BIOINFORMATICS DISCOVERY NOTE Vol. 27 no. 13 2011, pages 1749–1753
doi:10.1093/bioinformatics/btr280

Genome analysis

Advance Access publication May 5, 2011

A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes

Yong E. Zhang[1,2,*], Maria D. Vibranovski[1], Benjamin H. Krinsky[1,3] and Manyuan Long[1,3,*]

[1]Department of Ecology and Evolution, The University of Chicago, 1101 E 57th Street, [2]Department of Molecular Genetics and Cell Biology, The University of Chicago, 920 E 58th Street and [3]Committee on Evolutionary Biology, The University of Chicago, 1025 E 57th Street, Chicago, IL 60637, USA

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Retrocopies are important genes in the genomes of almost all higher eukaryotes. However, the annotation of such genes is a non-trivial task. Intronless genes have often been considered to be retroposed copies of intron-containing paralogs. Such categorization relies on the implicit premise that alignable regions of the duplicates should be long enough to cover exon–exon junctions of the intron-containing genes, and thus intron loss events can be inferred. Here, we examined the alternative possibility that intronless genes could be generated by partial DNA-based duplication of intron-containing genes in the fruitfly genome.

Results: By building pairwise protein-, transcript- and genome-level DNA alignments between intronless genes and their corresponding intron-containing paralogs, we found that alignments do not cover exon–exon junctions in 40% of cases and thus no intron loss could be inferred. For these cases, the candidate parental proteins tend to be partially duplicated, and intergenic sequences or neighboring genes are included in the intronless paralog. Moreover, we observed that it is significantly less likely for these paralogs to show inter-chromosomal duplication and testis-dominant transcription, compared to the remaining 60% of cases with evidence of clear intron loss (retrogenes). These lines of analysis reveal that DNA-based duplication contributes significantly to the 40% of cases of single exon gene duplication. Finally, we performed an analogous survey in the human genome and the result is similar, wherein 34% of the cases do not cover exon–exon junctions. Thus, genome annotation for retrogene identification should discard candidates without clear evidence of intron loss.

Contact: mlong@uchicago.edu; zhangy@uchicago.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 9, 2011; revised on April 4, 2011; accepted on April 20, 2011

*To whom correspondence should be addressed.

1 INTRODUCTION

Mechanisms that generate new gene duplicates can be roughly divided into two categories: DNA-based mechanisms and retroposition (Zhang *et al*., 2010a,b; Zhou *et al*., 2008). Unlike DNA-based duplicates, retrogenes have less evolutionary constraint, given that they lose the *cis*-regulator elements of their parental loci, and thus might more often undergo neofunctionalization (Brosius, 2003). The empirical data confirming this hypothesis are that retrogenes are more often fast evolving relative to DNA-based duplicates (Cusack and Wolfe, 2007). Thus, because DNA- and RNA-based duplications likely have different functional and evolutionary consequences, it is necessary to disentangle the origination mechanisms of duplicated genes.

Since retroposition uses mature mRNA as the template instead of intron-containing pre-mRNA, the signature of intron loss has been proposed as a hallmark to differentiate these two duplication mechanisms for a derived duplicate with an intron-containing parental gene (Brosius, 2003). It is standard annotation practice to perform an all-against-all protein alignment to search paralogous gene pairs. If one copy has at least one intron while the other copy is intronless, the latter is defined as a retrogene, with the assumption that the alignment should be long enough to cover exon–exon junctions, thus allowing intron loss to be inferred. A more conservative method is to check the actual alignments and discard cases where the alignment is too short and no parental introns are covered (Betran *et al*., 2002; Emerson *et al*., 2004). It is difficult to compare the performance of these two strategies since it remains unknown how many alignments between single-exon new genes and intron-containing parents do not cover introns and whether DNA-based duplication contributes to many of these cases.

Here, we performed a computational survey of intronless genes with intron-containing paralogs in the genomes of both fruitfly and human. The purpose of our work is 2-fold. First, we attempted to provide an evaluation of how we classify retrogene and DNA-based duplicates. The directionality of intron loss can be used to define parental/daughter gene relationships in the case of the retroposition-based copying mechanism. If there is actually no evidence for intron loss, it is difficult to define which paralog is the derived copy and which is the ancestral copy. Such information is critical for the study of evolutionary novelty contributed by

new genes. Second, since DNA-based duplicates and retrogenes are subject to different evolutionary trajectories (Brosius, 2003; Cusack and Wolfe, 2007), we wanted to address whether partial DNA-level duplication of intron-containing genes makes a significant contribution to the presence of intronless genes.

