



LTR retrotransposons create transcribed retrocopies in metazoans

Shengjun Tan, Margarida Cardoso-Moreira, Wenwen Shi, et al.

Genome Res. published online October 20, 2016

Access the most recent version at doi:[10.1101/gr.204925.116](https://doi.org/10.1101/gr.204925.116)

P<P	Published online October 20, 2016 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution-NonCommercial 4.0 International license), as described at http://creativecommons.org/licenses/by-nc/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Comprehensive immune receptor profiling.
Discover the **DriverMap™ AIR Assay** difference.



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

1 **LTR retrotransposons create transcribed retrocopies in metazoans**

2 Shengjun Tan¹, Margarida Cardoso-Moreira^{2,3}, Wenwen Shi¹, Dan Zhang^{1,4}, Jiawei Huang^{1,4},
3 Yanan Mao¹, Hangxing Jia^{1,4}, Yaqiong Zhang¹, Chunyan Chen^{1,4}, Yi Shao^{1,4}, Liang Leng¹,
4 Zhonghua Liu⁵, Xun Huang⁵, Manyuan Long⁶, Yong E. Zhang^{1,4,7}

5 ¹Key Laboratory of Zoological Systematics and Evolution & State Key Laboratory of
6 Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy
7 of Sciences, Beijing 100101, China.

8 ²Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland

9 ³Present address: Zentrum für Molekulare Biologie der Universität Heidelberg (ZMBH),
10 69120 Heidelberg, Germany.

11 ⁴University of Chinese Academy of Sciences, Beijing 100049, China

12 ⁵State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and
13 Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

14 ⁶Department of Ecology and Evolution, The University of Chicago, Chicago 60637, IL, USA.

15 ⁷Corresponding author. E-mail: zhangyong@ioz.ac.cn.

16 **Running title:** LTR retrotransposon mediated retroposition

17 **Keywords:** LTR retrotransposon, retrocopy, template switch, microsimilarity, exon shuffling

18

19

20 **Abstract**

21 In a broad range of taxa, genes can duplicate through an RNA intermediate in a process
22 mediated by retrotransposons (retroposition). In mammals, L1 retrotransposons drive
23 retroposition but the elements responsible for retroposition in other animals have yet to be
24 identified. Here, we examined young retrocopies from various animals which still retain the
25 sequence features indicative of the underlying retroposition mechanism. In *Drosophila*
26 *melanogaster* we identified and *de novo* assembled 15 polymorphic retrocopies and found
27 that all retroposed loci are chimeras of internal retrocopies flanked by discontinuous LTR
28 retrotransposons. At the fusion points between the mRNAs and the LTR retrotransposons we
29 identified shared short similar sequences that suggest the involvement of microsimilarity-
30 dependent template switches. By expanding our approach to mosquito, zebrafish, chicken and
31 mammals, we identified in all these species recently originated retrocopies with a similar
32 chimeric structure and shared microsimilarities at the fusion points. We also identified
33 several retrocopies that combine the sequences of two or more parental genes, demonstrating
34 LTR-retroposition as a novel mechanism of exon shuffling. Finally, we found that LTR-
35 mediated retrocopies are immediately co-transcribed with their flanking LTR
36 retrotransposons. Transcriptional profiling coupled with sequence analyses revealed that the
37 sense-strand transcription of the retrocopies often lead to the origination of in-frame proteins
38 relative to the parental genes. Overall, our data show that LTR-mediated retroposition is
39 highly conserved across a wide range of animal taxa and combined with previous work from
40 plants and yeast that it represents an ancient and ongoing mechanism continuously shaping
41 gene content evolution in eukaryotes.

42

43

44 **Introduction**

45 Retrotransposons are pervasive in eukaryotic genomes and consist of two major classes: long
46 terminal repeats (LTRs) and non-LTR retrotransposons (mainly long interspersed nuclear
47 elements or LINEs). In each class, they can be further divided into autonomous and non-
48 autonomous retrotransposons. Autonomous retrotransposons encode the proteins required for
49 their own mobilization, while non-autonomous retrotransposons are trans-mobilized by the
50 autonomous elements. For example, in mouse non-autonomous *MaLR* elements can be
51 copied by the autonomous *MERVL* elements since they share similar LTRs (McCarthy and
52 McDonald 2004). A key contribution of retrotransposons to evolution is their capacity to
53 mediate the formation of new genes, in a process termed retroposition (Soares et al. 1985;
54 Brosius 1991; Eickbush 1999). Herein, the retrotransposon machinery reverse transcribes
55 messenger RNAs (mRNAs) into complementary DNAs (cDNAs) and inserts them back into
56 the genome as retrocopies (Kaessmann et al. 2009). Unlike DNA-based duplicates, functional
57 retrocopies or retrogenes inherently lack the ancestral regulatory sequences, having to recruit
58 new ones from the vicinity and/or by evolving them *de novo* (Kaessmann et al. 2009; Carelli
59 et al. 2016). The contribution of retrogenes to phenotypic evolution is well recognized
60 ranging from the avoidance of male-male courtship in fruitfly to the short-legged phenotype
61 in domestic dogs (Parker et al. 2009; Chen et al. 2013).

62 Currently, our understanding of the mutational mechanisms underlying retroposition is
63 largely limited to mammalian systems where it has been conclusively shown that retrocopies
64 can be generated by the reverse transcriptase (RT) of LINE1 or L1 retrotransposons
65 (Eickbush 1999; Moran et al. 1999; Esnault et al. 2000; Wei et al. 2001). L1-mediated
66 retrocopies harbor several hallmark sequences in their flanking regions: a TTTT/AA
67 endonuclease cleavage site, a poly(A) tail priming the reverse transcription, target site
68 duplications (TSDs) generated when fixing the insertion nick overhang, and a bias favoring
69 truncation of 5' than 3' end (Kaessmann et al. 2009; Richardson et al. 2014).

70 In principle, LTR retrotransposons could also mediate the formation of retrocopies.
71 Specifically, LTR retrotransposons have a close evolutionary relationship with retroviruses
72 (Wicker et al. 2007), which replicate by undergoing two rounds of template switches during
73 their viral DNA synthesis, and use a low affinity RT enzyme (Delviks-Frankenberry et al.
74 2011). Retroviruses can capture host mRNAs (Hajjar and Linial 1993) and their RT can use

the mRNA as a template through template switching induced by microsimilarity (between short similar sequences shared by the mRNA and the retrovirus, also called microhomology) (Goodrich and Duesberg 1990; Luo and Taylor 1990). Sometimes sequence-similarity independent mechanisms (e.g., 3' transduction) (Swain and Coffin 1992) may also work. Thus, it is conceivable that mRNAs could be captured in the virus-like particles (VLP) where LTR retrotransposons replicate (Yoshioka et al. 1990; Havecker et al. 2004) and that the latter could mediate the retroposition of these mRNAs in a manner analogous to that of retroviruses. The evidence for this process is, however, very limited. In yeast, reporter assays demonstrated that the LTR retrotransposon *Ty* can capture partial mRNA sequences and form a chimera consisting of internal retrocopies flanked by the retrotransposon, and that this is often accompanied by microsimilarity at the fusion points (Derr et al. 1991; Schacherer et al. 2004). In maize, a chimera consisting of partial sequences derived from three unlinked cellular genes flanked by the LTR retrotransposon *Bs1* has been described (Elrouby and Bureau 2001; Elrouby and Bureau 2010), and in rice 27 retrocopies were found to be located within LTR retrotransposons (Wang et al. 2006). Finally, a single case study has been described in human where the LTR retrotransposon *HERV-K* captured the gene *FAM8A1* (with microsimilarity present at the fusion points) resulting in a currently untranscribed processed pseudogene (Jamain et al. 2001).

Here we performed a genome-wide study across a broad range of taxa (fruitfly, mosquito, zebrafish, chicken, mouse and human) to determine whether LTR retrotransposons can generally drive retroposition in animals and to identify the features associated with LTR-mediated retrocopies.

Results

***De novo* assembly reveals a conservative but highly reliable dataset of 15 polymorphic retrocopies in the *Drosophila melanogaster* Genetic Reference Panel (DGRP)**

We started our analyses by focusing on polymorphic retrocopies, which are encoded by at least one inbred line but are absent from the reference genome. Specifically, we analyzed the 40 core DGRP lines whose average sequencing depth reaches 52x (Supplemental Table 1). Based on previous work (Schrider et al. 2011; Ewing et al. 2013; Schrider et al. 2013), we developed a refined pipeline to identify polymorphic retrocopies that combines both exon-

exon and exon-intron junction read mapping, and that includes targeted local *de novo* assembly (Supplemental Fig. 1, Methods).

We identified a total of 31 candidate retrocopies supported by more than one read. Of these, we assembled 15 retrocopies and their flanking regions (Supplemental Dataset 1, Table 1 and Supplemental Table 2). The remaining 16 candidate retrocopies could not be assembled because of the low sequencing depth of the retroposed loci (for 12 candidate retrocopies, Supplemental Table 3) or because of mapping problems (for 4 retrocopies). We chose to focus on the set of 15 retrocopies because knowing their flanking regions is critical to infer the underlying mutational mechanism. For the 10 out of 15 retrocopies for which we could design probes, we confirmed their existence by PCR experiments (see Methods, Supplemental Fig. 2 and Supplemental Table 4). We further randomly chose three retrocopies for Sanger sequencing, and confirmed the assembled sequences (Supplemental Dataset 2).

Based on the presence/absence of exon-exon junction reads across the 40 lines, we estimate the lower-bound population frequencies of the 15 retrocopies to range from 1 to 31 (Supplemental Table 3). Notably, for both *CG4799_CG11924_r* and *CG5119_CR42443_r*, the sequences of two parental genes were retroposed together generating new chimeric gene structures, each containing the partial sequences of two unlinked genes. In total, 17 parental genes were involved in the formation of 15 polymorphic retrocopies (Table 1).

