J Mol Evol (2009) 69:81–93
DOI 10.1007/s00239-009-9254-1

Evolutionary Dynamics of Recently Duplicated Genes: Selective Constraints on Diverging Paralogs in the *Drosophila pseudoobscura* Genome

Richard P. Meisel

Received: 14 December 2008 / Accepted: 26 May 2009 / Published online: 18 June 2009

Abstract Duplicated genes produce genetic variation that can influence the evolution of genomes and phenotypes. In most cases, for a duplicated gene to contribute to evolutionary novelty it must survive the early stages of divergence from its paralog without becoming a pseudogene. I examined the evolutionary dynamics of recently duplicated genes in the *Drosophila pseudoobscura* genome to understand the factors affecting these early stages of evolution. Paralogs located in closer proximity have higher sequence identity. This suggests that gene conversion occurs more often between duplications in close proximity or that there is more genetic independence between distant paralogs. Partially duplicated genes have a higher likelihood of pseudogenization than completely duplicated genes, but no single factor significantly contributes to the selective constraints on a completely duplicated gene. However, DNA-based duplications and duplications within chromosome arms tend to produce longer duplication tracts than retroposed and inter-arm duplications, and longer duplication tracts are more likely to contain a completely duplicated gene. Therefore, the relative position of paralogs and the mechanism of duplication indirectly affect whether a duplicated gene is retained or pseudogenized.

Keywords Drosophila · Gene duplication · Pseudogene · Copy number polymorphism

Introduction

Gene duplication is an important evolutionary process (Ohno 1970) that allows for the expansion of gene families (e.g., Nozawa and Nei 2007), for the evolution of genes involved in the regulation of development (Sidow 1996), and for the evolution of cellular biochemical pathways (Wang and Gu 2001). If a newly duplicated gene is to contribute to evolutionary novelty as a protein coding sequence, it must survive the early stages of divergence from its paralog without becoming a pseudogene. I examined the early evolution of duplicated genes in the *Drosophila pseudoobscura* genome. Recently duplicated genes were interrogated to understand the factors influencing the nucleotide divergence between paralogs and the forces governing the retention of their open reading frames.

Immediately following the duplication of a gene, the derived copy will be found on a single chromosome. If it survives transmission to the next generation, the duplicated gene will segregate as a copy number polymorphism (CNP) until it is either lost or fixed in the population (Dopman and Hartl 2007; Emerson et al. 2008; Sebat et al. 2004; Turner et al. 2008). As paralogs diverge, they may experience one of a number of possible evolutionary trajectories (Lynch and Conery 2000; Moore and Purugganan 2005), including pseudogenization, subfunctionalization (Force et al. 1999; Hughes 1994), or neofunctionalization (Byrne and Wolfe 2007; He and Zhang 2005). The evolutionary fate of a duplicated gene may be influenced by the molecular mechanism responsible for the duplication. For example, retroposition will duplicate an open reading frame, but not

Electronic supplementary material The online version of this article (doi:10.1007/s00239-009-9254-1) contains supplementary material, which is available to authorized users.

R. P. Meisel
Intercollege Graduate Program in Genetics and Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA

R. P. Meisel (✉)
Department of Molecular Biology and Genetics, Cornell University, 227 Biotechnology Bldg., Ithaca, NY 14853, USA
e-mail: rpm16@cornell.edu

untranscribed *cis* regulatory regions responsible for transcriptional control. DNA-based mechanisms, on the other hand, may duplicate *cis* regulatory regions along with the coding region. In this case, the mechanism of duplication could affect the expression profile of a duplicated gene, which may influence whether the derived copy becomes a pseudogene. Additionally, the functional properties of a coding sequence can influence the likelihood that a duplication of the coding sequence will be retained (Kondrashov et al. 2002; Papp et al. 2003; Seoighe and Gehring 2004). Furthermore, the relative location of the two paralogs has also been shown to affect the selective constraints on duplicated genes (Cusack and Wolfe 2007; Zhou et al. 2008).

Comparative studies of genome wide patterns of duplication are now possible with the availability of multiple sequenced genomes from closely related species (e.g., Demuth et al. 2006; Hahn et al. 2007; Zhou et al. 2008). Previous work on the sequence divergence between paralogs within the *Drosophila melanogaster* genome suggests that the majority of newly arising duplicated genes are pseudogenized (Lynch and Conery 2000), although this conclusion is not unanimous (Long and Thornton 2001). With multiple sequenced *Drosophila* genomes (Adams et al. 2000; *Drosophila* 12 Genomes Consortium 2007; Richards et al. 2005) we can now infer duplication events along individual lineages (Hahn et al. 2007; Heger and Ponting 2007; Zhou et al. 2008). The ancestral karyotype in the genus consists of a telocentric X-chromosome, four telocentric major autosomes, and a small dot chromosome. The six chromosome arms are referred to as Muller elements A-F (Muller 1940). Throughout the evolution of the genus, chromosome arms have fused and gene order has been shuffled within arms, but very little gene movement occurs between arms (Bhutkar et al. 2007; *Drosophila* 12 Genomes Consortium 2007; Heger and Ponting 2007). This facilitates the mapping of gene duplications in these genomes. Additionally, many species, such as *D. pseudoobscura*, harbor inversion polymorphisms (Krimbas and Powell 2000), which may prevent the spread of duplicated genes between different chromosomal arrangements within a single species (Popadic et al. 1995).

I present an analysis of recently duplicated genes in the *D. pseudoobscura* genome using a dataset containing the duplications of complete open reading frames and those that appear to have been pseudogenized. I also have information regarding the structure of gene duplications, as well as high confidence assignments of ancestral versus derived copies. Paralogs with low divergence between copies were interrogated for CNP. My results indicate that nucleotide divergence between paralogs is not an adequate proxy for the age of very recent duplications because of gene conversion between paralogs. I am able to classify duplication events by their mechanism, and I consider the role that the duplication mechanisms play in the selective constraints on duplicated genes. Additionally, I examine how the relative positions of paralogs, size of duplications, nucleotide divergence between paralogs, and functional properties of the genes influence the likelihood of pseudogenization.

Methods

Identifying, Annotating, and Aligning Duplicated Genes

Duplicated protein coding genes were identified in the *D. pseudoobscura* genome by comparisons with the *D. melanogaster* genome (Meisel 2009). In the preliminary annotation of the *D. pseudoobscura* genome, one to one best hit orthologs to *D. melanogaster* protein coding genes were identified (Richards et al. 2005). The sequence covered by the orthologs (both exonic and intronic) was classified as "genic" sequence, and the sequence between the identified genes was classified as "intergenic". If a gene was duplicated in the *D. pseudoobscura* genome after the split with *D. melanogaster*, there should be a second homolog located in one of the sequences initially categorized as intergenic. Genes may also be duplicated into the introns of other genes, and my approach will miss these duplications.