By building protein-, transcript- and genome-level alignments, and performing transcriptional analysis, we found that intron loss was less likely for 40% (47) of cases in fruitfly and 34% (97) of cases in human. In these cases, the alignable regions do not cover exon–exon junctions and are clearly the consequence of partial duplicates relative to the parental genes. In 38% (18 out of 47) cases in *Drosophila*, the duplication blocks cover introns, intergenic regions and neighboring genes that are not compatible with retroposition. Duplication direction and transcriptional profiling further support the notion that these 47 cases have different biological features in comparison with the remaining cases whose alignments span at least one exon–exon junction. This set of analyses supports a conservative method of retrogene identification and suggests that DNA-based duplication of intron-containing genes contributes to the formation of many new intronless genes.

2 METHODS

We started with the parental and daughter gene pairs generated in our previous duplicate databases extracted from the *D.melanogaster* and human genomes (Zhang *et al*., 2010a,b). We excluded cases of genes encoding an intronless isoform and an intron-containing isoform to prevent ambiguity. We built protein and transcript alignments with BLAST (Altschul *et al*., 1997) and parsed the result with the chained SearchIO module provided by BioPerl (Stajich *et al*., 2002). In order to avoid arbitrary parameters, we called an intron loss event only if the corresponding exon–exon junction was covered by the alignable region. It is possible that some duplicates were falsely categorized as retrogenes if the exon–exon junctions were near the end of alignments, where the alignment tends to be less reliable. Thus, if this were the case, the retrogene dataset may contain some copies that are in fact DNA-level duplicates. There are, however, not many such dubious cases. In human, 150 (80%) out of 188 cases of duplications with intron loss involve at least two intron losses. In fruitfly, this proportion drops to 63% (43 out of 68). We further manually checked the remaining 25 cases in *Drosophila* involving one intron loss and found only three cases where the junction is near to the alignment border (less than 10 amino acids away from one alignment end). Thus, there are not many duplicates misidentified as retrogenes due to dubious intron loss. Actually, even if there were some such cases, it would only mean that our main conclusion that 30–40% single-exon genes lacking evidence of intron loss is actually conservative.

We generated the self netted and chained genome alignment for fruitfly by following UCSC's pipeline (Kent *et al*., 2003; Kuhn *et al*., 2007; Schwartz *et al*., 2003), where the best alignable paralogous genomic region for each genomic locus was identified with a scoring matrix allowing longer gaps.

Following a similar procedure in (Zhang *et al*., 2010a), we processed FlyAtlas microarray data (Chintapalli *et al*., 2007) using the Bioconductor platform (Gentleman *et al*., 2004). In brief, we used the customized array annotation file to filter probes mapping to both parental and daughter genes (Dai *et al*., 2005), the GCRMA package to generate gene level summary and gplots package to generate heatmaps.

We built codon-based alignments between parental genes and daughter copies by aligning the protein first by BLAST followed by conceptual translation (Suyama *et al.*, 2006). We used CODEML in the PAML package (Yang, 2007) to infer the pair-wise *Ka/Ks*, the ratio between non-synonymous substitution rate and synonymous substitution rate. We further performed a likelihood ratio test to investigate whether *Ka/Ks* is significantly ($P < 0.05$) smaller than 0.5.

We followed the previous pipeline in (Zhang *et al.*, 2007) and mapped UniGene (Wheeler *et al.*, 2008) ESTs sequences unambiguously to genes.

3 RESULTS

3.1 Expectations of the RNA- and DNA-based duplications

We investigated the origination mechanism of single exon duplicate genes by examining their different expected gene structures, chromosomal distributions and tissue-biased expression patterns.

For gene structures, a RNA-based duplications will more likely create more complete coding structures since mature transcripts are generally duplicated (Brosius, 2003; Kaessmann *et al*., 2009). In contrast, DNA-based duplication can copy only a part of a parental gene region, and thus does not necessarily duplicate complete coding structures (Emerson *et al*., 2008). In addition, longer fragment DNA-based duplication may involve parts of more than one gene (Fan *et al*., 2008).