LTR retrotransposons mediate retroposition in *Drosophila* populations

The 15 retrocopies tend to be partial with a median length of 26% of their parental genes' size, a consequence of both 5' and 3' truncations with the latter being more extensive (median truncation 419 bp vs 1,387 bp, Wilcoxon rank test $P = 0.04$; Supplemental Table 2). In addition to the lack of the 5' truncation bias, all other hallmarks suggestive of L1-mediated retroposition (i.e., presence of a poly(A) tail, TSDs and a recognition site) are absent from the 15 retrocopies. When we aligned the retrocopies' flanking sequences against the reference genome, we found that all flanking sequences consisted of LTR retrotransposons (Table 1). Most notably, this association was not created by the insertion of the retrocopies into preexisting LTR retrotransposons but instead by recombination events mediated by microsimilarity between the retrocopies and the LTR retrotransposons. Specifically, no retrotransposons flanking the retrocopies showed the signature of a previously continuous element that was interrupted by the insertion of a foreign sequence. Instead, in all cases, a

portion of the internal retrotransposon open reading frames (ORFs) was replaced by the retrocopies (Supplemental Dataset 1). Moreover, for 25 of the 29 (86%) junctions identified, we found stretches of microsimilarity shared between the LTR retrotransposons and the mRNAs (median length of 6 bp, Table 1 and Supplemental Dataset 1). For three junctions, *de novo* sequences (2-26 bp) were found inserted between the two elements. Simulations show that the extent of microsimilarity cannot be explained by the random insertion of retrocopies within LTR retrotransposons (Supplemental Fig. 3 and Methods).

Thus, retroposition in *Drosophila* seems to occur via a microsimilarity-dependent template switch mechanism of LTR retrotransposons, which is analogous to the one reported for retroviruses and yeast (Goodrich and Duesberg 1990; Luo and Taylor 1990; Derr et al. 1991; Hajjar and Linial 1993; Schacherer et al. 2004). One implication of this mechanism is that both 5' and 3' LTRs should co-exist in the retroposed product because the integrase has to recognize 15-20 bp sequences located at each terminal (Hindmarsh and Leis 1999; Engelman et al. 2009). Consistently, we found in the assembled sequences of nine (60%) retrocopies the presence of the 5' LTR (the data are unclear for the other retrocopies because the assembled sequences are too short). In addition, for five retrocopies, for which we obtained sequences at the two terminals, we confirmed the presence of both 5' and 3' LTRs (Supplemental Datasets 3-5, Methods). Although various families of LTR retrotransposons are associated with the retrocopies (Table 1), for the five retrocopies associated with the *Gypsy BLASTOPIA*, we found a unique pattern: the RT always switches back from the mRNA to the retrotransposon at one particular site: TAAAAACAAGC-GGTTG, which is always near an exon-exon junction (Fig. 1A and Supplemental Dataset 1). Our data also show that more than two switches can occur. For example, a segment from the 5' untranslated region (UTR) of *CG4799* and a segment from the coding sequence (CDS) of *CG11924* were co-retroposed by the *Pao BATUMI* element to form a new chimeric retrocopy, where all three switch points are associated with stretches of microsimilarity (Fig. 1B). Thus, two or more genes can be involved in this process to form a chimeric gene, which could be viewed as a new mechanism of exon shuffling (Gilbert 1978; Long et al. 1995; Li 1997).

In *Drosophila*, 44% of recently inserted LTR retrotransposons are located within or nearby genes, which is an overrepresentation compared to a random insertion model (Ganko et al. 2006). One would expect that retrocopies created by LTR retrotransposons should show a similar pattern. We could not test this hypothesis with the assembled retrocopies because they

do not include the sequences outside the flanking LTR retrotransposons. So, we generated whole-genome sequencing data with longer read sizes and mobile element targeted sequencing data via the ME-Scan methodology (Witherspoon et al. 2010) (Supplemental Table 5, Methods). With these additional data we identified 18 insertion sites for 9 retrocopies (Table 2 and Supplemental Table 6). Twelve insertion sites were further validated by PCR followed by Sanger sequencing. In agreement with the biased insertion pattern of LTR retrotransposons, most (8/12) retrocopies were inserted within genes.

Intriguingly, two retrocopies (*CG17604_r* and *CG10212_r*) are associated with more than one insertion site (Table 2), even though they appear to be derived from a single template switch event because they share the same gene structure. For *CG17604_r* the evidence for a single template switch event is even stronger because the different retrocopies share eight nucleotide differences that distinguish them from the parental gene (Supplemental Dataset 6). Considering that more than 80% of DNA-based segmental duplication (SD) occur in tandem in *D. melanogaster* (Zhou et al. 2008), the presence of multiple identical retrocopies located dispersedly in the genome is unlikely to be a consequence of independent SDs. One possible scenario is that the whole locus acts as a non-autonomous LTR retrotransposon that can be further retroduplicated by the enzymes of autonomous LTR retrotransposons. This scenario is not unlikely considering that non-autonomous LTR retrotransposons are much more likely to proliferate than their corresponding autonomous elements, presumably because the former are less likely to be recognized as retroviruses and thus repressed by the host machinery (Maksakova et al. 2006).

In summary, we have described three lines of evidence that identify LTR retrotransposons as capable of mediating retroposition in *Drosophila*: 1) all polymorphic retrocopies are flanked by LTR retrotransposons and the switch points between the LTR retrotransposons and the retrocopies share stretches of microsimilarity suggesting the occurrence of microsimilarity-mediated recombination; 2) the insertion sites are mostly genic, the preferential insertion site of LTR retrotransposons; and 3) these new loci show signs of acting as non-autonomous LTR retrotransposons, being subjected to further retroposition cycles.

LTR retrotransposons mediate retroposition across animals

Since *Drosophila* polymorphic retrocopies show the signatures of LTR-mediated retroposition, we also expect to find these features in recently originated retrocopies encoded

by the *D. melanogaster* reference genome. By modifying the strategy in (Bai et al. 2007) (Methods), we identified five retrocopies present in the reference genome but absent from the closely related *D. simulans* and *D. sechellia* (Table 3 and Supplemental Dataset 7). The youngest, *CG33932_r* and *CG16728_r*, are flanked by LTR retrotransposons and *CG16728_r* additionally has microsimilarities at both junctions. Consistently, no hallmarks of L1-mediated retroposition exist in the flanking regions of these two retrocopies (Supplemental Table 7). In contrast, the relatively older retrocopies *CG12324* and *CR12628*, are not flanked by LTR retrotransposons and are instead associated with candidate poly(A) tails and/or TSDs (Toba and Aigaki 2000) suggesting a non-LTR mediated retroposition. The final retrocopy (*CG40354*) is ambiguous because it harbors a 2 bp microsimilarity and a 3' truncation (121 bp) but also candidate TSDs (Supplemental Fig. 4A and Supplemental Dataset 7). Since this locus situates in a repeat-rich region that does not have a long synteny with closely related species (Supplemental Fig. 4B), we could not infer if it was formed through an LTR-mediated or non-LTR mediated retroposition followed by secondary mutations.

However, no LTR retrotransposon remnants were found flanking 96 old retrogenes which predate the species split of *D. melanogaster* and *D. simulans* (Bai et al. 2007). Similarly, only three retrogenes (*CG1924*, *CG13732* and *CG5650*) are associated with either candidate poly(A) tails or TSDs (Supplemental Table 7). The absence of the sequence features diagnostic of the type of retroposition event is expected for old retrogenes. It is a consequence of the rapid removal of flanking hallmark sequences by the strong deletion bias of *D. melanogaster* (Petrov and Hartl 1998) and by the accumulation of substitutions after the retroposition event.

Because LTR retrotransposons are widely shared across species (Havecker et al. 2004; Eickbush and Jamburuthugoda 2008), we searched additional animal genomes for signs of LTR-mediated retroposition. Specifically, we applied the pipeline used to identify recently originated retrocopies in *Drosophila* to five other species (mosquito, zebrafish, chicken, mouse and human) representing a broad taxonomic breadth (Methods). We found retrocopies potentially generated by LTR retrotransposons in all species. In mosquito, zebrafish and chicken we identified 3, 7 and 10 retrocopies (respectively) still retaining sequence features indicative of the underlying retroposition mechanism (Supplemental Table 8). We found that LTR retrotransposons are responsible for ~50% of retrocopies (1/3 in mosquito, 5/7 in zebrafish and 4/10 in chicken). One example is the mosquito retrocopy

ENSANGG00000011308, which is flanked by the LTR retrotransposon *BEL18_AG*. It is associated with 5 bp and 7 bp microsimilarity stretches at the switch points (Fig. 1C), an extensive 3' truncation (~1 kb), and no signs of a poly(A) or TSDs (Supplemental Table 8).

In mammals, L1 retrotransposons are known to be a major driver of retroposition (Esnault et al. 2000; Kaessmann et al. 2009; Abyzov et al. 2013), and have generated thousands of retrocopies (Carelli et al. 2016). We implemented multiple filters (Methods) to identify the possibly small number of retrocopies created by LTR retrotransposons. We identified 3 retrocopies in human (including the previously described *FAM8A1_r* (Jamain et al. 2001)) and 8 retrocopies in mouse that are flanked by LTR retrotransposon remnants and are not associated with any L1 hallmark sequences (Supplemental Table 8). The sequence structures of these retrocopies match those observed in *Drosophila*, mosquito, zebrafish and chicken. For example, in mouse a partial transcript (549/1646 bp) of *ENSMUSG00000021752* was retroposed by the LTR retrotransposon *IAPeY*, mediated by 10 bp of microsimilarity and 2 bp of *de novo* sequence (Fig. 1D). In an exceptional example, we identified a “Super” retrocopy where mRNAs derived from four distinct genes and one *Alu* element were co-retroposed into a single locus (Supplemental Table 8 and Supplemental Fig. 5). Overall, despite the older age of the mouse retrocopies, we were still able to identify stretches of microsimilarity for 9 of the 22 switch points. Unlike in *Drosophila* where most retrocopies are located within genes (Table 2), in mouse 7 of the 8 retrocopies are intergenic (Supplemental Table 8, Fisher’s exact test $P = 0.03$), which is consistent with the well-known depletion of LTR retrotransposons within or in the vicinity of genes in mammals (Jern and Coffin 2008).

