization. Genomic DNAs of the eight species were extracted using the Puregene DNA isolation kit (Gentra Systems). DNAs digested with restriction enzymes were separated on agarose gels and transferred to nylon membranes (Roche Molecular Biochemicals) by Southern blotting. The DIG-labeled probes were hybridized to the membrane to further confirm the copy number in different species. In addition, homology searches were carried out for those new genes that fell within sequenced genomes (<http://genome.ucsc.edu>).

**Breakpoint analyses.** To identify breakpoints and examine the type of sequence surrounding them, the genomic sequences of each pair of duplicate and parental copy, along with 5' and 3' flanking sequences, were aligned using the *bl2seq* software with default settings [63]. The length of the 5' and 3' flanking sequences for each pair was chosen to ensure that it extends 1kb beyond the point where sequence identity disappears. Breakpoints of duplicates were determined as the last nucleotide showing sequence identity between parental and new copy. For a multiple-copy gene family, the parental copy was defined as the copy that has the highest similarity to the new copy. RepeatMasker (<http://www.repeatmasker.org/>) was used to identify whether there is repetitive sequence within a 100 bp window centered at each breakpoint.

**Substitution analyses.** To examine the evolutionary forces operating on the new duplicates, we calculated synonymous (Ks) and non-synonymous (Ka) divergence between all paralogs except for the pseudogene CR33318 (we included the putative pseudogene CR9337 and CR9337-r because they are still intact in the *D. simulans* complex). In addition, we also conducted substitution analyses between orthologous copies in different species. For 11 young duplicates we retrieved their orthologs from a second species's genome, and therefore also calculated Ka and Ks between the orthologous pairs. Estimates were obtained using MEGA 3.1 [64]. A Z-test implemented in MEGA 3.1 was used to test if Ka/Ks ratios deviate from the neutral expectation (Ka/Ks = 1). We tested functional constraint in the whole gene coding region and the functional domain separately. To define the functional domains, the coding sequences of genes were translated into the protein sequences. Then we performed rps-BLAST to detect whether the newly translated protein sequences have functional domains using a cutoff line  $E < 0.01$  on NCBI website <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>.

**RACE, RT-PCR, and gene structure analyses.** Our approach is capable of observing three kinds of new duplicates, (1) direct duplicates that still keep intron(s) or flanking non-coding sequences,



(2) retroposed copies that have lost ancestral introns, and (3) copies that have no obvious sequence features identifying them as either created by retroposition or direct duplication. Tandem duplication can be resulted from either replication slippage or DLR, but the assumption is that those dispersed duplicates across long chromosome distance, or between chromosomes, have originated through DLR. In this study, we only considered direct dispersed duplicates that were derived through DLR. For each of these duplicate genes, we designed copy-specific RT-PCR primers. RT-PCR experiments were carried out using cDNA from 5 developmental stages: embryo, instar larva 2 (L2), instar larva 3 (L3), pupa and adult. Total RNA was extracted from these samples using RNeasy Mini RNA extraction kit (Qiagen). To avoid contamination of genomic DNA, total RNA was treated with Dnase I (amplification grade, Invitrogen) prior to first strand synthesis. First strand cDNA was synthesized using Oligo-dT and SuperScript II Rnase H- reverse Transcriptase (Invitrogen). All RT-PCR products were sequenced for verification.

To establish the gene structures of the new genes, four types of data were used: (1) the draft genomes of *D. simulans* (droSim1), *D. yakuba* (droYak1), *D. sechellia* (droSec1), and *D. erecta* (droEre1) (<http://genome.ucsc.edu>) were queried and provided addition verification and gDNA for primer design; (2) For those duplicates whose full length cDNAs are available in public databases (<http://www.ncbi.nlm.nih.gov/Database/>), we mapped the cDNA to their genomic positions if draft sequence was available; (3) For those duplicates without cDNA, and whose sequences have diverged enough to allow copy-specific primers, we carried out rapid amplification of cDNA ends (RACE); (4) For those duplicate pairs that are too similar to allow copy-specific primers, and for those that resulted in no RACE product (possibly due to low expression levels or long ends), we used the Softberry software [65] to obtain a tentative chimeric gene structure prediction. We then tested these predictions using RT-PCR.

