Gene 438 (2009) 1–7

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Repeat mediated gene duplication in the *Drosophila pseudoobscura* genome

Richard P. Meisel *

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

ARTICLE INFO

Article history:
Received 2 January 2009
Received in revised form 16 February 2009
Accepted 23 February 2009
Available online 9 March 2009

Received by E. Eichler

Keywords:
Retrotransposition
Chromosomal inversion
Double strand break repair
Non-homologous end joining
Synthesis dependent strand annealing
Non-allelic homologous recombination

ABSTRACT

Genetic mutations can occur on a wide variety of scales, including those that change single nucleotides, those that add or remove content to/from a genome, and those that change the organization of a genome. Gene duplications are a specific class of mutations that add content to a genome, and they can arise via a wide variety of mechanisms. I examined the mechanisms responsible for recently duplicated genes in the *D. pseudoobscura* genome, and I observed both retroposed and DNA duplications. Many duplicated genes lack signatures of either retroposition or DNA-based mechanisms, but other features of these ambiguously duplicated genes suggest that most were generated via retroposition. Furthermore, close examination of sequences flanking DNA duplications and those found at the breakpoints of chromosomal inversions suggests a connection between these two events. In *Drosophila*, duplicated genes near inversion breakpoints can arise via unequal genetic exchange during the non-allelic crossing over event giving rise to the inversion. I observed one duplicated gene in the *D. pseudoobscura* genome that appears to have been generated by this mechanism. Additionally, many DNA duplications in the *D. pseudoobscura* genome are flanked by a repetitive sequence also found at the breakpoints of chromosomal inversions. This suggests that the molecular mechanisms responsible for chromosomal rearrangements and some duplicated genes have overlapping processes.

1. Introduction

Gene duplication has long been recognized as an important evolutionary process (Ohno, 1970). In *Drosophila* genomes, genes can be duplicated via retroposition and DNA-based mechanisms (Bai et al., 2007; Fiston-Lavier et al., 2007; Yang et al., 2008; Zhou et al., 2008); although polyploidization is also common in eukaryotes (Otto, 2007), it is not discussed here because it is not a source of duplicated genes in *Drosophila* genomes. Retroposition occurs when messenger RNA (mRNA) from a protein coding gene is reverse transcribed into DNA by an enzyme encoded by an endogenous retroelement, and the reverse transcript is inserted into the genome (Esnault et al., 2000; Kaessmann et al., 2009). Adjacent DNA duplications may occur by unequal crossing over of repetitive sequences flanking the duplicated region (Zhang, 2003) or non-homologous end joining (NHEJ) (Zhou et al., 2008). Other DNA-based duplications also appear to be driven by repeat mediated processes, such as non-allelic homologous recombination (NAHR) between transposable elements (TEs) (Bailey et al., 2003; Fiston-Lavier et al., 2007; Yang et al., 2008). Multiple studies have reported duplicated genes at the boundaries of inversion breakpoints in *Drosophila* (Matzkin et al., 2005; Ranz et al., 2007). These duplications are thought to arise via unequal genetic exchange during the crossing over event that gives rise to an inversion. The crossing over events may be driven by the repair of a double-strand-break (DSB) by NAHR (Engels and Preston, 1984; Matzkin et al., 2005) or by staggered single-strand-breaks (SSB) which are repaired using NHEJ (Kehrer-Sawatzki et al., 2005; Ranz et al., 2007).

I present an analysis of the mechanisms responsible for generating recently duplicated genes in the *D. pseudoobscura* genome using data rich in information regarding the structure of gene duplications. I find evidence for both retroposed and DNA duplications. Additionally, I observe a single example of a gene duplication within an inversion breakpoint region. While duplications along with inversions appear to be rare, many DNA duplications are flanked by a repeat sequence also found within inversion breakpoint regions. This suggests that the initial steps responsible for DNA duplications and chromosomal inversions are similar.

2. Materials and methods

2.1. Identifying, aligning, and annotating duplicated genes

I identified genes that had been duplicated in the *D. pseudoobscura* genome after the divergence from the *D. melanogaster* lineage. These two species' lineages diverged approximately 55 million years ago (Tamura et al., 2004), and synonymous and non-coding sites are saturated between the two genomes (Richards et al., 2005). One to

Abbreviations: FET, Fisher's Exact Test; *TppII*, tripeptidyl-peptidase II; *TppIIψ*, tripeptidyl-peptidase II pseudogene.

* Present address: Department of Molecular Biology and Genetics, Cornell University, 227 Biotechnology Bldg., Ithaca, NY 14853, USA. Tel.: +1 607 255 1707; fax: +1 607 255 6249.
E-mail address: meisel@cornell.edu.

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.02.019

one best hit orthologs to *D. melanogaster* protein coding genes were taken from the initial published annotation of the *D. pseudoobscura* genome (Richards et al., 2005). If a gene was duplicated after the split between the *D. melanogaster* and *D. pseudoobscura* lineages, only one copy of the duplicated gene would have been identified in the collection of one to one orthologs. I performed the following *in silico* analyses to identify the second copy of duplicated genes in the *D. pseudoobscura* genome.