LTR retrotransposons drive the initial transcription of retrocopies in both fly and mouse

The prevalence of polymorphic retrocopies with more than one insertion site in *Drosophila* suggests that retrocopies mediated by LTR retrotransposons can be retroposed again as non-autonomous retrotransposons (Table 2). This also implies that the internal retrocopies and flanking retrotransposons are co-transcribed as chimeric transcripts by the regulatory sequences of the LTR retrotransposons. Hence, we tested whether the retroposed loci are transcribed as a single unit and whether the retrocopies share the transcriptional profile of the corresponding LTR retrotransposons. First, we performed RT-PCR on the 10 retrocopies for which we could design primers spanning the switch points (Supplemental Table 9). We detected the expression of all 10 retrocopies plus the flanking LTRs in both male and female

adult bodies (Fig. 2A). In addition, both long range RT-PCR for retrocopy *CG5020_r* and rapid amplification of cDNA ends (RACE) for *CG17604_r*, confirmed the transcription of the whole chimeric unit from the 5' to the 3' LTRs (Supplemental Datasets 3 and 4, Methods). Second, we profiled the expression of the 10 retrocopies and corresponding retrotransposons in testis, ovary and head (Methods). We targeted these three organs because in *Drosophila* a large fraction of retrogenes are testis-specific or testis-dominant (Supplemental Fig. 6) (Betran et al. 2002; Vibranovski et al. 2009), but LTR retrotransposons tend instead to be abundantly transcribed in the head (Perrat et al. 2013). The ovary was included for comparison. We found that 9 of the 10 retrocopies show a strong positive correlation between the expression of the retrocopy and the corresponding LTR retrotransposon with 8 retrocopies showing their highest expression in the head, more than expected by chance (Fig. 2B, Binomial test $P = 0.02$). RT-PCR further confirmed the expression in the head of four retrocopies encoded by line 208 (Fig. 2C). Interestingly, the expression intensity of LTR retrotransposons is generally higher than that of the corresponding retrocopies except for *STALKER2* (*CG5119_CR42443_r*) and *Gypsy5* (*CG3631_r*) (Fig. 2B), which could be explained by the fact that these two retrotransposons have a lower copy number in the genome relative to the others (1-3 vs 4-27, Wilcoxon rank test one-tailed $P = 0.03$; Supplemental Table 10). This also suggests that retrocopies as non-autonomous retrotransposons could be less repressed than autonomous elements (Maksakova et al. 2006).

The transcriptional characteristics of the mouse LTR-mediated retrocopies are similar to those of *Drosophila* with respect to both the transcription structure and expression profile. Specifically, we found ENCODE RNA-sequencing (RNA-seq) reads (Yue et al. 2014) spanning the switch points between the LTR retrotransposons and the retrocopies for 4 of the 8 LTR-mediated retrocopies suggesting that they are transcribed as a single transcript (Supplemental Table 11 and Supplemental Fig. 7, Methods). In the case of the *Super* retrocopy, long-range RT-PCR experiments showed that the fragments derived from the four different parental genes are transcribed together, also suggesting the whole locus represents a single unit (Supplemental Fig. 8). Furthermore, unlike in the fly where LTR retrotransposons tend to be broadly transcribed (Supplemental Fig. 9), LTR retrotransposons in mouse are generally tightly controlled at the transcriptional level and tend to be tissue-specific with a strong bias towards early embryogenesis (Maksakova et al. 2006; Faulkner et al. 2009; Macfarlan et al. 2012). Consistently, 4 of the 8 retrocopies show the highest expression at the

two-cell or eight-cell stages across 17 tissues/developmental stages, which is not expected by chance (Binomial test $P = 0.009$, Fig. 2D and Supplemental Table 11). Actually, 4 could be an underestimation given that genes without high transcriptional levels (Fragments Per Kilobase of exon per Million fragments mapped, FPKM > 100) like LTR-mediated retrocopies (Supplemental Table 11) tend to be neglected in single-cell RNA-seq data (Marinov et al. 2014). Overall, these lines of evidence in mouse again support the hypothesis that the expression of LTR-mediated retrocopies is driven by the LTR retrotransposons' regulatory sequences.

LTR-mediated retrocopies are more often transcribed from the same strand as the parental genes and capable of generating in-frame proteins

Since LTRs (at the two ends) are known to have the potential to act as bidirectional promoters (Feuchter and Mager 1990; Dunn et al. 2006; Faulkner et al. 2009), it is unclear whether the transcription of the retrocopies occurs from the sense strand relative to the parental genes. In *Drosophila*, a RACE experiment demonstrated that *CG17604_r* is transcribed from the sense strand (Supplemental Dataset 4). In order to identify the direction of transcription for more retrocopies, we generated strand-specific RNA-seq data for testis, ovary and head for line 208 (Supplemental Table 5 and Supplemental Fig. 10, Methods). Given the high sequence identity between polymorphic retrocopies and their parental genes, we evaluated the strandedness using only the reads spanning the switch points. We did not find any spanning reads for *CG17604_r* and *CG33957_r*, but we detected reads supporting the transcription of *CG14796_r* and *CG4799_CG11924_r* across all three tissues (Fig. 3A). Specifically, we found that *CG14796_r* is transcribed from the antisense strand in head, but from the sense strand in ovary. In agreement with the relatively low expression of *CG14796_r* in testis (Fig. 2B), we only found weak expression in this tissue. *CG4799_CG11924_r* is bidirectionally transcribed in ovary and head, and additionally shows very high levels of sense transcription in testis (Fragment count Per Million fragments mapped, FPM = 14,111), a result also consistent with the qRT-PCR result (Fig. 2B). Thus, the RACE and RNA-seq experiments showed that 3 out of 4 retrocopies are capable of transcription from the sense strand relative to the parental genes.

Analogously, by taking advantage of public strand-specific RNA-seq data from mosquito, chicken, zebrafish and mouse (Methods), we found that recently evolved retrocopies in these species are also transcribed more often or more strongly from the sense strand (relative to

their parental genes). Overall, 12 out of 18 retrocopies are moderately transcribed (FPKM > 0.1) from the sense strand in at least one tissue, while only 6 reach the same intensity from the antisense strand (Fig. 3).

The transcription of the retrocopies from the same strand as the parental genes suggests the possibility that the retrocopies could lead to in-frame translation. In support of this idea, for 14 (82%) of the 17 *Drosophila* polymorphic or recently evolved retrocopies mediated by LTR-retrotransposons, the predicted longest ORF is in-frame relative to the parental proteins (Supplemental Table 12) possibly because they are too young (nucleotide divergence < 1%, Table 3) to have accumulated loss-of-function mutations. For relatively older retrocopies in other animals (nucleotide divergence $\geq$ 1%, Supplemental Table 8), 14 out of 21 have accumulated at least one loss-of-function mutation (Supplemental Dataset 8). We examined directly the level of constraint at the protein-level by calculating K_A/K_S values (the ratio between the non-synonymous and the synonymous substitution rates) for the longest ORF predicted relative to the parental protein. We found that *ENSDARG00000076317_r* from zebrafish and *ENSGALG00000011896_r* from chicken have a K_A/K_S ratio of 0.145 and 0.239, respectively, which is significantly lower than the conservative cutoff of 0.5 and suggestive of constraint ($P = 0.041$ and 0.004 , Supplemental Table 13) (Betran et al. 2002). *ENSANGG00000011308* from mosquito also reached a marginal significance ($K_A/K_S = 0.261$, $P = 0.075$). Therefore, these retrocopies, together with genes like *Bs1* from maize (Elrouby and Bureau 2001; Elrouby and Bureau 2010), suggest that a fraction of LTR mediated retrocopies can encode functional proteins.

Discussion

LTR-mediated retroposition is deeply conserved across eukaryotes

By looking at young retrocopies in a broad range of taxa we showed that retrogenes can be created in both invertebrates and vertebrates through an LTR-mediated retroposition mechanism. Our data suggest a model where the RT starts synthesizing cDNA using the template from the RNA of the LTR retrotransposon, but induced by a stretch of microsimilarity between the LTR retrotransposon and a cellular mRNA captured within the VLP, the enzyme moves to the latter and takes the cellular mRNA as the template (Fig. 4). A second stretch of microsimilarity between the cellular mRNA and the LTR retrotransposon

leads to another switch, this time back to the LTR retrotransposon to end the reverse transcription process.

This mechanism creates retrocopies with four unique features. First, because the stretches of microsimilarity can occur anywhere, mRNAs are often only partially duplicated (Table 1). Second, because only small stretches of microsimilarity are required, LTR-mediated retroposition can give rise to chimeras involving multiple parental genes. It can therefore be considered as a novel mechanism of gene recombination, including domain- or exon-shuffling to capture the essence whereby the sequences of previously existing genes are combined as novel gene structures (Gilbert 1978; Moran et al. 1999; Graur and Li 2000). Third, a key feature of LTR-mediated transcription (that is notably absent from L1-mediated retroposition) is the immediate transcription of the retrocopies by the LTR's promoters. Fourth, LTR-mediated retrocopies can act as non-autonomous LTR retrotransposons, and be subject to further duplication (Fig. 4).

This mechanism is compatible with the results from the limited studies on LTR-mediated retroposition in plants and yeast. In plants, although the retrocopies described are generally old (Elrouby and Bureau 2001; Wang et al. 2006; Elrouby and Bureau 2010), they show the typical chimeric structure with internal retrocopies flanked by retrotransposons. Specifically, in maize, the gene *Bs1* combines the partial exonic sequences of three parental genes and is flanked by an LTR retrotransposon (Johns et al. 1985; Elrouby and Bureau 2001; Elrouby and Bureau 2010). The copy number of *Bs1* ranges from one to five across various species of the genus *Zea* and all individual copies share the same switch points (Elrouby and Bureau 2010). Thus, *Bs1* likely has undergone secondary retropositions as seen for *CG17604_r* or *CG10212_r* (Table 2). In contrast to these efforts in plants and our work on animals, the work in yeast studied retroposition by inducing or screening artificial retrocopies via cellular assays (Derr et al. 1991; Schacherer et al. 2004; Maxwell and Curcio 2007). Still, these experiments recapitulated the unique features associated with LTR-mediated retroposition such as the presence of microsimilarity and the transcriptional capability (Schacherer et al. 2004; Maxwell and Curcio 2007).