**Chromosomal mapping.** To establish an approximate chromosomal position (interstitial or not) for each of these genes, we used the *D. melanogaster* genome as a reference. We carried out BLAST queries of the *D. melanogaster* genome using sequence flanking each of the genes. These flanking regions were then used to query available genome draft sequence (<http://www.ncbi.nlm.nih.gov/Database/>) in order to determine orthologous chromosomal regions. The cytological positions were then extracted using NCBI's MapView (<http://www.ncbi.nlm.nih.gov/mapview/>). Two new copies (CG7635 and klg), fell between sequence gaps. For these two we determined their approximate position based on our FISH images.

**TE association simulation.** To assess the significance of our observed association between TE sequences and the flanking regions of the paralogs, we carried out simulations based on the known frequencies of TEs in *D. melanogaster* [44]. The mean TE-free region (TFR) is 23,878, with a median of 1,992. The difference between the mean and the median results from the clustering of TEs within the pericentric regions and the fourth chromosome. However, the identified new genes are non-pericentromeric regions in which the density of TEs is much lower and there are few cases of non-random insertions to one particular locus. Therefore, we carried out simulations over a range of normally distributed TFRs in a conservative assumption of the 15 kb average. The length of each TE was normally distributed with a mean of 4 kb. The total length of simulated chromosomes was kept at ~ 20 Mb. 14 breakpoints were introduced randomly into the sequence (seven paralog pairs where only one copy is associated with TE sequence) and an association was considered if the breakpoint was within 300 bp. This distance was also chosen to be conservative, given the distances observed in our data.

## References

- Kimura M (1983) The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
- Ohno S (1970) Evolution by gene duplication. New York: Springer.
- Ohta T (1983) On the evolution of multigene families. *Theor Popul Biol* 23: 216–240.
- Gilbert W (1987) The exon theory of genes. *Cold Spring Harb Symp Quant Biol* 52: 901–905.
- Long M, Betrán E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet* 4: 865–875.
- Betran E, Thornton K, Long M (2002) Retroposed new genes out of the X in *Drosophila*. *Genome Res* 12: 1854–1859.
- Emerson JJ, Kaessmann H, Betran E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. *Science* 303: 537–540.
- Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biology* 3: e357. doi:10.1371/journal.pbio.0030357

10,000 iterations were run and the upper 5% tail was calculated from the resulting distribution.

## Supporting Information

**Figure S1.** The Alignments of Gene Duplicate Copies and Their Flanked Repetitive Sequences

Found at doi:10.1371/journal.pgen.0040003.sg001 (114 KB PDF).

**Figure S2.** The Simulation Results of the TE Association with Gene Duplications

Vertical red line indicates the observed TE-associated genes in our paralog set. The distribution is from simulation where the mean TE-free regions are 15 kb, the mean distance at which our observation is significant at the 0.05 level [44].

Found at doi:10.1371/journal.pgen.0040003.sg002 (3 KB PDF).

**Table S1.** Expression Pattern of the New Genes and Their Parental Genes

Found at doi:10.1371/journal.pgen.0040003.st001 (135 KB DOC).

**Table S2.** Substitutions between Paralogous Copies

The *p*-values in black are for the tests of the null hypothesis that *Ka/Ks* is significantly lower than 1. The *p*-values in red are for the null hypothesis that *Ka/Ks* is significantly lower than 0.5. *p*-Values for paralog comparisons (red) are shown only when the *Ka/Ks* value is lower than 0.5.

Found at doi:10.1371/journal.pgen.0040003.st002 (56 KB DOC).

**Table S3.** Substitutions between Orthologous Copies

The *p*-values are for the tests of the null hypothesis that *Ka/Ks* is significantly lower than 1.

Found at doi:10.1371/journal.pgen.0040003.st003 (52 KB DOC).