Regarding chromosomal distribution, DNA-based duplication more often leads to duplicates within the same chromosomes, with a lower frequency of inter-chromosomal events (Zhang *et al*., 2010a).

It has also been observed that RNA-based duplicates are more often associated with testis-dominant expression, possibly because retrogenes lose their original promoters and the testis is transcriptionally permissive (Kaessmann, 2010). Since DNA-based duplicates can carry the original promoter, they should be less frequently dominantly transcribed in testis.

We examined these expectations by analyzing the actual single exon young gene databases in fruitfly and human we previously built (Zhang *et al*., 2010a,b).

3.2 DNA-based duplication contributes to single exon genes in fruitfly

We have previously identified 115 single exon genes in fruitfly whose parental duplicates contain at least one intron (Zhang *et al*., 2010a). Out of these 115 intronless genes, we identified explicitly at least 1 intron loss event in 67 cases based on the protein-level alignment (e.g. Fig. 1A). In addition, the alignment between transcripts showed that CG12324 was derived from CG2033 where two intron losses occurred in the 5′-UTR region (Fig. 1B). These 68 cases together represent typical retroposition events, which will be hereafter referred as duplication with intron loss (DIL) (Supplementary Material Table S1). For the remaining 47 cases (Supplementary Material Table S2; Fig. 1C and 1D), no molecular signatures of intron loss could be detected either by protein-, transcript- or genome-level alignments, since no exon–exon junctions were covered in the alignments. These cases will be hereafter referred as duplication without intron loss (DWIL).

Further comparative analyses revealed multiple lines of evidence showing that DNA-based duplication obviously contributes to the formation of the 47 single exon genes or genes that have no intron loss signatures. First, the derived copies are frequently partial duplicates of the parental proteins. As shown in Figure 2, the sequence coverage of the parental gene in more than 50% of the protein alignments is lower than 52%, a percentage much lower than the median coverage of the 68 cases of DIL (85%, Wilcoxon rank test $P = 0.01$). One example is shown in Fig. 1C where CG33797 is

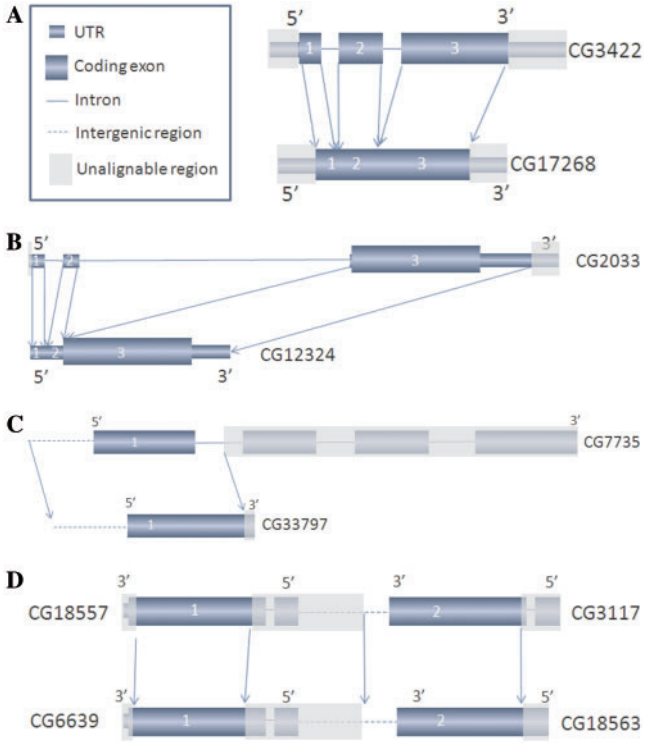

Fig. 1. Schematic alignments between parental and daughter genes. Arrowed lines together with numbers in white (1–3) mark the correspondingly alignable regions between two genes while shaded boxes indicate non-alignable regions. 5′ and 3′ indicate transcription direction. Notably, one alignable region can consist of one exon together with flanking regions (introns or intergenic regions) (**A**) Protein alignment between CG17268 and CG3422. Clearly, there are two intron loss events in the protein-coding region of CG17268. (**B**) Transcript alignment between CG2033 and CG12324. Since introns only exist in the 5′-UTR region of CG2033, intron loss events could be only identified based on the transcript alignment. (**C**) Genomic alignment between CG7735 and CG33797. (**D**) Genomic alignment between CG3177 and CG18563. Notably some portion of the flanking regions (CG3117 and CG18563) is also alignable.

mainly derived from the first coding exon of CG7735 (30% of the parental protein).