Thus, if we combine our current work in animals with previous work on individual genes in plants and artificially induced systems in yeast, we see a striking consistence of the features that characterize LTR-mediate retroposition: chimeric structures between partial duplicates and LTR retrotransposons, microsimilarity at the junction points, the capacity for

transcription, and the existence of secondary retroposition events (Fig. 4). Template-switch dependent LTR-mediated retroposition is therefore highly conserved across multiple eukaryotic lineages.

Although LTR-mediated retroposition is active across different lineages, its contribution to the formation of retrogenes varies widely across taxa. Our previous work in plants has found that LTR and non-LTR-mediated retroposition appear to occur with roughly equal rates in dicots (*Arabidopsis thaliana* and *Manihot esculenta*) (Zhu et al. 2016). In this work, we confirm that L1-mediated retroposition is the dominant mechanism of retroposition in human and mouse (Kaessmann et al. 2009). In mosquito, zebrafish and chicken the contribution is more balanced with each contributing roughly equally to the formation of retrocopies. Although in *Drosophila* we found similar numbers of recently originated retrocopies created by both mechanisms, all polymorphic retrocopies were associated with LTR mediated events. Through simulations (Supplemental Table 14, Methods) we show that our pipeline works equally well detecting LTR-mediated and non-LTR mediated retroposition. Therefore, the enrichment of LTR-mediated retroposition among polymorphic retrocopies when compared to those recently originated (Fisher's exact test $P = 0.04$) likely reflects the recent increase in activity of LTR retrotransposons in flies (Bergman and Bensasson 2007; Kofler et al. 2015).

The different evolutionary trajectories of L1-mediated and LTR-mediated retrocopies

L1-mediated retrocopies are long known to be mostly dead-on-arrival because of the absence of the parental genes' regulatory sequences (Mighell et al. 2000; Kaessmann et al. 2009). A recent work identified ~5,000 retrocopies across various placental mammals, and found that only 4% were robustly expressed (FPKM ≥ 1) (Carelli et al. 2016). The expressed retrocopies were found to have evolved their promoters *de novo* or to have recruited proto-promoters from nearby regions (Carelli et al. 2016). In our study, the two recently evolved non-LTR mediated retrocopies in *Drosophila* also demonstrate that the fate of these retrocopies depends on whether a promoter is fortuitously gained. *CG12324* is broadly transcribed likely because it hitchhiked regulatory sequences already present at the insertion site, as suggested by the correlation between its expression and that of its neighboring gene, *CG12323* (Pearson $r = 0.76$, $P = 0.005$) (Supplemental Table 15). The other one, *CR12628*, is not transcribed presumably because of the absence of regulatory sequences at the insertion site (Table 3 and Supplemental Fig. 11). In contrast, we found that LTR-mediated retrocopies are transcribed immediately after the retroposition (Fig. 2). This phenomenon may be restricted to the

fixation stage and to a short period after fixation. Consistently, we only detect reads spanning switch points for four retrocopies encoded by the mouse reference genome suggesting that their co-expression may eventually get decoupled. More importantly, four retrocopies (Supplemental Table 8) encoded by zebrafish and chicken already lost the LTR retrotransposon remnants at one side suggesting that LTR-mediated retrocopies eventually diverge from non-autonomous LTR retrotransposons.

A significant proportion of LTR-mediated retrocopies may play novel functions

Our work shows that LTR retrotransposons can create new transcribed retrocopies. Therefore, from the viewpoint of transcription, the potential of LTR-mediated retroposition to create functional retrocopies is higher than that of L1-mediated retroposition. But are any of these retrocopies actually functional (i.e., bona fide retrogenes)? Although the young age of the retrocopies makes it difficult to test the levels of protein constraint, we have identified two or three retrocopies showing evidence for constraint at the protein-level (Supplemental Table 13).

We should note that translation does not absolutely require that the retrocopies are in the same frame as the parental genes. In the case of *Bs1* in maize, the region derived from one parental gene is out-of-frame, while the regions derived from the other two parental genes are in-frame (Elrouby and Bureau 2001; Elrouby and Bureau 2010). Moreover, loss-of-function mutations can be rescued by the evolution of new introns within the retrocopies (Tan et al. 2014). In addition, evidence has accumulated suggesting the general significance of long noncoding RNAs (lncRNA) (Ponting et al. 2009; Guttman and Rinn 2012; Rinn and Chang 2012). The *Drosophila* retrocopy *CG4799_CG11924_r*, is more highly expressed than 99.5% of the genes transcribed in testis (Supplemental Fig. 12) and has therefore the potential to act as a functional lncRNA if it could not encode a protein. Consistently, we detected reduced levels of nucleotide diversity and increased levels of linkage disequilibrium (LD) around this locus suggesting that positive selection could be driving its spread (Supplemental Fig. 13).

If some LTR-mediated recopies are indeed functional, their chimeric structure implies that they should have a radically different functional role compared to their parental genes. The most extreme examples are those of retrocopies involving several parental genes. *Bs1* in maize shows that these complex chimeric genes can indeed sometimes become functional hybrid proteins (Jin and Bennetzen 1994; Elrouby and Bureau 2001; Elrouby and Bureau

2010). Therefore, it is reasonable to conclude that LTR retrotransposons are capable of creating retrocopies across a wide range of eukaryotes, which could subsequently evolve to be neofunctionalized retrogenes given their novel gene structure and expression profile.

Methods

Public *Drosophila* population resequencing data

We focused on the subset of 40 inbred lines from the DGRP project (NCBI SRA database: SRP000694 and SRP000224), which were also sequenced by the *Drosophila* Population Genomics Project (DPGP) project (Langley et al. 2012).

Terminology

Traditionally, “retrocopy” is used to refer to an intronless copy derived from a host mRNA. This terminology was developed for L1-mediated retrocopies, which generally have no additional flanking sequences. However, for LTR-mediated retroposition, the retroposed locus consists not only of the sequence derived from the host mRNA but also of the flanking LTR retrotransposon sequences. To be consistent with the previous literature we use the term retrocopy to refer only to the host mRNA derived sequence. We refer to retrocopies using the ID of the parental gene followed by “_r” except when the gene is already annotated.

Detection of *Drosophila* polymorphic retrocopies

To create the search library, we extracted exon-exon junction sequences by pooling 2x100 bp from each pair of consecutive exons (Supplemental Fig. 1A). All reads were mapped to this library with Novoalign (<http://www.novocraft.com>) and only uniquely mapping reads were retained (Supplemental Methods). We called candidate retroposition or intron loss events if there were at least two reads spanning one exon-exon junction. Since the introns of parental genes are not deleted by retroposition, we retained 31 candidates with reads mapping to both exon-intron and intron-exon junctions (Supplemental Fig. 1B).

Local assembly and validation of the polymorphic retrocopies

For each candidate retrocopy, we extracted the reads that mapped uniquely to the parental genes, 500 bp flanking regions, and to exon-exon junctions and performed *de novo* local assembly with MIRA (Chevreux et al. 1999) (Supplemental Fig. 1C). MIRA will export at

least two different contigs corresponding to the parental and retrocopy locus (Supplemental Fig. 1D). If the contig encoded by the retrocopy was too short to identify its exact switch points, it was further assembled with PRICE (Ruby et al. 2013). We assembled 15 retrocopies derived from 17 parental genes as well as their flanking regions (10-200 bp) (Supplemental Dataset 1). We validated this dataset by PCR (Schrider et al. 2011; Schrider et al. 2013) in which we used Primer3 (Untergasser et al. 2012) to design primers targeting the exons shared by parental genes and retrocopies that flank at least one intron in the parental genes (Supplemental Table 4). If retrocopies are present we should amplify two PCR products of different sizes, the larger corresponding to the parental gene (introns present) and the smaller to the retrocopy (introns absent).

Evaluation of potential bias in the detection of LTR-mediated and non-LTR mediated retrocopies

We ran simulations approximating the process of retroposition in *Drosophila*. Because there are more full-length LTR retrotransposons in this species (Supplemental Table 16), we randomly selected 105 genes out of the top 600 most highly transcribed genes in testis and ovary as the parental genes to simulate 90 retrocopies mediated by LTR retrotransposons and 15 retrocopies mediated by non-LTR (L1) retrotransposons (Supplemental Table 14). We determined the size of the retroposed sequences, of the replaced regions of the LTR retrotransposons, and of the microsimilarities following the distributions learned from the empirical set of 15 retrocopies. We simulated non-LTR mediated retrocopies with 5' truncation (≤ 500 bp), poly(A) tails (≤ 100 bp) and TSD (≤ 10 bp) based on previous literature (Linhaire and Bergman 2012; Subtelny et al. 2014). We randomly inserted the simulated retrocopies into intergenic or intronic regions. Finally, we simulated sequencing reads with a coverage of 50x and with variable read length as in the DGRP data (Supplemental Table 14). We were able to recover 85/90 simulated LTR-mediated retrocopies and 15/15 simulated L1-mediated retrocopies (Supplemental Table 14 and Supplemental Methods) demonstrating that our pipeline is not biased in the detection of LTR-mediated and non-LTR mediated retrocopies.

Estimation of the non-randomness of the microsimilarity length

To test whether the extent of microsimilarity could have occurred purely by chance, we performed simulations for all 11 *Drosophila* retrocopies consisting of only two rounds of

template switches. For each retrocopy we randomly selected switch points. We selected the size of the retroposed sequences and of the replaced regions of the LTR retrotransposons following the normal distribution learned from our empirical set of retrocopies. For each retrocopy, we generated 100 random samples and recorded the size of the microsimilarity observed at both switch points. Supplemental Fig. 3 shows the distribution of the microsimilarity length at both ends across 100 simulated events. The observed length (the red dot) is an outlier ($P < 0.01$) for 10 retrocopies. The exception is *CG8567_r* but even for this retrocopy, the total length of microsimilarity was only smaller or equal to the value observed in 8 of the 100 simulations ($P = 0.08$). Overall, these data show that the extent of microsimilarity observed between the retrocopies and the LTR retrotransposons cannot be explained by chance.