To identify genes duplicated along the *D. pseudoobscura* lineage, the *D. pseudoobscura* genic regions were searched against the intergenic regions using MegaBLAST (Zhang et al. 2000), after masking for all known *Drosophila* transposable elements and a common repetitive sequence found throughout the *D. pseudoobscura* genome (Richards et al. 2005) (Genbank accessions AY693425 and AY693426) using RepeatMasker (Smit et al. 2004). The *D. pseudoobscura* intergenic regions were also searched against all *D. melanogaster* proteins using BLASTX (Altschul et al. 1997) to mimic the initial search for one to one orthologs. (The parameters used in these BLAST searches are available in the Supplementary Methods.) All intergenic sequences that matched orthologous *D. pseudoobscura* and *D. melanogaster* genes were retained for further analysis. While this approach will not identify duplications of *D. pseudoobscura* genes that were missed by the initial search for one to one orthologs, these types of genes make up a minority of the genic content of the *D. pseudoobscura* genome (*Drosophila* 12 Genomes Consortium 2007). However, duplicated genes with degenerated open reading frames will be identified with this method because it relies solely on sequence identity; this differs from other collections of duplicated genes in

Drosophila genomes which rely on gene prediction algorithms (e.g., Hahn et al. 2007).

The duplication endpoints were identified using the nucleotide BLAST alignments of the intergenic and genic regions. The genome was partitioned into genic and intergenic regions prior to the BLAST search, and alignments would often reach the end of the partitioned regions. To remedy this, flanking partitions were added until the duplication endpoints no longer reached the end of the outermost partitions. As each partition was added, the endpoints were re-identified using BLAST.

The protein coding sequences of the paralogs were annotated using the following approaches. First, the initial published annotation often contained a predicted open reading frame for the one to one best hit ortholog in the *D. pseudoobscura* genome (Richards et al. 2005). Additionally, BLASTX (Altschul et al. 1997) searches of the duplicated sequence against the *D. melanogaster* proteome were used to identify protein coding sequences with homology to the *D. melanogaster* ortholog. Finally, the duplicated sequences were used as queries in Genscan (Burge and Karlin 1997), and the predicted open reading frames were also used in the annotation. The open reading frames identified using these three approaches were used to create a protein coding sequence for each copy with maximal overlap with the *D. melanogaster* gene. Predicted exons were included if they had a homologous exon in the *D. melanogaster* gene. Once the protein coding sequences were inferred, the amino acid sequences of the two *D. pseudoobscura* paralogs were aligned with the *D. melanogaster* ortholog using the CLUSTALW (Thompson et al. 1994) implementation in MEGA 3.1 (Kumar et al. 2004). Finally, the nucleotide sequences of the *D. pseudoobscura* genes were overlaid on the amino acid alignment, and the intergenic and intronic nucleotide sequences were aligned using CLUSTALW. (The alignments between the *D. pseudoobscura* paralogs and the *D. melanogaster* orthologous coding sequence are available as Supplementary Material.)

Draft genome assemblies have many assembly gaps. If an assembly gap was found in a duplicated region, that region was also extracted from the reconciled assembly of the *D. pseudoobscura* genome (http://rana.lbl.gov/drosophila/caf1.html). In some cases, the reconciled assembly also contained an assembly gap or was missing the duplication entirely; in these cases the duplicated gene was removed from the dataset. If the reconciled assembly contained the duplication without assembly gaps, the sequence from the reconciled assembly was annotated as described earlier. The duplications were further confirmed to not be assembly artifacts by MegaBLAST (Zhang et al. 2000) against the trace sequences from the whole genome sequencing project. Duplications with multiple traces spanning the endpoints were consider real, while those that did not have multiple traces spanning their endpoints were excluded. Finally, the ten paralogs with the highest sequence identity were reconstructed as follows to confirm that they are not assembly artifacts. Trace sequences from within each copy were obtained by using the mate-pairs located outside the duplicated region. These traces were assembled and aligned to the genomic sequence used in the analysis.

Estimating Nucleotide Divergence Between Paralogs and the Lengths of Duplicated Regions

Nucleotide divergence between *D. pseudoobscura* paralogs was calculated for all sites aligned between paralogs, non-coding sites, synonymous sites within coding exons (Nei and Gojobori 1986), and non-coding plus synonymous sites. Duplicated genes with frameshift mutations were shifted to be in the proper reading frame for these calculations. Paralogs were excluded from the analysis if they had less than 80% sequence identity at all sites, non-coding sites, synonymous sites, or non-coding plus synonymous sites and if there were at least 100 nucleotides in that particular class of sites (the 100 site cutoff was chosen to prevent spurious results due to a small sample size of nucleotides). The 80% sequence identity cutoff was chosen for four reasons: (1) it ensures that the duplications arose after the split between *D. pseudoobscura* and *D. melanogaster*; because divergence between these species is saturated at synonymous and non-coding sites (Richards et al. 2005), 80% sequence identity is a conservative cutoff. (2) It allows for reliable alignment of non-coding sequences (introns and intergenic regions). (3) Recent duplications are more likely to currently be located at the same genomic location in which they arose, which decreases the confounding effect of secondary relocation in the analysis of divergence between paralogs. (4) The BLAST settings should be able to identify a large majority of paralogs with >80% sequence identity (Gotea et al. 2003). The length of each duplication event was estimated using the number of nucleotide sites in each copy, the average of the lengths of the two copies, and the number of nucleotide sites in the alignment of the two copies (excluding gaps in the alignment). Finally, I excluded genes that had been recently duplicated multiple times to ensure the phylogenetic independence of all genes in the dataset. The excluded genes had more than one paralog with at least 80% sequence identity.

Completeness of Genes and Degeneration of Open Reading Frames

Amino acid alignments between paralogs were used to classify duplicated genes as “complete” or “partial”. If

both copies of a duplicated gene contain the same beginning- and end-points of their annotated protein coding sequence, the duplicated gene was considered complete—*cis* regulatory regions were not considered because they are poorly annotated for *D. pseudoobscura* genes (Richards et al. 2005). Conversely, partially duplicated genes are missing either the 5′ end, the 3′ end, or both ends of the annotated coding sequence. Coding sequences containing a frameshift mutation, premature stop codon, or mutation at a 5′ or 3′ intron splice site were categorized as "degenerated". Degenerating mutations were confirmed with comparisons to the original trace files. All duplicated genes fall into one of four mutually exclusive categories: (1) complete non-degenerated, (2) complete degenerated, (3) partial non-degenerated, and (4) partial degenerated.