The *D. pseudoobscura* genome was partitioned into genic regions (sequences identified as exons or introns in the initial annotation) and intergenic regions (sequence between the genic regions). (The coordinates of the genic and intergenic regions are provided as Supplementary Material.) The genic and intergenic sequences were masked for all known *Drosophila* TEs (downloaded from http://flybase.org) and the *D. pseudoobscura* rearrangement breakpoint motif (Richards et al., 2005) (Genbank accessions AY693425 and AY693426) using RepeatMasker (Smit et al., 2004). (The library containing these repeat sequences is available as Supplementary data.) Each intergenic region was searched against all of the genic regions using MegaBLAST (Zhang et al., 2000) and all *D. melanogaster* proteins using BLASTX (Altschul et al., 1997) (the parameters used in the BLAST searches are given in the Supplementary data). This approach will not identify duplications of *D. pseudoobscura* specific genes, but there are few genes in the *D. pseudoobscura* genome without orthologs in *D. melanogaster* relative to those with orthologs (Drosophila 12 Genomes Consortium, 2007). Some duplicated regions contain assembly gaps, which were closed by extracting the region from the reconciled assembly of the *D. pseudoobscura* genome (http://rana.lbl.gov/drosophila/caf1.html). If the reconciled assembly also contains the assembly gap or is missing the duplication, that paralog was excluded from the dataset. The duplications were also confirmed to not be assembly artifacts by MegaBLAST (Zhang et al., 2000) against the trace files from the initial sequencing of the *D. pseudoobscura* genome (Richards et al., 2005). If multiple trace sequences span the duplication end-points and cover the duplicated sequence, the duplication was retained.

The following steps were taken to align the paralogs. First, the MegaBLAST alignments of the intergenic and genic regions were used to determine the end-points of the duplicated regions; contiguous alignments were appended to each other until all collinear BLAST hits were used. Because the genome was partitioned prior to BLAST, some of the alignments terminated at the end-points of those partitioned genic or intergenic regions. In these situations, the flanking partition was added, and this longer sequence was used to identify the end-points (via BLAST). This process was repeated until the alignment no longer terminated at the end of the partition, and the duplication end-point was identified.

Protein coding sequences were inferred using the initial published annotation of the *D. pseudoobscura* genome (Richards et al., 2005), BLASTX (Altschul et al., 1997) searches against the *D. melanogaster* proteome, and Genscan predictions (Burge and Karlin, 1997). Coding sequences from the two copies of duplicated genes identified using these three methods were aligned to the *D. melanogaster* orthologous coding sequence, and the protein coding sequences of the *D. pseudoobscura* paralogs were constructed manually by creating the largest open reading frame with homology to the *D. melanogaster* ortholog. Nonsense and frameshift mutations were tolerated in the protein coding sequences, and the coding regions were shifted to ensure the *D. melanogaster* open reading frame was maintained. Amino acid sequences of the two *D. pseudoobscura* paralogs and the orthologous *D. melanogaster* protein were aligned using the default settings of the CLUSTALW (Thompson et al., 1994) implementation in MEGA 3.1 (Kumar et al., 2004). Nucleotide sequences of the *D. pseudoobscura* paralogs were overlaid on the amino acid alignment, and non-coding sequences were aligned using CLUSTALW.

Because the search for paralogs is based on sequence identity (rather than gene structure), it will identify duplicated genes with both "complete" and "partial" open reading frames. Completely duplicated genes include the beginning and end of the annotated *D. pseudoobscura* coding sequence — *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 the 5′ end, the 3′ end, or both ends of the annotated coding sequence. This classification does not consider whether the internal portion of the coding sequence has acquired frameshift or nonsense mutations.

2.2. Final trimming of the dataset

Recently duplicated genes are more likely to maintain evidence of their mechanism of origin and to be located in the same region in which they arose. Therefore, an 80% sequence identity cut-off was established for all paralogs to ensure that only recently duplicated genes were included in the dataset. Nucleotide divergence between paralogous sequences was measured at all sites, non-coding sites only, synonymous sites within coding exons (Nei and Gojobori, 1986), and non-coding and synonymous sites together. If paralogs differed at more than 20% of sites in any of these classes of sites (and if there were at least 100 nucleotides in that particular class) that duplication was excluded from the analysis; the 100 site cut-off was chosen to prevent spurious results due to a small sample size of nucleotides. The identity cut-off is also optimal for the MegaBLAST parameters (Gotea et al., 2003). The remaining paralogs were further trimmed to remove any *D. pseudoobscura* genes duplicated more than once in the dataset; this was done to ensure phylogenetic independence of all genes in the dataset. The trimmed dataset contains 88 duplications, containing a total of 101 genes (Supplementary Table S1). All but 11 of the duplications are of a single gene, while 9 contain two genes and 2 contain three genes. 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). The results presented are robust to different measures of duplication length.

2.3. Duplication mechanisms

The mechanism giving rise to each duplicated gene was inferred based on intron–exon structure. If a duplication event contains multiple genes, it was classified as a DNA-based duplication. If both copies of a single duplicated gene made up of multiple exons contain the same intron–exon structure (over the region of the gene that was duplicated), the duplication was classified as DNA-based. If one copy is missing the introns present in the duplicated region of the other copy, the duplication was classified as retroposed. Duplications of single intron-less genes and duplications of single exons were classified as ambiguous. Additionally, one duplicated gene (the ortholog of CG7730) contains two of the three introns present in the other copy — this duplication was classified as ambiguous. Other hallmarks of retroposition (i.e., short flanking repeats and poly-A tails) were not found in the ambiguous duplications.