Inference of the retrocopies' insertion sites

In order to identify the insertion sites, we generated whole-genome sequencing data using longer reads (2x250 bp) for six lines (e.g., 208) and mate-pair sequencing data for line 437 (Supplemental Table 5). We detected heterozygosity in these sequencing data that suggested contamination (admixture) between the lines sequenced. Thus in all analyses these data were treated as pooled data. To confirm that the contamination occurred exclusively during the production of the novel sequencing data, we selected three loci with high SNP density and performed PCR experiments that confirmed that our fly stocks were not contaminated, showing the expected levels of homozygosity (Supplemental Fig. 14). With these whole-genome sequencing data, we were able to extend the assembled retrocopies from a median of 163 bp to 210 bp, and to locate four retrocopies. To identify the genomic insertion sites of more retrocopies, we further performed targeted high-throughput sequencing following the mobile element scanning (ME-Scan) protocol (Witherspoon et al. 2010) and located nine retrocopies (Table 2). Then, by PCR and Sanger experiments, we validated 12 insertion sites for 6 retrocopies (Table 2 and Supplemental Table 6).

We cloned and sequenced the two terminals of five retroposed loci and found that they include LTRs that are required for their integration into the genome (Supplemental Datasets 3-6 and Supplemental Methods).

Identification of retrocopies in reference genomes

In order to detect recently originated retrocopies in the reference genome of *D. melanogaster*, we refined a previous strategy (Bai et al. 2007) and searched the exon-exon junction library against the reference genome with BLASTN (Altschul et al. 1990). We retained only uniquely mapped regions and identified five *D. melanogaster*-specific retrocopies. Similarly, we scanned five other species: mosquito (UCSC assembly anoGam1), zebrafish (danRer7), chicken (galGal4), mouse (mm9) and human (hg19). In mouse, we found 208 retrocopies flanked by LTR retrotransposons on both sides. We excluded retrocopies inserted into preexisting LTR retrotransposons by implementing four filters: 1) the flanking LTR retrotransposons have to belong to the same LTR retrotransposon type or the non-autonomous retrotransposon mated with autonomous retrotransposon (*ORR1/MaLR* with *MERVL*, *ETn* with *MusD*) (McCarthy and McDonald 2004); 2) both the internal mRNAs and the flanking LTR retrotransposons have to be absent from the rat genome (rn4); 3) the LTR retrotransposons have to be discontinuous with a gap > 20 bp; and 4) there are no L1 hallmarks. Analogously, we filtered 518 human retrocopies flanked by LTR retrotransposons with the whole retroposed locus being required to be absent in the non-primate placental mammals. In Supplemental Table 8, only the retrocopies that passed these filters are shown for mouse and human.

To approximate the evolutionary age, we calculated the overall nucleotide divergence between parental gene and retrocopy via BLAT (Kent 2002) for young retrocopies. For old retrogenes reported in (Bai et al. 2007), we directly downloaded the synonymous substitution rate (K_S) values.

Analysis of hallmark sequences of L1-mediated retroposition

We called L1-mediated retrocopies by the presence of one of the following three features: 1) a poly(A) tail at the 3' end; 2) TSDs at both ends; and 3) bias favoring truncation of 5' than 3' ends. The presence of a TTTT/AA endonuclease cleavage site has been shown to not be reliable (Richardson et al. 2014). We inferred the boundary between retrocopies and flanking regions by BLAT. In order to generate a conservative dataset of LTR-mediated retrocopies, we specified relatively loose cutoffs for L1 hallmark sequences: a minimum of 3 continuous As in a 5 bp long sequence at the 3' end, and a minimal TSD not less than 5 bp with an identity $\geq 80\%$. When retrocopies flanked by LTR retrotransposons were also associated with candidate hallmark sequences for L1-mediated retroposition, we determined whether the poly(A) tail and TSDs were part of the LTR retrotransposons, and whether the LTR

retrotransposons were discontinuous given the canonical annotations in Repbase (Kaminker et al. 2002).

Transcriptional profiling of *Drosophila* polymorphic retrocopies

We investigated the expression profile of polymorphic retrocopies using RT-PCR, qRT-PCR, RACE and RNA-seq (Supplemental Methods). For LTR retrotransposons, we designed primers on the basis of the canonical sequence (Supplemental Table 9). To determine the internal controls in qRT-PCR, we followed the protocol in (Gan et al. 2013), and chose the 3 most transcriptionally constant genes (*eIF-1A*, *Rap2l* and *SdhA*) out of 26 in (Ling and Salvaterra 2011; Ponton et al. 2011). Strand-specific RNA-seq data for line 208 were generated by following the dUTP protocol (Levin et al. 2010). Reads were mapped following a Novoalign-based RNA-seq pipeline, with the 15 retrocopies added to the reference sequence set. We only kept the reads spanning the switch points.

Transcriptional profiling of recently originated retrocopies across various animals

The FPKM data of the 96 *Drosophila* retrogenes were downloaded from FlyBase (Marygold et al. 2013). Only the genes *CG5150* and *CG9518* were absent from this dataset. To investigate the expression profiles of the 3 older *Drosophila* retrocopies and of 8 mouse retrocopies, we downloaded the FASTQ-format unstranded modENCODE RNA-seq data (Celniker et al. 2009) and stranded ENCODE mouse transcriptome data (<http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeCshlLongRnaSeq/>). Since LTR retrotransposons are highly active in early mouse embryogenesis (Macfarlan et al. 2012), we also downloaded the single-cell RNA-seq data for two-cell and eight-cell stages from the NCBI GEO database (GSE44183). To better represent the transcriptome diversity, 12 representative fly tissues and 15 mouse tissues were chosen according to a tissue-level clustering analysis (Supplemental Fig. 15). Because retrocopies are highly similar ($\geq 90\%$ identity) to parental genes, we used only uniquely mapping reads and calculated the effective length of the retrocopy defined in (Zhang et al. 2011) when calculating the FPKM values.

Furthermore, we analyzed public strand-specific datasets from mosquito (ArrayExpress ERP006838), zebrafish (Lee et al. 2013) and chicken (Necsulea et al. 2014). We estimated the quality of strandedness based on exon-exon junction spanning reads. Overall, all datasets are reasonably accurate, with $> 89\%$ reads (Supplemental Fig. 10) supporting the standard splicing site “GT-AG”.

In order to determine whether a gene is called as present, we specified two FPKM cutoffs. Because it has been shown that FPKM values from 0.1 to 0.4 could reflect active transcription (Hart et al. 2013), we took the high cutoff of 1 to define robust expression as done in (Carelli et al. 2016) and the low cutoff of 0.1 to define modest expression. For read counts, a cutoff of 0.1 (FPKM) corresponds to at least 9 reads in Fig. 3.

Evolutionary analyses

To examine protein-level constraints, we performed K_A/K_S tests between the parental genes and corresponding retrocopies. By following (Betran et al. 2002), we tested whether K_A/K_S is significantly smaller than 0.5, which suggests functionality for both genes. First, we translated retrocopies by aligning parental proteins against retrocopies with exonerate (Slater and Birney 2005). Then, we aligned the proteins (for retrocopies with candidate loss-of-function mutations, the longest continuous peptide was used) with MAFFT (Kato and Toh 2008) and then converted protein-level alignments into codon-level alignments via PAL2NAL (Suyama et al. 2006). Finally, we performed likelihood-ratio-tests via the codeml program in the PAML package (Yang 2007).

We tested whether *CG4799_CG11924_r* is under positive selection following (Cardoso-Moreira et al. 2016). All SNPs located within 50 kb upstream and downstream of 3L: 21,967,546 (the insertion site, Table 2) were extracted from the DGRP variant annotation file (VCF) using VCFtools (Danecek et al. 2011). We then calculate levels of LD (as measured by r^2) and estimated diversity levels using VCFtools (Danecek et al. 2011).

Data access

The sequence data generated for this study have been submitted to the NCBI BioProject database (<http://www.ncbi.nlm.nih.gov/bioproject/>) under accession numbers PRJNA282433, PRJNA285001, PRJNA325420 and PRJNA325579, and Genome Sequence Archive (GSA; <http://gsa.big.ac.cn/>) under accession numbers PRJCA000123, PRJCA000124, PRJCA000279 and PRJCA000280 (Supplemental Table 5). Sanger trace files have been submitted to the NCBI Trace Archives (<http://www.ncbi.nlm.nih.gov/Traces/home/index.cgi>) with TI numbers in (2344039996-2344040077).

Acknowledgements

We would like to thank three reviewers for their comprehensive and insightful comments. We thank the DGRP and DPGP groups for sharing the genomic data. We thank Bin He, Yang Shen and Zhang lab members for helpful discussions. We also thank Wanzhu Jin for providing mouse samples. This research was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13010400), the National Key Basic Research Program of China (2015CB943001, 2013CB531202), the National Natural Science Foundation of China (91331114, 31322050) and the 1000 Young Talents Program to Y.E.Z. M.C.M. was supported by a Marie Curie incoming postdoctoral fellowship. Computing was jointly supported by the HPC Platform in BIG, CAS and by HPC Platform, Scientific Information Center, Institute of Zoology, CAS, China.

Author Contributions

Y.E.Z. conceived and designed the projects. S.T., C.C, S.Y. and L.L. performed computational analyses. S.T., W.S., D.Z., J.H., Y.M., H.J., Y.Z., Z.L. and X.H. performed the experiments. S.T., Y.E.Z., M.C.M. and M.L. analysed the data. Y.E.Z., S.T., M.C.M. and M.L. wrote the paper.

References

- Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L, Genomes Project C, Lee C, Gerstein M. 2013. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res.* **23**: 2042-2052.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.
- 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.
- Bergman CM, Bensasson D. 2007. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* **104**: 11340-11345.

- 662 Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*.
663 *Genome Res.* **12**: 1854-1859.
- 664 Brosius J. 1991. Retroposons--seeds of evolution. *Science* **251**: 753.
- 665 Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016.
666 Evidence for the fixation of gene duplications by positive selection in *Drosophila*.
667 *Genome Res.* doi:10.1101/gr.199323.115.
- 668 Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history
669 of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* **26**:
670 301-314.
- 671 Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai
672 EC, Lieb JD, MacAlpine DM et al. 2009. Unlocking the secrets of the genome.
673 *Nature* **459**: 927-930.
- 674 Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat. Rev.*
675 *Genet.* **14**: 744-744.
- 676 Chevreux B, Wetter T, Suhai S. 1999. Genome Sequence Assembly Using Trace Signals and
677 Additional Sequence Information. In *German Conference on Bioinformatics*, pp. 45-
678 56, Hannover, Germany.
- 679 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter
680 G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools.
681 *Bioinformatics* **27**: 2156-2158.
- 682 Delviks-Frankenberry K, Galli A, Nikolaitchik O, Mens H, Pathak VK, Hu W-S. 2011.
683 Mechanisms and Factors that Influence High Frequency Retroviral Recombination.
684 *Viruses* **3**: 1650-1680.
- 685 Derr LK, Strathern JN, Garfinkel DJ. 1991. RNA-mediated recombination in *S. cerevisiae*.
686 *Cell* **67**: 355-364.