Mechanisms of Duplication and Inferring Ancestral and Derived Copies

The intron-exon structure of the duplicated genes, the relative position of the *D. pseudoobscura* paralogs, and the location of the *D. melanogaster* ortholog were used to classify each copy as ancestral or derived. First, duplications of single genes were classified as retroposed if one copy is missing all introns present in the other copy (over the region of the gene that was duplicated). The copy missing introns was considered derived. Intron loss via recombination between processed transcripts and the genes that encode them may also give rise to intronless genes (Coulombe-Huntington and Majewski 2007; Fink 1987). However, independent evidence suggests that the intronless genes in this dataset were generated via retroposition (Meisel 2009). For example, the ancestral and derived copies of all putatively retroposed duplications are found on different chromosome arms, and the ancestral copies are all found on the same arm as their *D. melanogaster* orthologs. Also, there are no ancestral copies that have lost introns, suggesting that intron-loss is a rare event in Drosophila. Duplications of multiple genes were classified as DNA-based duplications. Additionally, duplications of single genes were considered DNA duplications if both copies contain the same intron-exon structure and at least one intron is found in the duplicated region. Duplications of single exon genes and single exons were classified as ambiguous. A single duplicated gene (the orthologs of CG7730) contains two of the three introns present in the ancestral copy; this gene was classified as ambiguously duplicated.

Ancestral and derived copies were inferred for DNA duplications as follows. The location of the *D. melanogaster* copy was inferred to be the ancestral location of the gene prior to duplication. *D. pseudoobscura* paralogs were first classified based on their relative position; they were divided into those in which both paralogs are located on the same chromosome arm (intra-arm) and those in which the paralogs are located on different chromosome arms (inter-arm). A portion of Muller element A (chromosome arm XL) was relocated to the proximal region of chromosome XR (Muller element D) along the *D. pseudoobscura* lineage after the split with the *D. melanogaster* lineage (Schaeffer et al. 2008). My analysis is robust to whether this region is considered part of element A or element D. In the case of inter-arm duplications, the ancestral copy was inferred to be the one located on the same homologous chromosome arm as the *D. melanogaster* ortholog. Paralogs located on the same chromosome arm with no protein coding genes between them were classified as adjacent (regardless of their orientation), and intra-arm duplications with at least one gene between them were classified as non-adjacent. For non-adjacent intra-arm duplications, the ancestral copy was inferred to be the one with the same flanking genes as the *D. melanogaster* ortholog. If neither copy has the same flanking genes as the *D. melanogaster* ortholog, I was able infer ancestral and derived copies only if one copy has a partial coding sequence—the copy with a partial coding sequence was inferred to be the derived copy.

Adjacent duplications were assigned as ancestral and derived based on the completeness of the coding sequence and the conserved orientation with the *D. melanogaster* orthologs. Partial coding sequences may arise via either the partial duplication of a coding sequence or the complete duplication of a coding sequence followed by the subsequent loss of the 5′ or 3′ end of the coding sequence. An adjacent duplication containing the partial coding sequence of a gene was inferred to be the derived copy because that copy is missing a portion of the coding sequence found in the other copy and the *D. melanogaster* ortholog. If adjacent paralogs are in opposite orientations, the copy in the same orientation (relative to flanking genes) as the *D. melanogaster* ortholog was inferred to be the ancestral copy. In the case of adjacent duplications with complete coding sequences and both paralogs in the same orientation, neither copy was assigned as ancestral or derived. This approach is expected to give reliable assignments of ancestral and derived copies because it is independent of most of the evolutionary events that occur subsequent to gene duplication. However, there may be errors in the assignment of ancestral and derived copies for adjacent duplications because either one of the paralogs or the *D. melanogaster* ortholog may have changed its orientation (via a microinversion). Additionally, the assumption that partial duplications are derived may not be appropriate. The error associated with these potential flaws appears to be negligible—treating the ancestral and derived copies as unknown for adjacent duplications

results in the same conclusions as assigning copies as ancestral and derived.

Expression Profiles of *D. melanogaster* Orthologs

Expression data for the *D. melanogaster* orthologs of the *D. pseudoobscura* duplicated genes were downloaded from the FlyAtlas (Chintapalli et al. 2007), which contains expression data from 13 different body parts. Genes with an mRNA signal >100 in a particular body part were said to be expressed in that body part. *G* tests were used to assess whether expression in a particular body part is independent of whether completely duplicated genes are degenerated.

Copy Number Polymorphism

Duplicated genes with derived copies on one of two autosomes—Muller elements C and E (chromosomes 3 and 2, respectively)—were tested to see if the derived copy has fixed in a sample of chromosomes from natural populations. Duplicated genes with the highest sequence identity between paralogs are most likely to be segregating as CNPs. I selected paralogs with high sequence identity for which PCR primers could be designed such that they only amplify a region if the derived copy of a duplicated gene is present (primer sequences are available in Supplementary Table 1). No explicit sequence identity cutoff was used, and the sampled paralogs differ at 0.44–6.13% of non-coding and synonymous sites. Primers were placed in sequence that is conserved between *D. pseudoobscura* and *D. melanogaster* to minimize the possibility of false negatives. For two of the duplications (the orthologs of CG2412 and CG11552) two sets of primers were used to test for CNP.

A sample of 63 lines that had been made isochromosomal for the third chromosomes (Schaeffer et al. 2003) was assayed for CNP on the third chromosome. Each line carries one of six different arrangements differentiated by chromosomal inversions: Arrowhead (AR), Standard (ST), Pikes Peak (PP), Chiricahua (CH), Santa Cruz (SC), and Tree Line (TL). Duplicated genes were identified on the AR background because the strain sequenced in the *D. pseudoobscura* genome project was homozygous for an AR third chromosome (Richards et al. 2005). Only one chromosome was assayed for SC and TL, each, and at least 12 chromosomes were assayed for the other four arrangements. Twelve lines that had been inbred for ten generations using single-pair sib-matings were assayed for duplicated genes on chromosome 2, which harbors no inversion polymorphism. Each of these inbred lines was started from a single wild-caught female. DNA sequencing from loci not discussed here reveals that these lines have very little heterozygosity (data not shown).

Results and Discussion

I identified 88 duplications that occurred in the *D. pseudoobscura* genome after the split with the *D. melanogaster* lineage, containing a total of 101 genes (Supplementary Table 2). Nine duplications contain two genes, two duplications contain three genes, and the rest are duplications of single genes. I analyzed duplications of single genes separately and found similar results as when I analyzed all duplications; only the analysis of all duplications is presented. In the analysis below, nucleotide divergence between paralogs is measured at non-coding and synonymous sites and duplication length by the number of sites in the alignment of the paralogs (excluding gaps), but the results are robust to all measures of nucleotide divergence and duplication length.

On the Path to Fixation of a New Duplication and Gene Conversion Between Paralogs

The path to fixation of a duplicated gene will be influenced by a combination of stochastic and deterministic processes (Lynch et al. 2001). Recently duplicated genes may be segregating as CNPs if the derived copy has yet to fix in the population. Nucleotide divergence between paralogs can be used a proxy for the age a duplication. However, gene conversion between paralogs (Drouin 2002; Lazzaro and Clark 2001; Osada and Innan 2008; Petes and Fink 1982; Semple and Wolfe 1999; Slightom et al. 1980; Thornton and Long 2005) may slow the rate of divergence between the two copies, resulting in an underestimate of the age of the duplication (Teshima and Innan 2004). The data on divergence between paralogs and CNP can be used to indirectly infer the effect of gene conversion on the sequence divergence between paralogs.