2.4. Relative positions of paralogs

Drosophila genomes consist of five major chromosome arms and a dot chromosome, and each arm is referred to as a Muller element (Muller, 1940). In *D. pseudoobscura*, a portion of Muller element A is located on chromosome arm XR (which is mostly made up of Muller element D), but the rest of the chromosome arms correspond to individual Muller elements (Schaeffer et al., 2008). Each copy of all duplicated genes was assigned to a Muller element and a chromosome

arm; the results presented are not affected by whether Muller elements or chromosome arms are used. Duplicated genes were assigned to one of three classes based on the relative position of the paralogs: adjacent, non-adjacent intra-chromosome-arm, or inter-chromosome-arm. Adjacent duplications have no genes between them. Non-adjacent duplications have at least one gene between them, but both copies are located on the same chromosome arm. The current distance between paralogs on the same chromosome (in nucleotides) was not included in the analysis because this distance may not reflect the distance between the two loci at the time of the duplication event — as a result of the extensive amount of rearrangements that have occurred in the *D. pseudoobscura* genome (Richards et al., 2005; Bhutkar et al., 2008). The relative position of the paralogs was used to determine which copy of each duplicated gene is the ancestral copy and which is the derived copy (as described in the Supplementary methods).

Each copy of all of the duplications was also assigned as either within a region of conserved gene order between *D. pseudoobscura* and *D. melanogaster* (conserved linkage group, CLG) (Ehrlich et al., 1997) or a region between CLGs (rearrangement breakpoint region). Intergenic regions were classified as rearrangement breakpoint regions if the order of genes flanking those regions was not conserved between *D. pseudoobscura* and *D. melanogaster*. Rearrangement breakpoint regions flanking duplications were polarized along the *D. melanogaster* and *D. pseudoobscura* lineages (i.e., on which lineage the rearrangement giving rise to the breakpoint occurred) using the genome sequences of *D. willistoni* and *D. virilis* (Drosophila 12 Genomes Consortium, 2007) and annotations of one to one orthologs from those genomes to *D. melanogaster* genes (Bhutkar et al., 2007).

2.5. Repeats flanking duplications

I determined if the sequences flanking either copy of all the duplications contain any known *Drosophila* repetitive sequences. The 500 bp, 1 kb, 2 kb, and 5 kb flanking the 5′ and 3′ end of both copies of each duplicated region were extracted from the genome assembly in two ways: allowing the flanking sequences to overlap and not allowing overlap of the flanking sequences (see Supplementary methods). These flanking sequences were used as queries in RepeatMasker (Smit et al., 2004) searches against a library made up of the *D. pseudoobscura* rearrangement breakpoint motif sequence (Richards et al., 2005) and all known *Drosophila* TEs (RepeatMasker library available as Supplementary data). As a control, I determined whether random intergenic regions from the *D. pseudoobscura* genome measuring 500 bp, 1 kb, 2 kb, and 5 kb contain any repeat sequences. Two control datasets were examined: one containing sequences without assembly gaps (153366 sequences measuring 500 bp; 71028 sequences measuring 1 kb; 30936 sequences measuring 2 kb; 8828 sequences measuring 5 kb) and another containing sequences with a maximum of 10% uncalled bases in each sequence (154885 sequences measuring 500 bp; 72925 sequences measuring 1 kb; 33002 sequences measuring 2 kb; 10529 sequences measuring 5 kb). The control dataset containing assembly gaps was used because repetitive sequences tend to be associated with assembly gaps, and excluding assembly gaps from the control dataset may bias it for non-repetitive sequences. The region within 500 bp of each duplication end-point was also examined for repeat sequences. In this analysis, the results of the RepeatMasker search using the 500 bp flanking sequence was interrogated for the closest matching nucleotide to the duplication end-point. The location of this match was used to determine the distance from the duplication to the repeat sequence.

The sequences flanking duplications were also used as queries in BLASTN searches (Altschul et al., 1990) against the *D. pseudoobscura* rearrangement breakpoint motif sequence (Richards et al., 2005) using *E*-values of 1×10^{-5}, 1×10^{-15}, and 1×10^{-25}. Random intergenic regions of 500 bp, 1 kb, 2 kb, and 5 kb (both without assembly gaps and with minimal assembly gaps) were extracted from the *D. pseudoobscura* genome as a control. These sequences were used as BLASTN queries against the same repeat database described above, using the same *E*-values.

3. Results and discussion

3.1. Mechanisms of duplication

Many studies have examined duplicated genes in *Drosophila* genomes (e.g., Betrán et al., 2002; Thornton and Long, 2002; Dai et al., 2006; Bai et al., 2007; Hahn et al., 2007; Yang et al., 2008), while few have characterized the relative contribution of different mechanisms responsible for generating the duplications (e.g., Harrison et al., 2003; Zhou et al., 2008). Each of the 88 duplication events in my dataset was assigned to one of three classes: retroposed, DNA-based duplication, and ambiguous. Eight duplications in this dataset have positive evidence for retroposition, consistent with the finding that retroposition contributes to approximately 10% of all duplicated genes in the *D. melanogaster* species group (Zhou et al., 2008). Additionally, there are 46 DNA-based duplications and 34 ambiguously duplicated genes. The results presented below suggest that many of the ambiguously duplicated genes were generated by retroposition.