As the second line of evidence, the duplication blocks tend to cover non-mRNA sequences including introns or intergenic sequences. For 18 (38%) cases, the duplication blocks cover 5′ or 3′ flanking regions by at least 50 bp. For example, in the case of CG33797 (Fig. 1C), the alignment appears to cover some portion of the first intron together with the 5′ flanking region of the parental gene. The dot plot in Supplementary Figure S1 clearly shows that the duplication blocks extend to 5′ and 3′ non-mRNA regions.

Third, in the extreme case, the duplication block can involve neighboring genes. As shown in Figure 1D, CG18563 appears to be derived from an intron-containing parental gene CG3117. Interestingly, the neighbor of CG18563, CG6639, shares similarity with a gene adjacent to CG3117, namely CG18557. Since it is less likely that two independent duplication events are involved in one single pair of genomic loci, such an alignment suggests that one larger DNA-based duplication covers both genes. In other words, it is possible that a DNA-level duplication covering both CG18557 and CG3117 lead to two new genes, CG6639 and CG18563. Later on, rapid divergence or extensive recombination may have occurred in the middle part of the duplication block rendering this region unalignable.

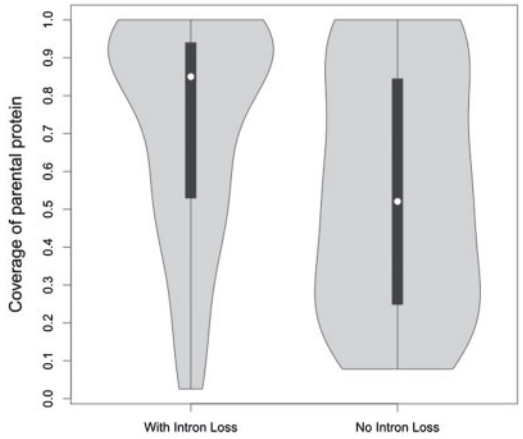

Fig. 2. Distribution of alignment coverage for parental genes in two scenarios, 68 DIL cases and 47 DWIL cases. The white circle marks the median. The thick black bar indicates the 25% and 75% percentiles, while the shape indicates the distribution density.

Table 1. Relative chromosomal location for duplication pairs

	Intra-chromosomal	Inter-chromosomal
DIL	26	42
DWIL	25	22
FET $P = 0.08$		

Although all these features are compatible with DNA-based duplication, retroposition may also create such duplicates, although with a lower probability. Specifically, partial duplication might be caused by incomplete retrotransposition, while duplication blocks covering introns or flanking regions might be explained by ancestral alternative isoforms. Moreover, duplication involving two neighboring genes may occur via retroposition of a readthrough transcript (Zhang *et al*., 2009).

However, it is possible to investigate other genomic features of DWIL cases in comparison to DIL (standard retrogene cases). In other words, genomic characteristics known to be distinct between RNA- and DNA-based duplications can be used to evaluate their contributions to the formation of duplications without intron loss. First, as mentioned above, it is predicted that parental and daughter genes without intron loss are more likely to be located in the same chromosome if they are generated by DNA-based duplication (Zhang *et al.*, 2010a). Consistently, as shown in Table 1, duplications where parental and daughter genes with intron loss are encoded in the same chromosomes constitute 38% of all such cases (26 out of 68). In DWIL cases, however, this proportion increases to 53% (25 out of 47, marginal significance by One-sided Fisher's Exact Test FET, $P = 0.08$).

Second, also as previously mentioned, if many of these 47 cases were generated via a DNA-based mechanism, we expect that

Table 2. Relative chromosomal location for duplication pairs

	Intra-chromosomal	Inter-chromosomal
DIL	16	172
DWIL	33	64
FET $P = 2 \times 10^{-7}$		

these genes would have lower testis transcription. As expected, for intronless daughter genes derived from intron-containing parental genes, 41 out of 66 (62%) show testis-dominant expression (Supplementary Fig. S2). In contrast, in DWIL cases, this proportion drops to 40% (17 out of 42, one-sided FET $P = 0.02$).