Copy Number Polymorphism

Duplicated genes with minimal divergence between copies have the highest likelihood of segregating as CNPs. I used PCR to determine if the derived copy is present for duplicated genes with low divergence between paralogs. Derived copies on chromosomes 2 and 3 (Muller elements E and C, respectively) were interrogated. While there are technical limitations with using PCR to infer CNP, these are more likely to lead to false negatives, rather than false positives. Therefore, in the analysis of CNP below, I focus primarily on what the presence of derived copies in the sampled chromosomes reveals about the evolutionary dynamics of the duplicated genes.

The *D. pseudoobscura* third chromosome harbors a rich inversion polymorphism, with over 30 different

arrangements segregating in natural populations (Powell 1992). Six different arrangements of chromosome 3 were sampled, although two of the arrangements only had one representative in the sampled chromosomes. The AR arrangement was sequenced in the *D. pseudoobscura* genome project (Richards et al. 2005). Therefore, we would expect any genes segregating as CNPs to be at the highest frequency on the AR arrangement because this is the arrangement upon which we identified the duplicated genes. Indeed, all four duplicated genes assayed on the third chromosome are fixed on the AR arrangement (Table 1). Some duplicated genes are fixed or near fixation on other backgrounds as well. This suggests that the inversions do not prevent the spread of duplicated genes between arrangements. Interestingly, the duplicated gene with the lowest amount of divergence between paralogs is the only one found on all third chromosomes sampled.

Five duplicated genes were assayed on the second chromosome, which harbors no inversion polymorphism. One duplicated gene on chromosome 2 was not found in the sampled chromosomes despite the ability of the PCR primers to amplify the derived copy from the genome strain. This paralog had the lowest nucleotide divergence between copies of all the duplications sampled for CNP, suggesting it may be the most recent duplication. The other four derived copies are fixed in the sampled chromosomes, despite very little nucleotide divergence between paralogs (Table 1). This suggests that the fixation of recently duplicated genes occurs rapidly. Alternatively, gene conversion between paralogs may slow the rate of divergence between paralogs, causing nucleotide divergence to be a poor indicator of the age of the duplications. The evidence for gene conversion between paralogs is examined below.

Table 1 Copy number polymorphism frequency of duplicated genes on two chromosomes

Gene(s)	Len[b]	Div[c]	Frequency[a]				
			Total	AR (15)	ST (20)	PP (14)	CH (12)
Chromosome 3—Muller element C (n = 63)							
CG8589	5109	0.0188	1.000	1.000	1.000	1.000	1.000
capt (CG5061)[d]	181	0.0473	0.698	1.000	0.750	0.286	0.667
CG5969	817	0.0555	0.714	1.000	0.850	0.071	0.833
Rad51C (CG2412)[d]	1533	0.0613	0.952	1.000	0.950	0.857	1.000
Chromosome 2—Muller element E (n = 12)							
CG11552	826	0.0044	0.000				
CG16734	4658	0.0060	1.000				
Hsp68 (CG5436)	2429	0.0073	1.000				
Hsp70B	2488	0.0123	1.000				
CG7262, CG14860	4926	0.0458	1.000				

[a] Frequency of the derived copy in the sampled chromosomes. Additionally, the CNP frequency was determined for four different arrangements of Muller element C: Arrowhead (AR), Standard (ST), Pikes Peak (PP) and Chiricahua (CH). Sample size for these arrangements are in parentheses

[b] Length of duplication measured by total alignable sites between paralogs

[c] Nucleotide divergence between paralogs measured at non-coding and synonymous sites

[d] Derived copy has partial or degenerated coding sequence

The Effect of Relative Position on the Divergence Between Paralogs

High sequence identity between paralogs may indicate a recent origin of the duplicated gene, gene conversion between paralogs, or selective constraint on the two copies. Intra-arm duplications have higher nucleotide sequence identity between paralogs than inter-arm duplications, and adjacent duplications are less diverged than non-adjacent duplications ($H = 13.34$, $P < 0.005$) (Fig. 1a). There is not a significant difference in the divergence between paralogs when one compares adjacent duplications in the same orientation with those in opposite orientations ($P = 0.81$, Wilcoxon test). One possible explanation for this pattern is that gene conversion between paralogs is

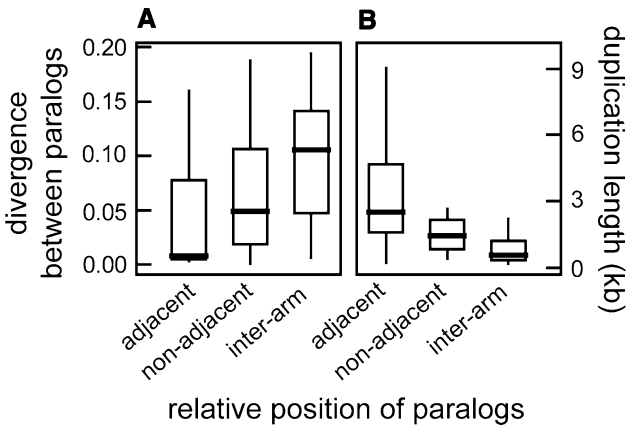

Fig. 1 Effects of relative position of paralogs on divergence between paralogs and lengths of duplicated regions. Box and whisker plots of **a** divergence between paralogs and **b** lengths of duplicated regions for adjacent, non-adjacent, and inter-arm duplications are presented. Boxes extend from the first quartile to the third quartile, with the median indicated by the *horizontal line* within the box. Whiskers extend to the smallest and largest non-outlier values, and outliers were not plotted. **a** Divergence is measured at non-coding and synonymous sites, and only paralogs with at least 100 non-coding and synonymous sites in their alignments are used in the plot of divergence between paralogs. **b** The lengths of the duplications are measured by the number of nucleotide sites in the alignments of the paralogs

more frequent for more proximally located duplications (Benovoy and Drouin 2009; Drouin 2002; Katju and Lynch 2003; Semple and Wolfe 1999), although experiments in yeast do not support this hypothesis (Haber et al. 1991). Unfortunately, these data do not allow for accurate identification of individual gene conversion events because of high sequence identity between paralogs. Also, the effect of gene conversion between paralogs should diminish as the two copies diverge (Teshima and Innan 2004). Therefore, nucleotide sequence divergence may be an adequate proxy for the age of a duplication when paralogs are well beyond the threshold where gene conversion may occur.