Duplicated genes were also classified based on their relative positions, and each duplication event was assigned to one of three classes: inter-chromosome-arm, non-adjacent on the same chromosome arm, and adjacent. Because only two copy gene families were considered, these results likely underestimate the number of adjacent duplications (Pan and Zhang, 2007). However, these data include both apparently functional genes and pseudogenes, which should capture the mutational processes giving rise to duplicated genes without as much distortion from the differential retention of duplicated genes. DNA duplications are more likely to be intra-arm events, while retroposed and ambiguous duplications are more likely to be inter-arm ($P=0.00029$, FET) (Table 1). The similarity in the relative positions of ambiguous and retroposed duplications suggests that many of duplicated genes in the ambiguous class were generated by retroposition. However, ambiguously duplicated genes are also more likely to be intra-arm events than retroposed duplications ($P=0.036$, FET), indicating that some of the ambiguously duplicated genes were probably generated by a DNA-based mechanism. It is unlikely that the non-adjacent duplications arose as tandem duplications and subsequently moved apart via secondary rearrangements because there is no evidence for such secondary rearrangements near non-adjacent duplications — at least one copy, and often both copies, tend to be located in regions of conserved gene order between *D. pseudoobscura* and *D. melanogaster*.

RNA mediated TEs and retroposed genes have a tendency to contain only the 3′ end of the ancestral sequence when a partial copy is duplicated (Lander et al., 2001; Bai et al., 2007). The synthesis of the complementary strand of DNA from an mRNA template begins at the

Table 1
Mechanisms of duplication, relative positions of paralogs, and completeness of derived coding sequences.

Mechanism[c]	Relative position[a]			Complete CDS	Partial CDS[b]	
	Inter-arm	Non-adj	Adjacent		3′ end	Not 3′
DNA dup.	14	18	14	26	2	18
Ambiguous	21	8	5	13	8	13
Retroposed	8	0	0	4	2	2

[a] Whether the ancestral copies are located on different chromosome arms (Inter-arm), on the same chromosome arm, but with at least one other gene in between (Non-adj), or adjacent with no genes in between (Adjacent).
[b] For partially duplicated coding sequences, whether only the 3′ end of the gene was duplicated, or if another region (Not 3′) was duplicated.
[c] Mechanism giving rise to the duplication: DNA-based duplication, ambiguous, or retroposition.

3′ end of the transcript and proceeds toward the 5′ end. If reverse transcription terminates prior to the 5′ end of the transcript, only the 3′ end will be copied. Duplication events that capture single partial genes were classified based on whether only the 3′ end of the coding sequence was duplicated. Very few DNA-based duplications of partial genes contain only the 3′ end, while nearly half of the ambiguous and retroposed partially duplicated genes contain only the 3′ end ($P = 0.025$, FET) (Table 1). This provides further evidence that many of the ambiguous duplications were generated by retroposition.

3.2. Role of repeats in generating duplications

Analysis of the sequences flanking duplications can reveal insights into their origins. DNA duplications tend to have repetitive sequences flanking the 5′ end, 3′ end, or both in a wide variety of eukaryotic genomes, and these repeats are thought to play a role in generating the duplications (Kim et al., 1998; Bailey et al., 2003; Fiston-Lavier et al., 2007; Yang et al., 2008). Both RepeatMasker (Smit et al., 2004) and BLASTN (Altschul et al., 1990) were used to identify repetitive sequences flanking the 5′ and 3′ ends of duplication tracts. The two approaches yield highly similar results, and only the results from the RepeatMasker queries are presented. Additionally, I obtain the same results whether or not I allow for overlap of sequences flanking adjacent duplications — only the results allowing overlapping flanking sequences are presented.

There is not a significant enrichment of TEs flanking the duplicated regions when compared to intergenic controls (Fig. 1A). However, the sequences flanking the duplications do have a higher frequency of matches to the *D. pseudoobscura* rearrangement breakpoint motif (Richards et al., 2005) than intergenic controls, regardless of the length of the flanking regions examined (Fig. 1B). The same results are observed whether or not I use control sequences with assembly gaps. The lack of an enrichment of TEs flanking *D. pseudoobscura* duplications suggests that the breakpoint motif plays a larger role than TEs in generating duplicated genes in the *D. pseudoobscura* genome. Alternatively, the database of TEs I used may not adequately represent the TEs found in the *D. pseudoobscura* genome. I focus my remaining analysis of flanking sequences on the breakpoint motif because this is the only sequence enriched around the duplications.

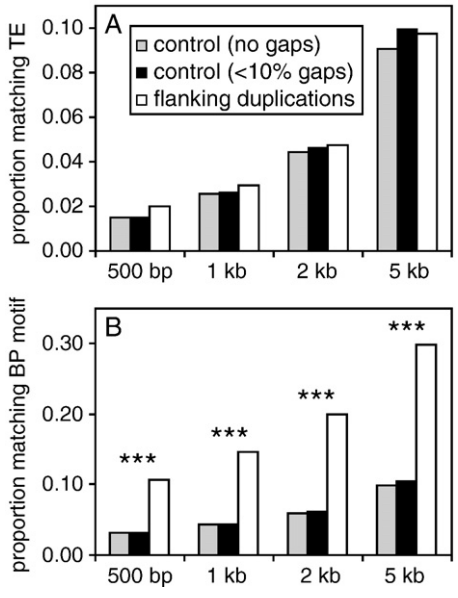

Fig. 1. Frequency of flanking sequences containing repetitive elements. The proportion of sequences flanking duplications matching (A) known *Drosophila* TEs and (B) the *D. pseudoobscura* breakpoint motif (Richards et al., 2005) are indicated by white bars. Flanking sequences of 500 bp, 1 kb, 2 kb, and 5 kb were analyzed. Control regions either have no assembly gaps (gray bars) or assembly gaps make up less than 10% of the entire control region (black bars). Comparisons between flanking sequences and controls for which $P < 0.0005$ using a z test are indicated by three asterisks.