## Acknowledgments

We would like to thank James Shapiro for insightful discussion regarding TEs; the M. Long lab for many helpful discussions; The University of Chicago sequencing center for sequencing PCR products; and the *Drosophila* Comparative Genome Sequencing, and Analysis Consortium for the genome sequences of the *melanogaster* subgroup.

**Author contributions.** ML and WW conceived and designed the experiments. SY, JRA, ML, and WW analyzed data. SY, XL, YD, QZ, YC, YZ, RZ, FB, LP, and WW performed molecular and cytological experiments. JRA conducted computer simulation. SY, JRA, ML, and WW wrote the paper.

**Funding.** This work was supported by a CAS-Max Planck Society Fellowship, a National Natural Science Foundation of China (NSFC) award (number 30325016), a NSFC key grant (number 30430400), and a 973 Program (number 2007CB815703–5) to WW; a US National Science Foundation CAREER award (MCB0238168) and US National Institutes of Health R01 grants (R01GM065429-01A1 and 1R01GM078070-01A1) to ML at the University of Chicago; a Graduate Assistance in Areas of National Need (GAANN) genomics grant supports JRA.

**Competing interests.** The authors have declared that no competing interests exist.

- Patthy L (1995) Protein evolution By exon-shuffling. New York: Springer-Verlag.
- Gilbert W (1978) Why genes in pieces? *Nature* 271: 44.
- Van Rijk A, de Jong WW, Bloemendal H (1999) Exon shuffling mimicked in cell culture. *Proc Natl Acad Sci U S A* 96: 8074–8079.
- Van Rijk A, Bloemendal H (2003) Molecular mechanisms of exon shuffling: illegitimate recombination. *Genetica* 118: 245–249.
- Kumatori A, Faizunnessa NN, Suzuki S, Moriuchi T, Kurozumi H, et al. (1998) Nonhomologous recombination between the cytochrome b(558) heavy chain gene (CYBB) and LINE-1 causes an X-linked chronic granulomatous disease. *Genomics* 53: 123–128.
- Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, et al. (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* 437: 94–100.
- Zucman-Rossi J, Legoix P, Victor J-M, Lopez B, Thomas G (1998) Chromosome translocations based on illegitimate recombination in human tumors. *Proc Natl Acad Sci U S A* 95: 11786–11791.