Another possible explanation for the low nucleotide sequence divergence between proximally located paralogs is that the more distantly located the two copies are, the greater the genetic independence between the paralogs. That is, linkage disequilibrium between paralogs located on different chromosome arms can be decreased via independent assortment, and associations between intra-arm paralogs can be broken via crossing over. This will make paralogs with less genetic linkage appear more diverged under the following scenario. The derived copy of a duplicated gene originates from a single allele of the ancestral locus, which had segregating polymorphisms at the time of the duplication event. Initially, the derived copy will have more sequence identity with that allele at the ancestral locus than to the other alleles at the ancestral locus. More tightly linked paralogs have a higher probability of the derived copy being sampled along with a descendant of the ancestral allele from which it arose, causing measures of divergence between genetically linked paralogs to be artificially lower than unlinked paralogs. Unfortunately, it is not possible to determine whether differences in rates of gene conversion or differences in the genetic independence of paralogs cause the relationship between divergence and relative positions.

The other explanations for the relationship between relative position and nucleotide divergence between paralogs are not as convincing. For example, derived copies may be located near the ancestral copy when they arise and disperse throughout the genome over time—a hypothesis that has been previously considered, and rejected, in *Caenorhabditis elegans* (Katju and Lynch 2003). This is also unlikely for the duplicated genes in the *D. pseudoobscura* genome because most distantly located paralogs contain no hallmarks of relocation; the majority of the non-adjacent and inter-arm duplications lie within regions of conserved gene order between *D. pseudoobscura* and *D. melanogaster*, so they could not have moved away via a simple rearrangement event. It is more parsimonious to assume that the current location of the derived copy is an adequate approximation of its location when it was generated. Codon usage bias has also been shown to slow the rate of divergence between paralogs (Lin et al. 2006). However, the effect of relative position on the divergence between paralogs is also observed when only intergenic and intronic sequences are examined. Therefore, the result is unlikely to be affected by codon bias. Furthermore, it has been suggested that proximal duplications in *Drosophila* are usually pseudogenized or lost because they are unlikely to gain a new function, while dispersed duplications are more likely to be retained because they can evolve a beneficial function (Zhou et al. 2008). Therefore, one expects adjacent duplications to be of a more recent origin than non-adjacent duplications. However, this model also predicts a higher rate of pseudogenization for adjacent duplications, which is not observed in the *D. pseudoobscura* genome (see below).

A negative correlation between nucleotide divergence between paralogs and lengths of duplications has been previously described for *D. melanogaster* gene duplications (Osada and Innan 2008). A similar correlation can be recovered by excluding the longest duplication in this dataset ($r = -0.319$, $P < 0.01$) (Fig. 2). Osada and Innan (2008) argued that the endpoints of duplication blocks decay faster than the central region of duplications because gene conversion rates are higher in the central region. However, analyses of variance (ANOVA) reveal a significant effect of the relative position of the paralogs (adjacent, non-adjacent, or inter-arm) on both divergence ($F = 6.74$, $P < 0.005$) and length ($F = 7.70$, $P < 0.005$), but no significant effect of divergence on length or length on divergence (divergence was arcsine transformed).

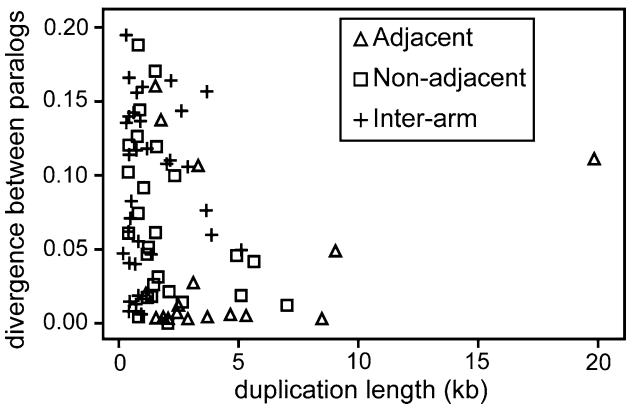

Fig. 2 Relationship of divergence between paralogs and lengths duplicated regions. The divergence between paralogs are plotted against the length of the duplicated region for each duplication. Adjacent duplications are indicated by *triangles*, non-adjacent intra-arm duplications are indicated by *squares*, and inter-arm duplications are indicated by a "+". Divergence is measured at non-coding and synonymous sites, and only paralogs with >100 non-coding and synonymous sites in their alignment are included. The lengths of the duplications are measured using the total number of sites in the alignment of the paralogs

Therefore, the correlation between divergence and length is the result of both divergence and length being affected by the relative position of the paralogs. Finally, the effect of relative position on the nucleotide divergence between paralogs may be the result of low gene conversion rates between the paralogs of retroposed genes (because of missing introns), as the retroposed genes are all located on different chromosome arms that their ancestral copies. Excluding retroposed genes from the ANOVA yields the same results as above, suggesting that relative position is the main factoring influencing the amount of sequence divergence between paralogs.

On the Likelihood of Pseudogenization

The Effect of Complete/Partial Coding Sequences

Although pseudogenes have been extensively studied in *Drosophila* genomes (e.g., Harrison et al. 2003; Petrov and Hartl 2000), the processes governing why particular duplicated genes become pseudogenes while others are retained are still unclear (Lynch et al. 2001; Zhang 2003). One hallmark of a pseudogenized gene is a nonsense, frameshift, or intron-splice-site mutation (Harrison and Gerstein 2002), and I refer to genes harboring at least one of these mutations as "degenerated". Two different ancestral copies of duplicated genes have accumulated mutations that disrupt their open reading frames (Supplementary Table 2). These degenerating mutations presumably occurred after the duplication event because the mutations are not shared by the derived copy. There is no difference in the conclusions of the analysis presented below if those two duplicated genes are treated as degenerated or not degenerated.

One expects that a completely duplicated coding sequence would be under more selective constraints than a partially duplicated gene. Indeed, the derived copies of partially duplicated genes are more likely to have degenerated open reading frames than completely duplicated genes ($P = 0.0003$, F.E.T.) (Table 2). Additionally, if both copies of a duplicated gene are under similar selective constraints, they should be evolving at approximately equal rates in their functional regions (Lynch and Katju 2004). Amino acid substitutions were polarized along the lineages leading to the ancestral and derived copies of all duplicated genes, using the *D. melanogaster* orthologs as outgroups. A relative rate test based on the χ^2 test (Tajima 1993) was used to determine if the number of amino acid substitutions differ between ancestral and derived copies (Supplementary Fig. S1). The test-statistic of the relative rate test can be used as a measure of the differences in rates of evolution between the two lineages. Gene pairs for which the derived copy has a partial coding sequence have more asymmetrical rates of amino acid evolution than completely duplicated genes (Fig. 3a). This is because the derived copies of many partial duplicates have accumulated an excess of amino acid substitutions relative to their ancestral paralogs (Supplementary Fig. S1). These results indicate that the derived copies of completely duplicated genes are under more selective constraints than partial duplicates. While this is unsurprising, it is interesting that the majority of completely duplicated genes have not degenerated.