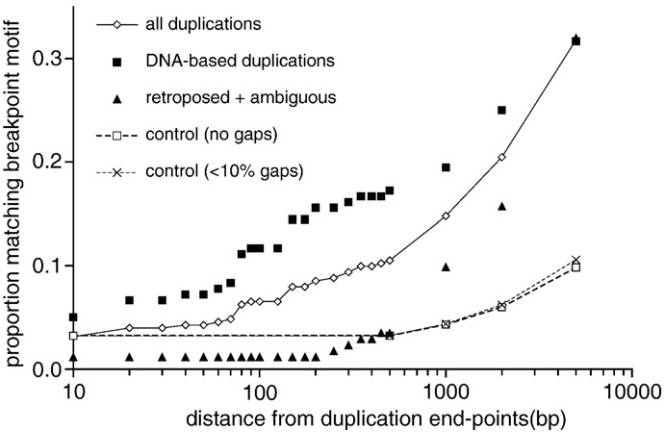

Fig. 2. Breakpoint motifs at progressive distances from duplication end-points. The proportion of 500 bp, 1 kb, 2 kb, and 5 kb regions flanking duplications with RepeatMasker (Smit et al., 2004) hits to the breakpoint motif (Richards et al., 2005) are shown for all duplications (white diamonds), DNA-based duplications (black squares), and retroposed duplications (black triangles). Additionally, the distance from the duplication of the first hit within the 500 bp region was determined, and the frequency of flanking sequences with a match at various distances from the duplication end-points is graphed. Intergenic regions measuring 500 bp, 1 kb, 2 kb, and 5 kb were used as controls and searched against the same RepeatMasker library. Control regions either have no assembly gaps (white squares) or assembly gaps make up less than 10% of the entire control region ("X"). The dashed line representing the control regions less than 500 bp is the fraction of 500 bp control regions containing the breakpoint motif.

If the rearrangement breakpoint motif sequence is responsible for generating DNA duplications and not retroposed genes, we expect it to only be enriched around the DNA duplications. The 500 bp regions flanking DNA duplications contain a significant excess of motif sequences when compared to those flanking ambiguous duplications ($P = 0.000065$, FET) and those flanking ambiguous and retroposed duplications together ($P = 0.000063$, FET) (Fig. 2). The same is true for 1 kb and 2 kb flanking intervals. There is not a significant excess of motif sequences flanking DNA duplications when compared to only retroposed duplications for any of the flanking intervals examined; this may be because of the small sample size of retroposed duplications. A significant excess of DNA duplications are flanked by the breakpoint motif within 500 bp relative to intergenic controls, but there is no evidence for an enrichment of the motif flanking ambiguous and retroposed duplications (Fig. 2). The lower frequency of the breakpoint motif flanking ambiguously duplicated genes (relative to DNA duplications and intergenic controls) further supports the hypothesis that many of the ambiguously duplicated genes were generated via retroposition. Also, the close proximity of the breakpoint motif to the DNA duplications (Fig. 2) suggests that the motif sequence is involved in the duplication events in the mechanism described below.

The model for DNA-based duplication via NAHR in *Drosophila* requires repeats flanking both the ancestral and derived copies (Fiston-Lavier et al., 2007). In this model, a DSB in one repetitive sequence is repaired using a non-allelic repeat sequence with high identity as a template via a synthesis dependent strand annealing (SDSA) pathway. These repetitive sequences are expected to be located near the duplication end-points. However, because the repeat sequence may be part of the duplication itself, and not in the flanking region, it is possible that only one of the copies will have the repeat sequence outside of the duplicated region (i.e., in the flanking sequence). I examined how many duplications had at least one copy flanked by the rearrangement breakpoint motif and how many had both copies flanked by the motif (Table 2; Supplementary Table S1). There is a significant excess of DNA duplications with at least one copy flanked by the motif within 500 bp when compared to ambiguous

Table 2
Rearrangement breakpoint motif flanking duplications[a].

Class[c]	500 bp[b]			1 kb[b]			2 kb[b]			5 kb[b]		
	Both	One	Neither	Both	One	Neither	Both	One	Neither	Both	One	Neither
DNA dup.	9	10	27	11	10	25	13	14	19	17	15	14
Ambiguous	0	4	30	2	9	23	5	10	19	11	12	11
Retroposed	1	0	7	1	1	6	1	1	6	2	1	5
DNA duplications												
Inter-arm	3	6	5	4	6	4	4	8	2	6	7	1
Non-adj	5	3	10	6	3	9	7	3	8	7	5	6
Adjacent	1	1	12	1	1	12	2	3	9	4	3	7

[a] Counts of duplications in which both copies, one copy, or neither copy are flanked by the rearrangement breakpoint motif.
[b] Size of interval flanking ancestral and derived copies of duplications.
[c] Whether a duplication is DNA-based, ambiguous, or retroposition; or whether a DNA duplications is inter-arm, non-adjacent intra-arm, or adjacent.