- 687 Dunn CA, Romanish MT, Gutierrez LE, van de Lagemaat LN, Mager DL. 2006.
688 Transcription of two human genes from a bidirectional endogenous retrovirus
689 promoter. *Gene* **366**: 335-342.
- 690 Eickbush T. 1999. Exon Shuffling in Retrospect. *Science* **283**: 1465-1467.
- 691 Eickbush TH, Jamburuthugoda VK. 2008. The diversity of retrotransposons and the
692 properties of their reverse transcriptases. *Virus Res.* **134**: 221-234.
- 693 Elrouby N, Bureau TE. 2001. A novel hybrid open reading frame formed by multiple cellular
694 gene transductions by a plant long terminal repeat retroelement. *J. Biol. Chem.* **276**:
695 41963-41968.
- 696 Elrouby N, Bureau TE. 2010. Bs1, a New Chimeric Gene Formed by Retrotransposon-
697 Mediated Exon Shuffling in Maize. *Plant Physiol.* **153**: 1413-1424.
- 698 Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the
699 mammalian X chromosome. *Science* **303**: 537-540.
- 700 Engelman A, Oztop I, Vandegraaff N, Raghavendra NK. 2009. Quantitative analysis of HIV-
701 1 preintegration complexes. *Methods* **47**: 283-290.
- 702 Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed
703 pseudogenes. *Nat. Genet.* **24**: 363-367.
- 704 Ewing A, Ballinger T, Earl D, Sequencing BIG, Program A, Platform, Harris C, Ding L,
705 Wilson R, Haussler D. 2013. Retrotransposition of gene transcripts leads to structural
706 variation in mammalian genomes. *Genome Biol.* **14**: R22.
- 707 Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N,
708 Steptoe AL, Lassmann T et al. 2009. The regulated retrotransposon transcriptome of
709 mammalian cells. *Nat. Genet.* **41**: 563-571.
- 710 Feuchter A, Mager D. 1990. Functional heterogeneity of a large family of human LTR-like
711 promoters and enhancers. *Nucleic Acids Res.* **18**: 1261-1270.

- 712 Gan H, Wen L, Liao S, Lin X, Ma T, Liu J, Song CX, Wang M, He C, Han C et al. 2013.
713 Dynamics of 5-hydroxymethylcytosine during mouse spermatogenesis. *Nat. Commun.*
714 **4**: 1995.
- 715 Ganko EW, Greene CS, Lewis JA, Bhattacharjee V, McDonald JF. 2006. LTR
716 retrotransposon-gene associations in *Drosophila melanogaster*. *J. Mol. Evol.* **62**: 111-
717 120.
- 718 Gilbert W. 1978. Why genes in pieces? *Nature* **271**: 501-501.
- 719 Goodrich DW, Duesberg PH. 1990. Retroviral recombination during reverse transcription.
720 *Proc. Natl. Acad. Sci. U. S. A.* **87**: 2052-2056.
- 721 Graur D, Li WH. 2000. *Fundamentals of Molecular Evolution*. Sinauer Associates,
722 Sunderland, Massachusetts.
- 723 Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature*
724 **482**: 339-346.
- 725 Hajjar AM, Linial ML. 1993. A Model System for Nonhomologous Recombination between
726 Retroviral and Cellular Rna. *J. Virol.* **67**: 3845-3853.
- 727 Hart T, Komori H, LaMere S, Podshivalova K, Salomon D. 2013. Finding the active genes in
728 deep RNA-seq gene expression studies. *BMC Genomics* **14**: 778.
- 729 Havecker ER, Gao X, Voytas DF. 2004. The diversity of LTR retrotransposons. *Genome Biol.*
730 **5**: 225.
- 731 Hindmarsh P, Leis J. 1999. Retroviral DNA Integration. *Microbiol. Mol. Biol. Rev.* **63**: 836-
732 843.
- 733 Jamain S, Girondot M, Leroy P, Clergue M, Quach H, Fellous M, Bourgeron T. 2001.
734 Transduction of the human gene FAM8A1 by endogenous retrovirus during primate
735 evolution. *Genomics* **78**: 38-45.

- 736 Jern P, Coffin JM. 2008. Effects of retroviruses on host genome function. *Annu. Rev. Genet.*
737 **42**: 709-732.
- 738 Jin YK, Bennetzen JL. 1994. Integration and nonrandom mutation of a plasma membrane
739 proton ATPase gene fragment within the Bs1 retroelement of maize. *Plant Cell* **6**:
740 1177-1186.
- 741 Johns MA, Mottinger J, Freeling M. 1985. A low copy number, copia-like transposon in
742 maize. *EMBO J.* **4**: 1093-1101.
- 743 Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic
744 and evolutionary insights. *Nat. Rev. Genet.* **10**: 19-31.
- 745 Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler
746 DA, Lewis SE, Rubin GM et al. 2002. The transposable elements of the *Drosophila*
747 *melanogaster* euchromatin: a genomics perspective. *Genome Biol.* **3**:
748 RESEARCH0084.
- 749 Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment
750 program. *Briefings in Bioinformatics* **9**: 286-298.
- 751 Kent WJ. 2002. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**: 656-664.
- 752 Kofler R, Nolte V, Schlötterer C. 2015. Tempo and Mode of Transposable Element Activity
753 in *Drosophila*. *PLoS Genet.* **11**: e1005406.
- 754 Langley CH, Stevens K, Cardeno C, Lee YC, Schrider DR, Pool JE, Langley SA, Suarez C,
755 Corbett-Detig RB, Kolaczowski B et al. 2012. Genomic variation in natural
756 populations of *Drosophila melanogaster*. *Genetics* **192**: 533-598.
- 757 Lee MT, Bonneau AR, Takacs CM, Bazzini AA, DiVito KR, Fleming ES, Giraldez AJ. 2013.
758 Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-
759 zygotic transition. *Nature* **503**: 360-364.

- 760 Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A,
761 Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA
762 sequencing methods. *Nat. Meth.* **7**: 709-715.
- 763 Li WH. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, Massachusetts.
- 764 Ling D, Salvaterra PM. 2011. Robust RT-qPCR data normalization: validation and selection
765 of internal reference genes during post-experimental data analysis. *PLoS One* **6**:
766 e17762.
- 767 Linheiro RS, Bergman CM. 2012. Whole genome resequencing reveals natural target site
768 preferences of transposable elements in *Drosophila melanogaster*. *PLoS One* **7**:
769 e30008.
- 770 Long M, Rosenberg C, Gilbert W. 1995. Intron phase correlations and the evolution of the
771 intron/exon structure of genes. *Proc. Natl. Acad. Sci. U. S. A.* **92**: 12495-12499.
- 772 Luo GX, Taylor J. 1990. Template switching by reverse transcriptase during DNA synthesis.
773 *J. Virol.* **64**: 4321-4328.
- 774 Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer
775 O, Trono D, Pfaff SL. 2012. Embryonic stem cell potency fluctuates with endogenous
776 retrovirus activity. *Nature* **487**: 57-63.
- 777 Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL. 2006.
778 Retroviral Elements and Their Hosts: Insertional Mutagenesis in the Mouse Germ
779 Line. *PLoS Genet.* **2**: e2.
- 780 Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. 2014.
781 From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and
782 RNA splicing. *Genome Res.* **24**: 496-510.

- 783 Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ, the
784 FlyBase c. 2013. FlyBase: improvements to the bibliography. *Nucleic Acids Res.* **41**:
785 D751-DD757.
- 786 Maxwell PH, Curcio MJ. 2007. Retrosequence formation restructures the yeast genome.
787 *Genes Dev.* **21**: 3308-3318.
- 788 McCarthy EM, McDonald JF. 2004. Long terminal repeat retrotransposons of *Mus musculus*.
789 *Genome Biol.* **5**: R14.
- 790 Mighell AJ, Smith NR, Robinson PA, Markham AF. 2000. Vertebrate pseudogenes. *FEBS*
791 *Lett.* **468**: 109-114.
- 792 Moran JV, DeBerardinis RJ, Kazazian HH. 1999. Exon Shuffling by L1 Retrotransposition.
793 *Science* **283**: 1530-1534.
- 794 Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F,
795 Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in
796 tetrapods. *Nature* **505**: 635-640.
- 797 Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC,
798 Elkahouloun A, Cargill M, Jones PG et al. 2009. An expressed *fgf4* retrogene is
799 associated with breed-defining chondrodysplasia in domestic dogs. *Science* **325**: 995-
800 998.
- 801 Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, Rosbash M, Waddell S. 2013.
802 Transposition-driven genomic heterogeneity in the *Drosophila* brain. *Science* **340**: 91-
803 95.
- 804 Petrov DA, Hartl DL. 1998. High rate of DNA loss in the *Drosophila melanogaster* and
805 *Drosophila virilis* species groups. *Mol. Biol. Evol.* **15**: 293-302.
- 806 Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs.
807 *Cell* **136**: 629-641.