Table 2 Counts of degenerated derived copies of duplicated genes for various classes of duplicated genes

	Derived copy degenerated?	
	No	Yes
All duplicated genes		
Complete	34	10
Partial	23	34
Completely duplicated genes		
Amb & retr[a]	12	5
DNA dup[b]	22	5
Adjacent	11	2
Non-adjacent	13	3
Inter-arm	9	6

[a] Ambiguous and retroposed duplications

[b] DNA-based duplications

It is possible that partially duplicated genes arose as complete genes and subsequently lost the 5′ or 3′ end of the duplicated region. If the loss of one end of a duplicated gene occurs via the fixation of neutral mutations, we would expect the probability that a complete duplication becomes a partial duplicate to increase with time. However, there is not a significant difference in nucleotide divergence between paralogs when one compares complete and partial duplications (Fig. 3b). To control for the selective constraints on non-degenerated genes, I looked at degenerated genes alone. There is not a significant difference in nucleotide divergence between partial and complete degenerated genes ($P = 0.33$, Wilcoxon test). While a small number of duplications that were classified as partial may have originally arisen as complete duplications (and subsequently lost their 5′ or 3′ end), this does not appear to be a common occurrence.

I further examined the selective constraints on completely duplicated genes. First, I compared the rates of amino acid evolution between completely duplicated genes in which the derived copied had not degenerated and those with degenerated derived copies. The two copies of completely duplicated genes for which one copy has degenerated evolve at more asymmetrical rates than duplicated genes that had not degenerated (Fig. 3a). The increased

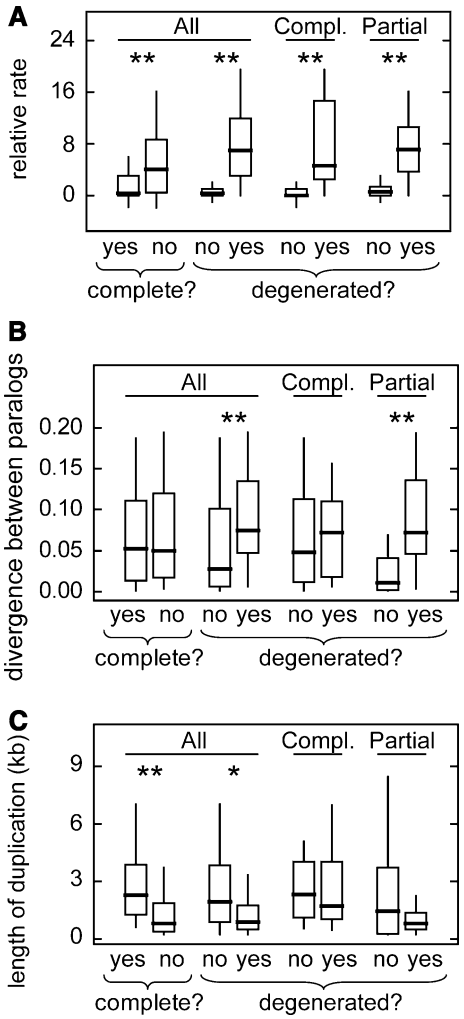

Fig. 3 The effects of completeness and degeneration on rates of evolution, nucleotide divergence, and duplication length. Graphs compare **a** relative rate of evolution between paralogs, **b** nucleotide divergence between paralogs, and **c** the length of the duplicated region. Comparisons are between: (1) completely and partially duplicated genes, (2) duplicated genes in which the derived copy has not degenerated and those in which the derived copy has degenerated, (3) completely duplicated genes that have not degenerated and those that have, and (4) partially duplicated genes that have not degenerated and those that have. Boxes extend from the first quartile to the third quartile, with the median indicated by the *horizontal line* within the box. Whiskers extend to the smallest and largest non-outlier values, and outliers were not plotted. Comparisons for which there is a significant difference at $P < 0.05$ using a Kruskal–Wallis test are indicated by a single *asterisk*, and those for which $P < 0.005$ are indicated by two *asterisks*. **a** "Relative rate" is the chi-square test statistic (Tajima 1993) of the difference in amino acid substitutions between the derived lineage and the ancestral lineage for each paralogous pair, with negative values indicating faster rates on the ancestral lineage. **b** Plots are shown for the divergence between paralogs using only non-coding and synonymous sites for paralogs with at least 100 sites in those classes in the alignment between paralogs. **c** Length estimates are the total number of nucleotides in the alignment of the paralogs

rate asymmetry is the result of faster amino acid evolution along the lineages leading to degenerated copies (Supplementary Fig. S1). This indicates that non-degenerated completely duplicated genes are under more selective constraints than degenerated complete genes—another unsurprising result. However, degenerated partially duplicated genes also have more asymmetrical rates of amino acid evolution than non-degenerated partial duplicates (Fig. 3a). Furthermore, no significant differences in the rates of amino acid evolution are detected between the two copies of non-degenerated partially duplicated genes (Supplementary Fig. S1). This suggests that there are more selective constraints on non-degenerated partially duplicated genes than on degenerated partial genes. But there is also less nucleotide divergence between paralogs for non-degenerated partially duplicated genes than degenerated partial genes (Fig. 3b). Therefore, the failure to detect asymmetrical rates of amino acid evolution between the two copies of non-degenerated partially duplicated genes is a by-product of their recent origin—not enough time has passed for them to accumulate amino acid differences. The same phenomenon cannot explain the differences in relative rates of amino acid evolution between degenerated and non-degenerated completely duplicated genes; degenerated and non-degenerated completely duplicated genes do not have significantly different amounts of nucleotide divergence between paralogs (Fig. 3b).

The Role Other Factors Play in Pseudogenization

The effects of the length of a duplication event, the ancestral expression profile, the duplication mechanism, and the relative position of paralogs on the pseudogenization of duplicated genes were examined. I find that none of these factors significantly affect the likelihood of pseudogenization for completely duplicated genes. The size of a duplication event may influence whether or not the complete coding sequence of a gene gets duplicated (Katju and Lynch 2003). Indeed, completely duplicated genes are contained in significantly longer duplication blocks than partially duplicated genes (Fig. 3c). Additionally, non-degenerated genes are also contained in longer duplications than degenerated genes (Fig. 3c). This suggests that the length of a duplication event influences the likelihood that the gene contained within it will degenerate. Furthermore, adjacent duplications are longer than non-adjacent duplications, and intra-arm duplications are longer than inter-arm duplications ($H = 21.67$, $P < 0.001$) (Fig. 1b). If duplication length influences the selective constraints on

completely duplicated genes, then the relative position of a duplication may influence the selective constraints on the coding sequence as well. However, there is not a significant difference in duplication length between non-degenerated and degenerated duplicated genes when one looks at completely or partially duplicated genes separately (Fig. 3c). Non-degenerated derived copies are contained in longer duplications than degenerated genes because longer duplications have a greater likelihood of containing completely duplicated genes, and complete genes are less likely to be degenerated.