duplications ($P = 0.0034$, FET) and ambiguous and retroposed duplications together ($P = 0.0018$, FET). There is also a significant excess of DNA duplications with both copies flanked by the motif within 500 bp when compared to ambiguous duplications ($P = 0.0048$, FET) and ambiguous and retroposed duplications together ($P = 0.011$, FET). The same is true when one looks at duplications in which both copies are flanked by the motif within 1 kb. Additionally, of the nine DNA duplications in which both copies are flanked by the motif within 500 bp, seven have a motif sequence in the same orientation in both copies. This strongly suggests that the rearrangement breakpoint motif is used as a non-allelic template to repair a DSB using the SDSA pathway, giving rise to duplicated genes (Fiston-Lavier et al., 2007).

It has previously been observed that tandem duplications in *Drosophila* genomes are less associated with repetitive sequences than non-tandem duplications (Zhou et al., 2008). Indeed, I also observe that there is a significant excess of non-adjacent DNA duplications (both intra- and inter-chromosome-arm) in which at least one copy is flanked by the breakpoint motif when compared to adjacent DNA duplications for 500 bp ($P = 0.014$, FET), 1 kb ($P = 0.0050$, FET), and 2 kb ($P = 0.039$, FET) flanking regions (Table 2). The same pattern is not observed when one looks at ambiguous duplications. These data provide support for the hypothesis that non-adjacent DNA duplications arise via NAHR, while adjacent DNA duplications are the result of NHEJ (Zhou et al., 2008).

3.3. The effect of relative position on the lengths of duplicated regions

The lengths of the duplications may reveal information regarding their origins. Intra-arm duplications are longer than inter-arm duplications, and adjacent duplications are longer than non-adjacent duplications ($H = 21.67$, $P < 0.001$) (Supplementary Table S1), regardless of how duplication length is measured. These results hold whether or not one large duplication, which is twice the size of the next largest duplication, is included in the analysis. The difference in length may be the result of different mechanisms giving rise to these two classes — intra-arm duplications tend to be DNA-based duplications (Table 1), which may allow for longer duplication events than retroposed duplications (which tend to be inter-arm). Differences between the

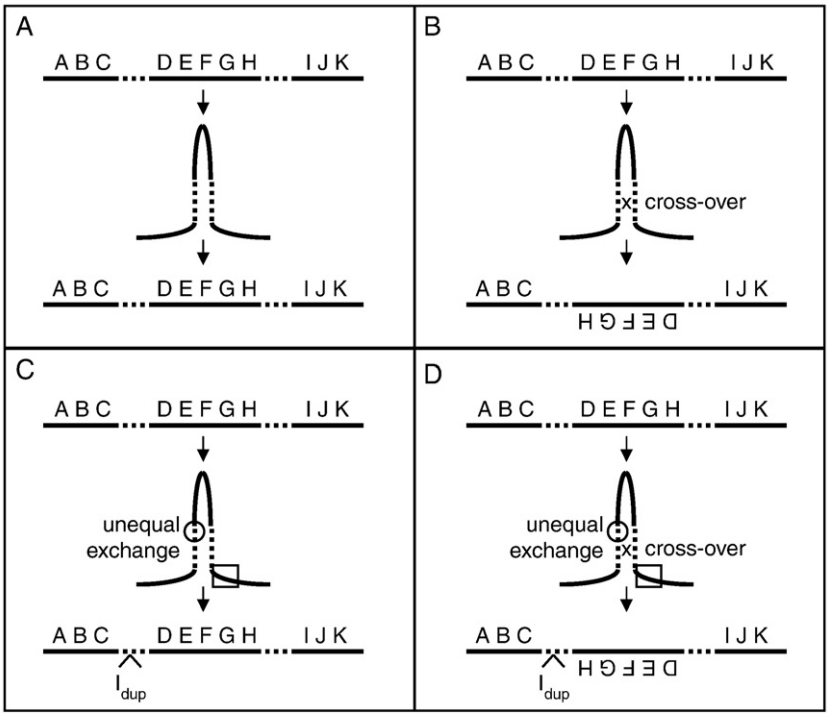

Fig. 3. Alternative outcomes of non-allelic DSB repair. Solid lines indicate regions containing genes, with the genes represented by single letters. Dashed lines show repeat sequences that pair during non-allelic DSB repair. The genomic features are not drawn to scale. Four possible outcomes of non-allelic DSB repair are presented: (A) gene conversion between repeats, (B) crossing over giving rise to a chromosomal inversion, (C) non-reciprocal exchange leading a duplicated gene, and (D) crossing over with non-reciprocal exchange causing an inversion with a gene duplication in a breakpoint. Locations of cross-over events (B, D) are indicated with an "x". Regions donating sequence in an unequal crossing over event (C, D) are shown with squares, while regions receiving sequence are circled. The location of the duplication event (C, D) is indicated by an arrowhead.

lengths of intra- and inter-arm DNA-based duplications (Fiston-Lavier et al., 2007) may also be responsible for the observed difference in length between all intra- and inter-arm duplications. Additionally, adjacent duplications are longer than non-adjacent duplications, which may be the result of different mechanisms of DNA-based duplication between these two processes — if adjacent duplications are the result of NHEJ, and non-adjacent duplications arise via a NAHR pathway utilizing SDSA (Fiston-Lavier et al., 2007; Zhou et al., 2008).