Notably, such a pattern could not be interpreted by confounding factors. First, it is known that younger genes are predominantly expressed in testis (Zhang *et al*., 2010a). Based on the age information in (Zhang *et al*., 2010a), daughter genes without intron loss are actually slightly younger although not significantly different from daughter genes with intron loss. Second, out-of-X duplicates, i.e. autosomal duplicates with X-linked parental genes, tend to be expressed in testis (Betran *et al*., 2002; Vibranovski *et al*., 2009). The two datasets (DIL and DWIL) show a similar proportion (∼26%) of out-of-X duplicates, which therefore also cannot explain their different trends regarding testis expression.

3.3 DNA-based duplication contributes to single exon genes in human

Consistently, in the 285 human intronless genes, which we previously identified as derived from intron-containing parental genes, we found 97 cases that had no intron loss signature (Supplementary Material Tables S3, S4). We further observed that these 97 cases have much lower parental gene coverage relative to the remaining cases (median 56% versus 98%, Wilcoxon rank test $P = 5 \times 10^{-10}$). Moreover, these cases also show significantly higher within-chromosomal duplications (34% versus 9%, Table 2).

3.4 Single exon DWIL cases appear to be functional

By definition, DWIL cases only duplicate one exon of the parental gene. Since the median peptide length encoded by one exon is only 72 for fruitfly and 41 for human (Supplementary Material Fig. S3A), it is possible that DWIL proteins are too short and thus non-functional.

In order to test this hypothesis, we first investigated the length distribution of duplicated regions. As shown in Supplementary Fig. S3B, the duplicated regions of parental proteins in DWIL cases are much longer than the background exon length distribution (163 amino acids for fruitfly and 143 for human). In other words, larger coding exons are preferentially copied.

Second, if DWIL genes are pseudogenic, they should be free of constraint. Thus, *Ka/Ks* between DWIL gene and their corresponding parental genes should be above 0.5 (Betran *et al.*, 2002; Emerson *et al.*, 2004). However, as shown in Supplementary Figure S4, pairwise *Ka/Ks* is small with a median of 0.06 in fruitfly and 0.12 in human, respectively. Likelihood-ratio tests show that *Ka/Ks* is significantly smaller than 0.5 in more than half of the cases (74% in fruitfly and 55% in human) suggesting protein-level constraint.

Finally, Supplementary Figure S2B shows that DWIL cases tend to be transcribed in at least one tissue in fruitfly. Consistently, UniGene EST data indicate that 59 out of 97 (61%) human DWIL cases are transcribed.

Thus, all these lines of evidence support the functionality of DWIL proteins.

4 DISCUSSION

We highlight the ambiguity of single exon genes with regard to differentiating DNA- and RNA-mediated duplications. Our results show that one conventional pipeline for identifying retrogenes as intronless genes with intron-containing paralogs ought to be revised. The existence of intron-containing paralogs does not directly ensure that intron loss has occurred. In order to study the origination mechanism of such duplicates, it is necessary to check whether or not the particular alignment spans exon–exon junctions. For the cases without intron loss, although we cannot give a quantitative estimate of the number of DNA-mediated duplicates, our analysis suggest that a significant number of previously identified candidate retrogenes have been miscategorized. Furthermore, we might expect that DNA-mediated duplication plays a large role in generating intronless duplicates given that it occurs at a much higher rate relative to retroposition (Zhang *et al.*, 2010a,b; Zhou *et al.*, 2008).

Our comprehensive survey also validates the efficiency of protein-alignment based pipeline searching for retrogenes (Bai *et al.*, 2007; Emerson *et al.*, 2004). Such a strategy captured the majority of duplications with intron losses since transcript-based alignment only added one out of the 68 cases in fruitfly and 7 out of 188 cases in humans.

Third, our result indicates the difficulty in differentiating processed pseudogenes and duplicated pseudogenes. Given the lack of evolutionary constraint, it is even more challenging to identify intron loss events in the case of pseudogenes. Therefore, it would be even more arbitrary to classify these two kinds of pseudogenes.