It has previously been observed that different functional classes of genes are preferentially retained following duplication, while other classes are underrepresented as multigene families (Kondrashov et al. 2002; Papp et al. 2003; Seoighe and Gehring 2004). I was unable to examine the prevalence of different gene ontology classes among degenerated and non-degenerated completely duplicated genes because of small sample sizes. Additionally, the only whole genome expression data available for *D. pseudoobscura* are from whole males and females (Zhang et al. 2007), and there is no evidence that sex-biased expression predicts degeneration (data not shown). As a proxy for the expression profile of the ancestral copy, data from 13 different body parts were obtained for the *D. melanogaster* orthologs of each completely duplicated gene (Chintapalli et al. 2007). Completely duplicated genes with orthologs expressed in the hindgut, midgut, and brain have a higher frequency of degeneration than completely duplicated genes not expressed in those tissues (Supplementary Fig. S2). A sequential Bonferroni correction for multiple tests (Sokal and Rohlf 1995), however, causes one to fail to reject the null hypothesis of independence between degeneration and tissue expression. It is also possible that the presence of *cis* regulatory sequences at the derived locus (either arriving with a DNA-based duplicated gene or acquired from sequences flanking a duplicated gene) influences whether a duplicated gene becomes a pseudogene. This could be studied by examining the expression profiles of recently duplicated genes in future experiments, but these experiments would be beyond the scope of the results presented here.

The process giving rise to a completely duplicated gene does not influence whether the derived copy has degenerated either. There is an insignificant excess of non-degenerated DNA-based completely duplicated genes compared to ambiguous and retroposed complete duplications ($P = 0.32$, F.E.T.) (Table 2). Additionally, the relative position of the paralogs does not significantly affect the probability of pseudogenization of completely duplicated genes; intra-arm duplications have an insignificantly lower frequency of degenerated derived copies than inter-arm duplications ($P = 0.10$, F.E.T.) (Table 2). In rodents, duplications in which the ancestral and derived copies are distantly located from each other evolve at more asymmetrical rates than duplications located in close proximity (Cusack and Wolfe 2007); this may be the result of relaxed constraints on relocated duplications. In *D. pseudoobscura*, however, there is not a significant difference in relative rates of amino acid evolution between complete non-degenerated duplications in different relative positions ($H = 2.34$, $P = 0.310$), nor is there a significant difference in the ratio of non-synonymous to synonymous substitutions between paralogs for intra- and inter-arm completely duplicated genes.

While none of the aforementioned factors alone significantly affect the probability that a completely duplicated gene will become a pseudogene, their interactions may. DNA duplications and intra-arm duplications produce longer duplication tracts, and longer duplications are more likely to contain a complete gene. Therefore, the duplication mechanism and relative position of paralogs seem to indirectly affect the selective retention of duplicated genes. However, DNA-based duplications are more likely to be on the same chromosome arm as the ancestral copy (Meisel 2009), so the mechanism of duplication and relative position are not independent.

The Neutrality of Pseudogenization

If none of the aforementioned factors contribute significantly to the likelihood that a completely duplicated gene will degenerate, it is possible that the degeneration of completely duplicated genes is primarily a neutral process. Using data from *D. melanogaster*, it was previously estimated that half of all duplicated genes are lost by the time the paralogs have diverged at 8.5% of their synonymous sites (Lynch and Conery 2000). This approach toward estimating a half-life for duplicated genes assumes that the probability of degeneration for duplicated genes that arrive with functional potential (i.e., completely duplicated genes) will increase with time. Contrary to this assumption, degenerated completely duplicated genes do not have significantly more nucleotide divergence between paralogs than non-degenerated completely duplicated genes (Fig. 3b). This suggests that a large fraction of completely duplicated genes are under enough selective constraints to prevent degeneration by mutational pressure alone. Therefore, there may not be an appropriate method for estimating the half-life of duplicated genes in Drosophila; the probability and frequency of pseudogenization may depend on specific properties of individual duplicated genes rather than on the nucleotide divergence between paralogs. Unfortunately, the attempts to identify those properties for the duplicated genes in this dataset were not fruitful.

The Possibility of Chimeric Genes

While partially duplicated genes are most likely pseudogenes, it is also possible that they make up part of a functional chimeric gene (Arguello et al. 2006, 2007; Long et al. 2003). It is unlikely that a considerable fraction of the partial duplications in the *D. pseudoobscura* genome are part of functional chimeric genes for three reasons. First, previously described chimeric genes have intact open reading frames over the portion of the gene that was duplicated (Arguello et al. 2006; Jones and Begun 2005; Long and Langley 1993; Nozawa et al. 2005; Yang et al. 2008), whereas the partially duplicated genes in this dataset tend to have degenerated open reading frames (Table 2). Second, even though unequal rates of evolution between ancestral and derived copies is also a hallmark of chimeric genes (Jones and Begun 2005), the partially duplicated genes presented here with unequal rates of amino acid evolution also have degenerated open reading frames (Supplementary Fig. S1). Third, partially duplicated genes with degenerated coding sequences are more diverged from their ancestral paralogs at the nucleotide sequence level than those with intact open reading frames (Fig. 3b). This indicates that partially duplicated genes with degenerated open reading frames are older than those with non-degenerated open reading frames. Therefore, partially duplicated genes probably degenerate via a neutral process because the probability of degenerating increases with time—in contrast to completely duplicated genes for which there is no relationship between degeneration and nucleotide divergence (Fig. 3b). That is not to say that chimeric genes are not evolutionarily important; however, it is unlikely that many of the partially duplicated genes in this dataset make up part of functional chimeric genes.

Acknowledgements N. Hasan, B. B. Hilldorfer, R. LeGros, and R. L. Zindren helped with sorting the BLAST hits, and N. Hasan and B. B. Hilldorfer assisted in testing the recently duplicated genes for CNP. S. W. Schaeffer and J. R. Arguello provided useful discussion and comments on the manuscript. V. Gotea and W. Makalowski provided assistance with RepeatMasker and MegaBLAST, and V. Gotea also commented on the manuscript. This material is partially based on work supported by the National Science Foundation under Grant No. 0608186, awarded to RPM. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers Y-HC, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor Miklos GL, Abril JF, Agbayani A, An H-J, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, Bd Pablos, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei M-H, Ibegwam C et al (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389–3402

Arguello JR, Chen Y, Yang S, Wang W, Long M (2006) Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. PLoS Genet 2:e77

Arguello JR, Fan C, Wang W, Long M (2007) Origination of chimeric genes through DNA-level recombination. Genome Dyn 3:131–146

Benovoy D, Drouin G (2009) Ectopic gene conversions in the human genome. Genomics 93:27–32

Bhutkar A, Russo SM, Smith TF, Gelbart WM (2007) Genome-scale analysis of positionally relocated genes. Genome Res 17:1880–1887

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

Byrne KP, Wolfe KH (2007) Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. Genetics 175:1341–1350

Chintapalli VR, Wang J, Dow JAT (2007) Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet 39:715–720

Coulombe-Huntington J, Majewski J (2007) Characterization of intron loss events in mammals. Genome Res 17:23–32