- 808 Ponton F, Chapuis MP, Pernice M, Sword GA, Simpson SJ. 2011. Evaluation of potential
809 reference genes for reverse transcription-qPCR studies of physiological responses in
810 *Drosophila melanogaster*. *J. Insect Physiol.* **57**: 840-850.
- 811 Richardson SR, Morell S, Faulkner GJ. 2014. L1 Retrotransposons and Somatic Mosaicism
812 in the Brain. *Annu. Rev. Genet.* **48**: 1-27.
- 813 Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu. Rev.*
814 *Biochem.* **81**: 145-166.
- 815 Ruby JG, Bellare P, Derisi JL. 2013. PRICE: Software for the Targeted Assembly of
816 Components of (Meta) Genomic Sequence Data. *G3 (Bethesda)* **3**: 865-880.
- 817 Schacherer J, Tournette Y, Souciet JL, Potier S, De Montigny J. 2004. Recovery of a function
818 involving gene duplication by retroposition in *Saccharomyces cerevisiae*. *Genome Res.*
819 **14**: 1291-1297.
- 820 Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza
821 SJ. 2013. Gene copy-number polymorphism caused by retrotransposition in humans.
822 *PLoS Genet.* **9**: e1003242.
- 823 Schrider DR, Stevens K, Cardeno CM, Langley CH, Hahn MW. 2011. Genome-wide analysis
824 of retrogene polymorphisms in *Drosophila melanogaster*. *Genome Res.* **21**: 2087-
825 2095.
- 826 Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence
827 comparison. *BMC Bioinformatics* **6**: 31.
- 828 Soares MB, Schon E, Henderson A, Karathanasis SK, Cate R, Zeitlin S, Chirgwin J,
829 Efstratiadis A. 1985. Rna-Mediated Gene Duplication - the Rat Preproinsulin-I Gene
830 Is a Functional Retroposon. *Mol. Cell Biol.* **5**: 2090-2103.

- 831 Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthes P, Kokkinaki M,
832 Nef S, Gnirke A et al. 2013. Cellular source and mechanisms of high transcriptome
833 complexity in the mammalian testis. *Cell Rep.* **3**: 2179-2190.
- 834 Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. 2014. Poly(A)-tail profiling
835 reveals an embryonic switch in translational control. *Nature* **508**: 66-71.
- 836 Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence
837 alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**: W609-
838 W612.
- 839 Swain A, Coffin JM. 1992. Mechanism of transduction by retroviruses. *Science* **255**: 841-
840 845.
- 841 Tan S, Zhu Z, Zhu T, Te R, Zhang YE. 2014. Chance and Necessity: Emerging Introns in
842 Intronless Retrogenes. *eLS* doi:10.1002/9780470015902.a0022886.
- 843 Toba G, Aigaki T. 2000. Disruption of the Microsomal glutathione S-transferase-like gene
844 reduces life span of *Drosophila melanogaster*. *Gene* **253**: 179-187.
- 845 Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012.
846 Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**: e115.
- 847 Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome
848 in the *Drosophila* genus. *Genome Res.* **19**: 897-903.
- 849 Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene
850 copies in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **103**: 3220-3225.
- 851 Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S et al. 2006.
852 High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell*
853 **18**: 1791-1802.

- 854 Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV.
855 2001. Human L1 Retrotransposition: cisPreference versus trans Complementation.
856 *Mol. Cell Biol.* **21**: 1429-1439.
- 857 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhou B, Flavell A, Leroy P,
858 Morgante M, Panaud O et al. 2007. A unified classification system for eukaryotic
859 transposable elements. *Nat. Rev. Genet.* **8**: 973-982.
- 860 Witherspoon DJ, Xing JC, Zhang YH, Watkins WS, Batzer MA, Jorde LB. 2010. Mobile
861 element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC*
862 *Genomics* **11**: 410.
- 863 Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.*
864 **24**: 1586-1591.
- 865 Yoshioka K, Honma H, Zushi M, Kondo S, Togashi S, Miyake T, Shiba T. 1990. Virus-Like
866 Particle Formation of Drosophila-Copia through Autocatalytic Processing. *EMBO J.*
867 **9**: 535-541.
- 868 Yue F Cheng Y Breschi A Vierstra J Wu WS Ryba T Sandstrom R Ma ZH Davis C Pope BD
869 et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome.
870 *Nature* **515**: 355-364.
- 871 Zhang YE, Landback P, Vibranovski MD, Long M. 2011. Accelerated recruitment of new
872 brain development genes into the human genome. *PLoS Biol.* **9**: e1001179.
- 873 Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008.
874 On the origin of new genes in Drosophila. *Genome Res.* **18**: 1446-1455.
- 875 Zhu Z, Tan S, Zhang Y, Zhang YE. 2016. LINE-1-like retrotransposons contribute to RNA-
876 based gene duplication in dicots. *Sci. Rep.* **6**: 24755.

877

878

Tables**Table 1. The 15 *D. melanogaster* polymorphic retrocopies.**

Parental gene	Parent/retrocopy transcript length (bp)	LTR retrotransposon annotation	Microsimilarity (bp)
<i>CG17604</i>	2651/975	<i>BLASTOPIA</i> , <i>Gypsy</i>	16, 2
<i>CG10212</i>	3687/1323	<i>BLASTOPIA</i> , <i>Gypsy</i>	6, 0
<i>CG32082</i>	5246/704	<i>BLASTOPIA</i> , <i>Gypsy</i>	13, 1
<i>CG4799</i> , <i>CG11924</i>	2571/243, 3530/917	<i>BATUMI</i> , <i>Pao</i>	6, 9, 12
<i>CR42443</i> , <i>CG5119</i>	1023/441 [#] , 3127/1294	<i>STALKER2</i> , <i>Gypsy</i>	2, -16
<i>CG33957</i>	3879/745	<i>BLASTOPIA</i> , <i>Gypsy</i>	9, 1
<i>CG14796</i>	5388/1073	<i>MAX</i> , <i>Pao</i>	10, 10
<i>CG31605</i>	2792/1564 [#]	<i>TABOR</i> , <i>Gypsy</i>	9
<i>CG8567</i>	2425/1683	<i>412</i> , <i>Gypsy</i>	-2, 3
<i>CG2662</i>	1647/377	<i>412</i> , <i>Gypsy</i>	-26, 15
<i>CG5020</i>	5625/1469	<i>BLASTOPIA</i> , <i>Gypsy</i>	7, 2
<i>CG3631</i>	2757/1695	<i>Gypsy5</i> , <i>Gypsy</i>	4, 7
<i>CG1242</i>	2713/915	<i>MAX</i> , <i>Pao</i>	16, 2
<i>CG10652</i>	513/205	<i>412</i> , <i>Gypsy</i>	16, 1
<i>CG8392</i>	846/394 [#]	<i>412</i> , <i>Gypsy</i>	12

Transcript length is the length of the parental gene's mRNA. When a parental gene encodes multiple alternative isoforms compatible with the retroposed sequences, one

isoform is randomly chosen (see Supplemental Dataset 1). The retrocopies where we failed to assemble one flanking region are marked with “#” and the length of the current contig is shown. LTR retrotransposon is described as element name followed by its family name, e.g., “*BLASTOPIA*, *Gypsy*”. Microsimilarity refers to the stretch of similar sequence present at the switch points shared by the retrocopies and the flanking LTR retrotransposons. The numbers in this column follow the order of retroposition, i.e., from retrotransposons’ 3’ terminal to their 5’ terminal. The three negative values refer to the length of *de novo* sequences inserted at the switch point. For the retrocopy derived from *CG10212*, there is no microsimilarity or *de novo* sequence at its 5’ end, which is therefore referred as “0”.

Table 2. Insertion sites of nine retrocopies.

Retrocopy	Position	Method
<i>CG17604_r</i>	SR:TATAA	2x250 bp
	Satellite [#]	2x250 bp
	2L:11489564 [#]	intergenic
	2R:21009193 [#]	intronic of <i>CG16778</i>
	3R: 3618946 [#]	3' UTR of <i>CG10086</i>
	3R:9536417 [#]	intronic of <i>CG33748</i>
<i>CG10212_r</i>	LINE:DOC_DM	2x250 bp
	2L:13226230 [#]	intergenic
	2R:3607631 [#]	intronic of <i>CG18812</i>
	3R:21723045 [#]	intergenic
<i>CG32082_r</i>	3L:230088887 [#]	intronic of <i>CG8779</i>
<i>CG4799_CG11924_r</i>	3L:21967546 [#]	intronic of <i>CG7470</i>
<i>CG33957_r</i>	2L:17134768 [#]	intronic of <i>CG34106</i>
<i>CG14796_r</i>	X:8740722-8741237 ^x	intronic of <i>CG12075</i>
<i>CG31605_r</i>	3L:17668235 ^x	intronic of <i>CG34251</i>
<i>CG5020_r</i>	3R:13789450 [#]	intronic of <i>CG42457</i>
<i>CG1242_r</i>	3L:2624322-2624753 ^x	intronic of <i>CG12002</i>
	3R:15303121-15303991 ^x	exonic/intronic of <i>CG34157</i>
		ME-Scan

All insertion sites (except *CG17604_r* in a simple repeat and *CG10212_r* in a LINE) were tested by PCR and Sanger sequencing. The sites validated and the false positives are marked with “[#]” and “^x”, respectively.

Table 3. The five newly originated *Drosophila* retrocopies.

Parent/ Retrocopy ^A	Retrocopy location	Parent/ retrocopy transcript length (bp)	Diverg- ence ^B	Microsi- milarity (bp)	Flanking transposon, transposon family, divergence ^C	Poly(A)	TSDs
<i>CG33932/</i> <i>NA</i>	Uextra:5521091- 5521902 ^D	1244/ 812	0.10%	-11 ^E	<i>MAX</i> , <i>Pao</i> , 3.70%	-	-
<i>CG16728/</i> <i>NA</i>	X:17689090- 17689601	2489/ 512	0.90%	7, 2	<i>BLASTOPIA</i> , <i>Gypsy</i> , 0.20%	-	-
<i>CG2033/</i> <i>CG12324</i>	2R:6709256- 6709889	678/ 634	2.30%	-	No associated repeat elements	AAAA	-
<i>CG1742/</i> <i>CR12628</i>	2L:22226151- 22226773	667/ 623	3.50%	-14, 9	<i>G3_DM</i> , <i>Jockey</i> , <i>LINE</i> , 10% ^F	AAAAA AA	AATCA/ AATCA
<i>CG3265/</i> <i>CG40354^G</i>	3LHet:687408- 688819	1873/ 1412	4.00%	2 ^H	<i>TV1</i> , <i>Gypsy</i> , 27.50%	-	ATTCTTCTT G/ATCTTCT TG

^ABecause the function of these genes (except *CG12324* or *Rps15ab*) is not characterized we use the term retrocopy rather than retrogene. ^BDivergence reflects the nucleotide differences between parental genes and retrocopies. ^CThe average divergence relative to the consensus sequence generated by RepeatMasker. ^DThe sequence of this retrocopy is absent from the new *D. melanogaster* Release 6 (Dm6). However, we could find exon-exon spanning reads in more than one DGRP line suggesting that it is genuine. ^EA sequencing gap flanks one end of the retrocopy preventing the identification of microsimilarity stretches. ^FThe presence of a continuous *Jockey* element in the closely related species, *D. simulans*, suggests that the retrocopy was inserted into this pre-existing element (Supplemental Fig. 11). ^G*CG40354* is annotated as a pseudogene in Dm6. ^H*TV1* is immediately adjacent to this

retrocopy at the 3' flanking region, while the *QUASIMODO* Gypsy element is located in the 5' flanking region 111 bp away (Supplemental Fig. 4B). So, only microsimilarity relative to *TV1* is examined.