Fourth, it is interesting to ask how a new open reading frame emerges given that DWIL cases are often partial duplications of parental proteins. In the case of CG33797 (Fig. 1C), the original stop codon of the parental protein is absent. The genomic alignment across multiple *Drosophila* genomes (Supplementary Fig. S5A) suggests that a new stop codon (TAA) emerged by *de novo* mutations. Similarly, CG18563 recruited a new start codon ATG by *de novo* mutations (Supplementary Fig. S5B). These cases suggest that DWIL cases can explore novel protein sequence space.

Finally, it should be pointed out that retrogenes do not necessarily lack introns. By recruiting flanking sequences, they can generate new introns usually in untranslated regions (Brosius, 1999a; Wang *et al.*, 2004). Fortunately, this feature does not affect the efficacy of the protein search-based retrogene identification strategy. Moreover, initial retrogenes can be subsequently amplified by DNA-level duplication (Brosius, 1999a; Wang *et al.*, 2004) and the ancestor of the present DNA-based genomes is RNA (Brosius, 1999b). From this standpoint, it is difficult to generate a clear-cut differentiation between DNA-level duplication and retroposition across all evolutionary time scales. However, our work provides for the first time a rough estimate of the number of single-exon genes with a multiple-exon paralogs that were generated by the DNA-based mechanism in recent evolutionary history.

Funding: The National Institutes of Health (R0IGM078070-01A1 and R01GM078070-03S1 to M.L., T32 GM007197 to B.H.K.); the National Science Foundation (MCB-1026200 to M.L.); the Chicago Biomedical Consortium with support from The Searle Funds at The Chicago Community Trust.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3390.

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

Betran,E. *et al*. (2002) Retroposed new genes out of the X in Drosophila. *Genome Res.*, **12**, 1854–1859.

Brosius,J. (1999a) Many G-protein-coupled receptors are encoded by retrogenes. *Trends Genet.*, **15**, 304–305.

Brosius,J. (1999b) Transmutation of tRNA over time. *Nat. Genet.*, **22**, 8–9.

Brosius,J. (2003) The contribution of RNAs and retroposition to evolutionary novelties. *Genetica*, **118**, 99–115.

Chintapalli,V.R. *et al*. (2007) Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nat. Genet.*, **39**, 715.

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

Dai,M. *et al*. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.

Emerson,J.J. *et al*. (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science*, **320**, 1629–1631.

Emerson,J.J. *et al*. (2004) Extensive gene traffic on the mammalian X chromosome. *Science*, **303**, 537–540.

Fan,C. *et al*. (2008) The subtelomere of Oryza sativa Chromosome 3 short arm as a hot bed of new gene origination in rice. *Mol. Plant*, **1**, 839–850.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics, *Genome Biol.*, **5**, R80.

Kaessmann,H. (2010) Origins, evolution, and phenotypic impact of new genes, *Genome Res.*, **20**, 1313–1326.

Kaessmann,H. *et al*. (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.*, **10**, 19–31.

Kent,W.J. *et al*. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes, *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.

Kuhn,R.M. *et al*. (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.

Schwartz,S. *et al*. (2003) Human-Mouse Alignments with BLASTZ. *Genome Res.*, **13**, 103–107.

Stajich,J.E. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Suyama,M. *et al*. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609–W612.

Vibranovski,M.D. *et al*. (2009) General gene movement off the X chromosome in the Drosophila genus. *Genome Res.*, **19**, 897–903.

Wang,W. *et al*. (2004) Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species, *Nat. Genet.*, **36**, 523–527.

Wheeler,D.L. *et al*. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.*, **24**, 1586-1591.

Zhang,Y. *et al*. (2007) NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res*, **35**, D156–D161.

Zhang,Y. *et al*. (2009) Positive selection for the male functionality of a co-retroposed gene in the hominoids. *BMC Evol Biol*, **9**, 252.

Zhang,Y.E. *et al*. (2010a) Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Res.*, **20**, 1526–1533.

Zhang,Y.E. *et al*. (2010b) Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol*, **8**, e1000494.

Zhou,Q. *et al*. (2008) On the origin of new genes in Drosophila, *Genome Res.*, **18**, 1446–1455.