16. Shapiro JA (2005) A 21st century view of evolution: genome system architecture, repetitive DNA, and natural genetic engineering. *Gene* 345: 91–100.
17. Voelker RA, Greenleaf AL, Gyurkovics H, Wisely GB, Huang S-M, et al. (1984) Frequent imprecise excision among reversions of a P element-caused lethal mutation in *Drosophila*. *Genetics* 107: 279–294.
18. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431: 569–573.
19. Kapitonov VV, Jurka J (2001) Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A* 98: 8714–8719.
20. Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC (2003) The maize genome contains a helitron insertion. *Plant Cell* 15: 381–391.
21. Alexander JRB, Schiestl RH (2000) Homologous recombination as a mechanism for genome rearrangements: environmental and genetic effects. *Hum Mol Genet* 9: 2427–2334.
22. Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, et al. (2003) Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Res* 13: 2519–2532.
23. Bailey J, Liu G, Eichler EE (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* 73: 823–834.
24. Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, et al. (2002) Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am J Hum Genet* 70: 38–100.
25. Sharp AJ, Cheng Z, Eichler EE (2006) Structural variation of the human genome. *Annu Rev Genomics Hum Genet* 7: 407–442.
26. Stankiewicz P, Lupski JR (2002) Molecular-evolutionary mechanisms for genomic disorders. *Curr Opin Genet Dev* 12: 312–319.
27. Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, et al. (2005) Complex genomic rearrangements lead to novel primate gene function. *Genome Res* 15: 343–351.
28. Inoue K, Lupski JR (2002) Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet* 3: 199–242.
29. Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55: 1–24.
30. Petrov DA, Lozovskaya ER, Hartl DL (1996) High intrinsic rate of DNA loss in *Drosophila*. *Nature* 384: 346–349.
31. Bailey J, Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* 7: 552–564.
32. Arguello JR, Chen Y, Yang S, Wang W, Long M (2006) Origination of an X-linked testes-specific chimeric gene by illegitimate recombination in *Drosophila*. *PLoS Genetics* 2: e77. doi:10.1371/journal.pgen.0020077
33. Long MY, Langley CH (1993) Natural-selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science* 260: 91–95.
34. Wang W, Brunet FG, Nevo E, Long M (2002) Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 99: 4448–4453.
35. Wang WY, Yu HJ, 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.
36. Betran E, Long M (2003) Dntf-2r, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. *Genetics* 164: 977–988.
37. Force A, Lynch M, Pickett FB, Amores A, Yan YL, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531–1545.
38. Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459–473.
39. Johnson ME; NISC Comparative Sequencing Program, Cheng Z, Morrison VA, Scherer S, et al. (2006) Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad Sci U S A* 103: 17626–17631.
40. Katju V, Lynch M (2006) On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Mol Biol Evol* 23: 11056–11067.
41. López-Correa C, Dorschner M, Brems H, Lázaro C, Clementi M, et al. (2001) Recombination hotspot in NF1 microdeletion patients. *Hum Mol Genet* 10: 1387–1392.
42. Kapitonov VV, Jurka J (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc Natl Acad Sci U S A* 100: 6569–6574.
43. Yang HP, Hung TL, You ZL, Yang ZH (2006) Genomewide comparative analysis of the highly abundant transposable element DINE-1 suggests a recent transpositional burst in *Drosophila yakuba*. *Genetics* 173: 189–196.
44. Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the *D. melanogaster* genome. *Genome Biol* 7: R112.
45. Galindo MI, Ladeveze V, Lemeunier F, Kalmes R, Periquet G, et al. (1995) Spread of the autonomous transposable element hobo in the genome of *Drosophila melanogaster*. *Mol Biol Evol* 12: 723–734.
46. Lorenc A, Makalowski W (2003). Transposable elements and vertebrate protein diversity. *Genetica*. 118: 183–191.
47. Nekrutenko A, Li WH (2001) Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* 17: 619–621.
48. Zhou Y, Mishra B (2005) Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc Natl Acad Sci U S A* 15: 4051–4056.
49. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 38: 75–81.
50. Redon RIS, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. (2006) Global variation in copy number in the human genome. *Nature* 444: 444–454.
51. Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL (1998) Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 396: 572–575.
52. Gao LZ, Innan H (2004) Very low gene duplication rate in the yeast genome. *Science* 306: 1367–1370.
53. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155.
54. Cordaux R, Udit S, Batzer MA, Feschotte C (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci USA* 103: 8101–8106.
55. Makalowski W, Mitchell GA, Labuda D (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet* 10: 188–193.
56. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, et al. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* 37: 997–1002.
57. Capy P (1998) A plastic genome. *Nature* 396: 522–523.
58. Schlenke TA, Begun DJ (2004) Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc Natl Acad Sci USA* 101: 1626–1631.
59. Brookfield JFY (2004) Evolutionary genetics: mobile DNAs as sources of adaptive change? *Curr Biol* 14: R344–R345.
60. Powell JR (1997) Progress and prospects in evolutionary biology—the *Drosophila* model. New York: Oxford University Press.
61. Lachaise D, Harry M, Solignac M, Lemeunier F, Benassi V, et al. (2000) Evolutionary novelties in islands: *Drosophila santomea*, a new melanogaster sister species from Sao Tome. *Proc R Soc Lond B Biol Sci* 267: 1487–1495.
62. Wang W, Zhang J, Alvarez C, Llopert A, 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
63. Tatusova TA, Madden TL (1999) Blast 2 sequences—a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 174: 247–250.
64. Kumar S, Tamura K, Nei M (2004) Brief Bioinform 5: 150–163.
65. Salamov AA, Solovyev VV (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* 10: 516–522.