3.4. Chromosomal inversion and gene duplication

The processes of gene duplication and chromosomal inversion appear to be coupled in a wide variety of taxa (Fischer et al., 2001; Katju and Lynch, 2003; Goidts et al., 2004; Kehrer-Sawatzki et al., 2005; Matzkin et al., 2005; Sharakhov et al., 2006; Ranz et al., 2007). Two models have been presented to explain the presence of DNA duplications at inversion breakpoint regions. Traditionally, it has been thought that inversions arise when a non-allelic intrachromatid template is used to repair a DSB (Fig. 3B) (Engels and Preston, 1984; Richards et al., 2005). This template is chosen because of high sequence identity between the DSB region and the template as the result of a shared repetitive sequence. If there is non-reciprocal genetic exchange during the crossing over event, a gene flanking one inversion breakpoint may be duplicated into the other breakpoint region (Fig. 3D) (Matzkin et al., 2005; Sharakhov et al., 2006). However, an alternative model has been presented by which intrachromatid staggered SSBs are repaired using NHEJ (Kehrer-Sawatzki et al., 2005; Ranz et al., 2007). When the sticky ends are filled in using the template sequence, genes flanking the SSBs can get duplicated.

I interrogated the duplicated genes in my dataset to see if any are found within the breakpoints of inversions that occurred along the *D. pseudoobscura* lineage after the divergence with the *D. melanogaster* lineage. Genes found in the same order in both *D. pseudoobscura* and *D. melanogaster* are referred to as CLGs (Ehrlich et al., 1997), and sequences in between CLGs are rearrangement breakpoint regions. Rearrangement breakpoints in *Drosophila* are generated by inversion events (Richards et al., 2005; Bhutkar et al., 2008). Gene order in two outgroup species (*D. willistoni* and *D. virilis*) was used to polarize inversion events along either the *D. melanogaster* or *D. pseudoobscura* lineages.

I found one duplicated gene that may have arisen via an inversion event along the *D. pseudoobscura* lineage (the ortholog of *TppII* [CG3991]). The derived copy is in a breakpoint region and an ancestral copy flanks the other breakpoint region of the same inversion (Supplementary Fig. S1). The derived copy of *TppII* in *D. pseudoobscura* (*TppIIψ*) contains a partial coding sequence (556 bp 5′ of the start codon, the first exon [169 bp], and 37 bp of the first intron), and it arose via a DNA-based duplication. Both the ancestral and derived copies are flanked by the breakpoint motif within 500 bp of the endpoints of the duplications, and the motif sequences are in the same orientation relative to the duplicated sequences. The paralogs differ at 10% of their non-coding and synonymous sites. The *obscura* species group consists of three subgroups: *pseudoobscura*, *affinis*, and *obscura*. The *Adh* and *Gpdh* genes from the *pseudoobscura* species subgroup differ from those from the *affinis* and *obscura* subgroups at approximately 20–40% of all synonymous sites (Russo et al., 1995; Wells, 1996). Therefore, if this DNA-based duplication is the result of unequal crossing over during an inversion event that was initiated by NAHR to repair a DSB, the inversion most likely occurred after the split between the *pseudoobscura* subgroup and the *affinis* and *obscura* subgroups. However, all duplicated genes found thus far at *Drosophila* and *Anopheles* inversion breakpoints have contained partial coding sequences (Matzkin et al., 2005; Sharakhov et al., 2006; Ranz et al., 2007) (Supplementary Fig. S1), suggesting that duplications generated within inversion breakpoints are not likely to generate much evolutionary novelty.

The mechanisms giving rise to inversions and non-adjacent DNA duplications in *D. pseudoobscura* both involve NAHR to repair a DSB. Inversions are generated by crossing over between repeat sequences on the same chromatid when a non-allelic sequence is used as a template to repair a DSB (Fig. 3B) (Engels and Preston, 1984; Richards et al., 2005). Non-adjacent DNA-based duplications in *Drosophila* can occur when a SDSA pathway is used to repair a DSB using a non-allelic sequence as a template (Fiston-Lavier et al., 2007). The non-allelic templates in *D. melanogaster* tend to be TEs, but the DNA-based duplications in *D. pseudoobscura* are flanked by the breakpoint motif (Figs. 1 and 2; Table 2). While inversions and duplications can both be initiated by DSB repair, the end result of the DSB repair event depends on the pathway used. Based on the results presented here, one can imagine four possible outcomes of using an intrachromatid non-allelic template to repair a DSB: 1) gene conversion between the repeat sequences (Slightom et al., 1980; Petes and Fink, 1982); 2) crossing over resulting in an inversion; 3) SDSA giving rise to a duplicated gene; or 4) crossing over with unequal genetic exchange giving rise to an inversion with a duplicated gene in the breakpoint region (Fig. 3).

4. Conclusions

Most of the recently duplicated genes in the *D. pseudoobscura* genome were generated by a DNA-based mechanism or lack the traditional evidence of retroposition. However, many of the ambiguously duplicated genes resemble retroposed genes in the relative position of the paralogs, the portion of the gene that was duplicated, the length of the duplication event, and the lack of flanking repetitive sequences. The non-adjacent DNA-based gene duplications are often flanked by the same repetitive sequence found at inversion breakpoint regions, suggesting that non-adjacent DNA duplications are generated via similar mechanisms as those responsible for chromosomal inversions.