Cusack BP, Wolfe KH (2007) Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol Biol Evol 24:679–686

Demuth JP, Bie TD, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. PLoS ONE 1:e85

Dopman EB, Hartl DL (2007) A portrait of copy-number polymorphism in *Drosophila melanogaster*. Proc Natl Acad Sci USA 104:19920–19925

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218

Drouin G (2002) Characterization of the gene conversions between the multigene family members of the yeast genome. J Mol Evol 55:14–23

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. Science 320:1629–1631

Fink GR (1987) Pseudogenes in yeast? Cell 49:5–6

Force A, Lynch M, Pickett FB, Amores A, Yan Y-l, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Gotea V, Veeramachaneni V, Makalowski W (2003) Mastering seeds for genomic size nucleotide BLAST searches. Nucl Acids Res 31:6935–6941

Haber JE, Leung WY, Borts RH, Lichten M (1991) The frequency of meiotic recombination in yeast is independent of the number and position of homologous donor sequences: implications for chromosome pairing. Proc Natl Acad Sci USA 88:1120–1124

Hahn MW, Han MV, Han S-G (2007) Gene family evolution across 12 *Drosophila* genomes. PLoS Genet 3:e197

Harrison PM, Gerstein M (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol 318:1155–1174

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

He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169:1157–1164

Heger A, Ponting CP (2007) Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. Genome Res 17:1837–1849

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci 256:119–124

Jones CD, Begun DJ (2005) Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci USA 102:11373–11378

Katju V, Lynch M (2003) The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics 165:1793–1803

Kondrashov F, Rogozin I, Wolf Y, Koonin E (2002) Selection in the evolution of gene duplications. Genome Biol 3:0008.1–0008.9

Krimbas C, Powell J (2000) Inversion polymorphisms in *Drosophila*. In: Singh RS, Krimbas CB (eds) Evolutionary genetics: from molecules to morphology. Cambridge University Press, Cambridge, pp 284–299

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150–163

Lazzaro BP, Clark AG (2001) Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the attacin genes of *Drosophila melanogaster*. Genetics 159:659–671

Lin Y-S, Byrnes JK, Hwang J-K, Li W-H (2006) Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. Proc Natl Acad Sci USA 103:14412–14416

Long M, Langley CH (1993) Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science 260:91–95

Long M, Thornton K (2001) Gene duplication and evolution. Science 293:1551a

Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865–875

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet 20:544–549

Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. Genetics 159:1789–1804

Meisel RP (2009) Repeat mediated gene duplication in the *Drosophila pseudoobscura* genome. Gene 438:1–7

Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol 8:122–128

Muller HJ (1940) Bearings of the 'Drosophila' work on systematics. In: Huxley J (ed) The new systematics. Clarendon Press, Oxford, pp 185–268

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Nozawa M, Nei M (2007) Evolutionary dynamics of olfactory receptor genes in *Drosophila* species. Proc Natl Acad Sci USA 104:7122–7127

Nozawa M, Aotsuka T, Tamura K (2005) A novel chimeric gene, *siren*, with retroposed promoter sequence in the *Drosophila bipectinata* complex. Genetics 171:1719–1727

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York

Osada N, Innan H (2008) Duplication and gene conversion in the *Drosophila melanogaster* genome. PLoS Genet 4:e1000305

Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197

Petes TD, Fink GR (1982) Gene conversion between repeated genes. Nature 300:216–217

Petrov D, Hartl D (2000) Pseudogene evolution and natural selection for a compact genome. J Hered 91:221–227

Popadic A, Popadic D, Anderson W (1995) Interchromosomal exchange of genetic information between gene arrangements on the third chromosome of *Drosophila pseudoobscura*. Mol Biol Evol 12:938–943

Powell JR (1992) Inversion polymorphisms in *Drosophila pseudoobscura* and *Drosophila persimilis*. In: Krimbas CB, Powell JR (eds) *Drosophila* inversion polymorphism. CRC Press, Boca Raton, pp 73–126

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, Couronne O, Hua S, Smith MA, Zhang P, Liu J, Bussemaker HJ, van Batenburg MF, Howells SL, Scherer SE, Sodergren E, Matthews BB, Crosby MA, Schroeder AJ, Ortiz-Barrientos D, Rives CM, Metzker ML, Muzny DM, Scott G, Steffen D, Wheeler DA, Worley KC, Havlak P, Durbin KJ, Egan A, Gill R, Hume J, Morgan MB, Miner G, Hamilton C, Huang Y, Waldron L, Verduzco D, Clerc-Blankenburg KP, Dubchak I, Noor MAF, Anderson W, White KP, Clark AG, Schaeffer SW, Gelbart W, Weinstock GM, Gibbs RA (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and *cis*-element evolution. Genome Res 15:1–18

Schaeffer SW, Goetting-Minesky MP, Kovacevic M, Peoples JR, Graybill JL, Miller JM, Kim K, Nelson JG, Anderson WW (2003) Evolutionary genomics of inversions in *Drosophila pseudoobscura*: evidence for epistasis. Proc Natl Acad Sci USA 100:8319–8324

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguade M, Anderson WW, Edwards K, Garcia ACL, Goodman J, Hartigan J, Kataoka E, Lapoint RT, Lozovsky ER, Machado CA, Noor MAF, Papaceit M, Reed LK, Richards S, Rieger TT, Russo SM, Sato H, Segarra C, Smith DR, Smith TF, Strelets V, Tobari YN, Tomimura Y, Wasserman M, Watts T, Wilson R, Yoshida K, Markow TA, Gelbart WM, Kaufman TC (2008) Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics 179:1601–1655

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528

Semple C, Wolfe KH (1999) Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J Mol Evol 48:555–564

Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends Genet 20:461–464

Sidow A (1996) Gen(om)e duplications in the evolution of early vertebrates. Curr Opin Genet Dev 6:715–722

Slightom JL, Blechl AE, Smithies O (1980) Human fetal $^{G}\gamma$- and $^{A}\gamma$-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21:627–638

Smit AFA, Hubley R, Green P (2004) RepeatMasker Open-3.0

Sokal RR, Rohlf FJ (1995) Biometry. W.H. Freeman and Co., New York

Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599–607

Teshima KM, Innan H (2004) The effect of gene conversion on the divergence between duplicated genes. Genetics 166:1553–1560

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Thornton K, Long M (2005) Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. Mol Biol Evol 22:273–284

Turner TL, Levine MT, Eckert ML, Begun DJ (2008) Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. Genetics 179:455–473

Wang Y, Gu X (2001) Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics 158:1311–1320

Yang S, Arguello JR, Li X, Ding Y, Zhou Q, Chen Y, Zhang Y, Zhao R, Brunet F, Peng L, Long M, Wang W (2008) Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. PLoS Genet 4:e3

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214

Zhang Y, Sturgill D, Parisi M, Kumar S, Oliver B (2007) Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. Nature 450:233–237

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