Figure legends

Figure 1. Schematic representations of a subset of the retrocopies identified. LTR retrotransposons and retrocopies are marked in grey and blue, respectively. Microsimilarity stretches are marked in yellow, and *de novo* sequence insertions in green. “CDS” and “UTR” refer to coding sequences and untranslated regions respectively. The red line marks the “TAAAAACAAGC-GGTTG” motif.

Figure 2. The expression of retrocopies in fruitfly and mouse. (A) Whole body RT-PCR results for 10 retrocopies validated in Supplemental Fig. 2, which are expressed in both male (M) and female (F). The primer sequences and expected product sizes are listed in Supplemental Table 9. For the chimeric genes *CG5119_CR42443_r* and *CG4799_CG11924_r*, we also designed primers flanking the fusion point between the two parental genes, the product of which is marked with “C”. (B) Tissue profiling via qRT-PCR. *CG2662_r* has no detectable expression in ovary. The Pearson correlation (r) between the retrocopies and corresponding retrotransposons across the three tissues is displayed above. (C) Tissue level RT-PCR results for 4 out of 10 retrocopies encoded by line 208 across Testis (T), Ovary (O) and Head (H). (D) Expression profile of retrocopies encoded by the mouse genome. Tissues are hierarchically clustered on the basis of expression similarity across genes. Retrocopies are not clustered but instead are sorted by age as approximated by nucleotide divergence. To reveal relative moderate expression, the color-code scheme is truncated at the FPKM cutoff of 1 (i.e., all expression higher than 1 is shown in dark blue).

Figure 3. Strand-specific RNA-seq based quantification. On the basis of newly generated and public strand-specific RNA-seq data, we quantified the expression of LTR-mediated retrocopies in fruitfly and mosquito (A), zebrafish (B), chicken (C) and mouse (D). For fruitfly, given the high similarity between polymorphic retrocopies and parental genes, only switch points spanning reads were counted and FPM was calculated. For all other cases, FPKM was calculated. For mosquito,

embryo, the 3rd instar larvae, the 4th instar larvae and pupae were profiled. For zebrafish, 4-hour, 6-hour and 8-hour post fertilization (pf) stages were profiled.

Figure 4. A schematic representation of the template switch model for LTR-mediated retroposition. VLP: virus-like particle; RT: reverse transcriptase; PR: protease; IN: integrase. The black boxes in LTR retrotransposons correspond to the LTRs and the light green box to the ORF. The blue boxes correspond to exons encoded by the host genome and the white boxes to introns. The double template switch occurs in the VLP, generating a chimeric cDNA (multiple switches can occur). After integration, the chimeric sequences can be transcribed as pseudo-LTR retrotransposons, which could be subject to further cycles of retroposition. Previous literatures are cited when appropriate.

A

CDS of *CG17604*

GCACAAAAGGAAAAGTTTGA **GG** CAAACATTTTACAATTAAGCA.....GAGCATTCTACTAAGGTGC **AACCGGAAAATACAGAC** GAATTTTCAAGCACTGCCAG
GAAGCAGATAAAAAACAAGC **GG** TTGAGTATTCCCAAATGTTT ATGAAGTTGTGGAAGTACAT **AGCCG AAAATGCAGAC** CTATTGTCATCCGGGTCGGA
 5' *BLASTOPIA* 3' *BLASTOPIA*

B

3' *BATUMI*

TCTCCGTGTTGATGCCTGAC **AGTTCT** AAGGCAGTCAAGCAACCTTG
 GCTTTTCAAATCAAACGTA **AATTCT** **GGCGGCGCCAAGCAGGCATT**.....
 ----- 5' UTR of *CG4799* -----
 ----- **AAAGTTTCACACATTTTCATC** **GCAGCAGCA** AACATCTAACTTTTAAAGA
 CAACAACAACAGCAGCAGCA **GCAGCAGCA** **GGA**ACTGCTGGAGCAGCAGC.....
 ----- CDS of *CG11924* -----
 ----- **ACGGGATGAGCAAACCAATA** **AATGAAAGACCA** AACATCTAACTTTTAAAGA
 GGAAAATTCCTAGTTTGCTG **AAAAAAAGACCA** CTGAAAAAACACTCGAAGA
 5' *BATUMI*

C

CDS of *AGAP000264*

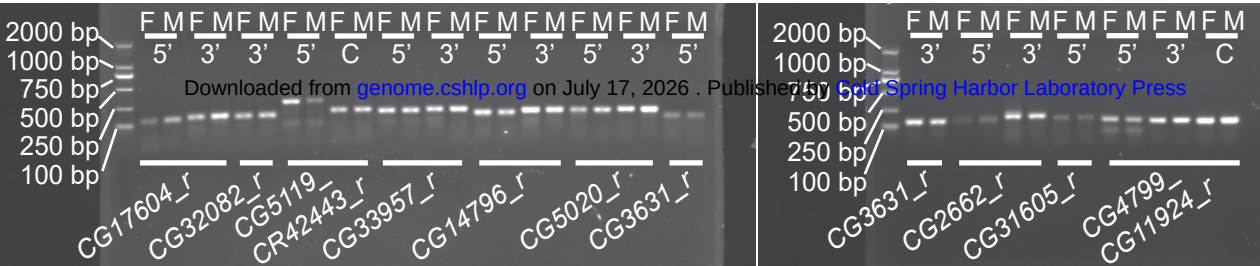
AACGCTGCTCGAACG **CAACAAA** **GACCTGGAGGCGACG**.....AGCTCAAGAGCACGC **AAAAG** TGTCTCCTGTCCGAG
TCCTGTCTTTAATCC **CAACAAA** CCAGGAAAGGTGCGA TCGTACAGTCGGTAA **AAAAG** **GTGTTAAGAGAGATA**
 5' *BEL18-I_AG* 3' *BEL18-I_AG*

D

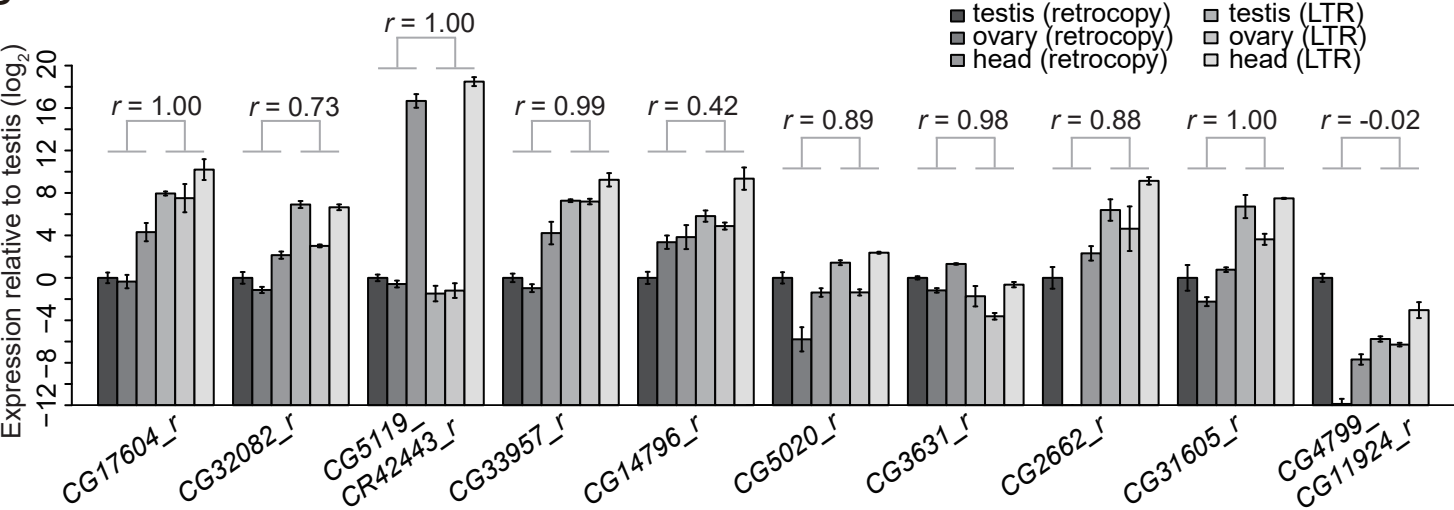
CDS of *ENSMUSG00000021752*

TCGCGCTCTCGGGCCGCCGG **CGTCTCCTGGCCAGGTGC**.....GTCCACTCCTTACTCGAA **GGCATCTCAA** ACTATTTACCAAGTGAAT
CCCACTTTATAGTGACTGCCAT TTATGGCCGGCCTTGA AAGACCTCTTTTACT **GGCATCTCAA** **GGTTACTGGA**ACTGCTCCAA
 5' *IAPEy* 3' *IAPEy*

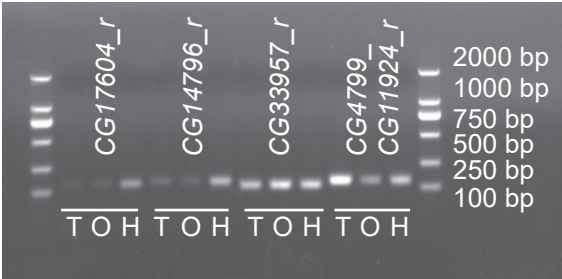
A



B



C



D