Acknowledgements

N. Hasan, B. B. Hilldorfer, R. LeGros, and R. L. Zindren helped with sorting the BLAST hits. S. W. Schaeffer, J. R. Arguello, and multiple anonymous reviewers 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. A. Bhutkar kindly provided the one to one ortholog calls for *D. willistoni* and *D. virilis*.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2009.02.019.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25, 3389–3402.
Bai, Y., Casola, C., Feschotte, C., Betran, E., 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. Genome Biol. 8, R11.
Bailey, J.A., Liu, G., Eichler, E.E., 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am. J. Hum. Genet. 73, 823–834.
Betrán, E., Thornton, K., Long, M., 2002. Retroposed new genes out of the X in *Drosophila*. Genome Res. 12, 1854–1859.
Bhutkar, A., Russo, S.M., Smith, T.F., Gelbart, W.M., 2007. Genome-scale analysis of positionally relocated genes. Genome Res. 17, 1880–1887.
Bhutkar, A., Schaeffer, S.W., Russo, S.M., Xu, M., Smith, T.F., Gelbart, W.M., 2008. Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. Genetics 179, 1657–1680.
Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94.
Dai, H., Yoshimatsu, T.F., Long, M., 2006. Retrogene movement within- and between-chromosomes in the evolution of *Drosophila* genomes. Gene. 385, 96–102.

Drosophila 12 Genomes Consortium, 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450, 203–218.
Ehrlich, J., Sankoff, D., Nadeau, J.H., 1997. Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics 147, 289–296.
Engels, W.R., Preston, C.R., 1984. Formation of chromosome rearrangements by *P* factors in *Drosophila*. Genetics 107, 657–678.
Esnault, C.C., Maestre, J.L., Heidmann, T., 2000. Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. 24, 363–367.
Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C., Dujon, B., 2001. Evolution of gene order in the genomes of two related yeast species. Genome Res. 11, 2009–2019.
Fiston-Lavier, A.-S., Anxolabehere, D., Quesneville, H., 2007. A model of segmental duplication formation in *Drosophila melanogaster*. Genome Res. 17, 1458–1470.
Goidts, V., Szamalek, J.M., Hameister, H., Kehrer-Sawatzki, H., 2004. Segmental duplication associated with the human-specific inversion of chromosome 18: a further example of the impact of segmental duplications on karyotype and genome evolution in primates. Hum. Genet. 115, 116–122.
Gotea, V., Veeramachaneni, V., Makalowski, W., 2003. Mastering seeds for genomic size nucleotide BLAST searches. Nucl. Acids Res. 31, 6935–6941.
Hahn, M.W., Han, M.V., Han, S.-G., 2007. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet. 3, e197.
Harrison, P.M., Milburn, D., Zhang, Z., Bertone, P., Gerstein, M., 2003. Identification of pseudogenes in the *Drosophila melanogaster* genome. Nucl. Acids Res. 31, 1033–1037.
Kaessmann, H., Vinckenbosch, N., Long, M., 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nat. Rev. Genet. 10, 19–31.
Katju, V., Lynch, M., 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics 165, 1793–1803.
Kehrer-Sawatzki, H., Sandig, C.A., Goidts, V., Hameister, H., 2005. Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. Cytogenet. Genome Res. 108, 91–97.
Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., Voytas, D.F., 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res. 8, 464–478.
Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. 5, 150–163.
Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Matzkin, L.M., Merritt, T.J.S., Zhu, C.-T., Eanes, W.F., 2005. The structure and population genetics of the breakpoints associated with the cosmopolitan chromosomal inversion *In(3R)Payne* in *Drosophila melanogaster*. Genetics 170, 1143–1152.
Muller, H.J., 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.
Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag, New York, NY.
Otto, S.P., 2007. The evolutionary consequences of polyploidy. Cell 131, 452–462.
Pan, D., Zhang, L., 2007. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. Genome Biol. 8, R158.
Petes, T.D., Fink, G.R., 1982. Gene conversion between repeated genes. Nature 300, 216–217.
Ranz, J.M., et al., 2007. Principles of genome evolution in the *Drosophila melanogaster* species group. PLoS Biol. 5, e152.
Richards, S., et al., 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and *cis*-element evolution. Genome Res. 15, 1–18.
Russo, C.A., Takezaki, N., Nei, M., 1995. Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. 12, 391–404.
Schaeffer, S.W., et al., 2008. Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics 179, 1601–1655.
Sharakhov, I.V., et al., 2006. Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (*2La*) in the *Anopheles gambiae* complex. PNAS 103, 6258–6262.
Slightom, J.L., Blechl, A.E., 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, A.F.A., Hubley, R. and Green, P., 2004. RepeatMasker Open-3.0.
Tamura, K., Subramanian, S., Kumar, S., 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol. Biol. Evol. 21, 36–44.
Thompson, J.D., Higgins, D.G., Gibson, T.J., 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., 2002. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. Mol. Biol. Evol. 19, 918–925.
Wells, R.S., 1996. Nucleotide variation at the Gpdh locus in the genus *Drosophila*. Genetics 143, 375–384.
Yang, S., et al., 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.
Zhou, Q., et al., 2008. On the origin of new genes in *Drosophila*. Genome Res. 18, 1446–1455.