SI Text

Intra-Element LTR-LTR Divergence and Patterns of Retrotransposon Nesting Support Recent Insertion of LTR Elements. For LTR elements, it is possible to independently confirm their age based on differences between LTRs within the same genomic copy, based on the fact that LTR sequences are identical at the time of integration (1, 2). This method is limited in that it can only be applied to LTR families and to genomic copies which retain both LTRs. Applying this method to the 270 elements with 2 LTRs, we find that the distributions of intra-element LTR divergences also demonstrate a recent origin for *D. melanogaster* LTR families (SI Fig. 3), confirming previous results based on unfinished Release 2 genome sequences (2). If both metrics accurately estimate the age since LTR integration, then terminal branch lengths should be positively correlated with intra-element LTR-LTR divergence. As shown in SI Fig. 4, we observe a strong positive correlation between terminal branch lengths with intra-element LTR-LTR divergence (Pearson's $r^2 = 0.3491$, $P < 2 \times 10^{-16}$). The slope of this relationship is slightly less than unity, since many intra-element LTR-LTR divergence values of zero are observed for LTR elements with non-zero terminal branch lengths, probably resulting from the shorter length of LTR sequences relative to ORF fragments in our alignments. If we assume that terminal branch lengths reflect the true age since insertion, ages of LTR integration based on LTR sequences in *Drosophila* may be underestimates either because of their relative short target length for mutation or because of homogenization by local gene conversion. Conversely, if we assume that intra-element LTR-LTR divergences are correct, this implies that we are overestimating terminal branch lengths for LTR elements using our method of unique substitutions. This latter possibility seems unlikely since the unique substitution method can only underestimate distances by the removal of non-unique sites that arise on internal branches or because of homoplasy. Whatever the causes of these slight discrepancies, both dating methods clearly support the conclusion that LTR elements have integrated in the recent past, and the strong correlation between age estimates from both methods provides independent confirmation for our method of unique substitution to estimate retrotransposon ages.

The recent LTR integration hypothesis makes predictions about the spatial organization of nested TEs, whereby one retrotransposon inserts inside a previously existing copy in the genome. A recent high-resolution annotation of TEs in *Drosophila* (3, 4) allows us to test whether LTR elements are more often the "inner" (newly inserted) rather than the "outer" (older) component of a TE nest, as would be expected if the majority of LTR elements have integrated in the recent past. Bergman *et al.* (3) observe 151 LTR→LTR, 39 LTR→non-LTR, 46 non-LTR→LTR and 46 non-LTR→non-LTR nests in the Release 4 genome annotation. Under the null hypothesis that all retrotransposons randomly integrated in the genome, we can calculate the expected proportion of outer elements as the relative proportion of genome sequence covered by LTR (3.29 Mb) and non-LTR (1.27 Mb) elements, and the expected proportion of inner elements as the relative proportion of LTR (1,321) and non-LTR (1,019) insertions in the genome. A test of random integration reveals strong departure from the null hypothesis for all four types of nests (G = 19.79, P = 0.00019, 3 d.f.) with an excess of nests among members of the same retrotransposon subclass (LTR→LTR and non-LTR→non-LTR). However, since it is possible that homing mechanisms cause recurrent insertion of the similar TE into the same genomic region and generate a preference for self-nesting, we repeated this test only on nests LTR→non-LTR and non-LTR→LTR nests by rescaling expected probabilities of these two type of nest. Controlling for the potential effects of homing, we find a significant excess of LTR→non-LTR nests relative to genome-wide expectation (G = 5.71, P = 0.0167, 1 d.f.). Thus global patterns of TE nesting in the *D. melanogaster* genome are consistent with the recent LTR element insertion hypothesis, supporting results based on both terminal branch length and intra-element LTR-LTR divergence.

Systematic Differences in Age Between LTR and Non-LTR Elements Are not Caused by Differences in Their Distribution with Respect to Recombination Rate or Transcription. Previous work has shown that non-LTR elements in regions of low recombination are older than non-LTR elements in regions of high recombination (5), and that TEs of all classes achieve high population frequencies (and even fixation) in regions of low recombination (6). Since our sample is enriched for non-LTR elements in regions of low recombination (especially on the 4th chromosome), we investigated whether overall differences in terminal branch lengths observed between LTR and non-LTR elements can be explained by differences in recombination rate (SI Fig. 5*A*). Our results confirm that the terminal branch lengths of non-LTR elements are on average longer in regions of low recombination than in regions of high recombination (Wilcoxon test, P = 6.114e-15) (5). Similarly, there is a slight but significant tendency for LTR elements to have longer terminal branch lengths in regions of low recombination (Wilcoxon test $P = 0.0004429$). These results indicate that the increased age of TE insertions in regions of low recombination is not restricted to one particular subclass of TE and may be a general phenomenon in *Drosophila* (see also ref. 7). Despite these general effects, LTR elements have significantly shorter branch lengths than non-LTR elements in regions of both high (Wilcoxon test $P = 0.01169$) and low (Wilcoxon test $P = 5.527e\text{-}15$) recombination, indicating that differences in age between LTR and non-LTR elements exist regardless of any effect of local recombination rates. In support of this non-parametric analysis, we find that TEs are older in low recombination regions regardless of whether these are LTR or non-LTR elements (linear mixed effects model, $F_{1,407}= 28.8$, $P < 0.0001$) and that non-LTR elements are older than LTR elements irrespective of recombination rate ($F_{1,25}= 28.0$, $P < 0.0001$, see main text).

Intriguingly, there are three non-LTR elements in regions of high recombination that have estimated ages older than the divergence of *D. melanogaster* from *D. simulans*. Two of these are *Cr1a* elements that are also present in the *D. simulans* and *D. sechellia* genomes and therefore likely to be fixed in the species. Both fixed *Cr1a* elements are present in intronic regions of transcription units (FBti0019203 in *CG10188* and FBti0020079 in *CG33926*) and may contribute functional sequences to the genome. The third (FBti0019530), a *baggins* element, is not present in the *D. simulans* or *D. sechellia* genomes and may represent a divergent *baggins* subfamily that did not exhibit significant constraint in our pseudogene test.

It is a well-established fact that *Drosophila* TEs occur less frequently in exons and introns relative to intergenic regions, consistent with negative selection occurring on TE insertions that occur in transcription units (8-11). Thus, we tested whether ages of TEs differ among genomic compartments and whether differences in age distribution between LTR and non-LTR elements are still observed while controlling for the effects of transcription (SI Fig. 5*B*). Despite small numbers because of their extreme rarity (LTR, $n = 16$; non-LTR, $n = 4$), we observe that only young retrotransposons can be found in exons, as is expected if TE insertions in exons are generally strongly deleterious and do not persist in natural populations long enough to accumulate unique substitutions. We also found a slight tendency for LTR elements to be more frequent in introns and exons relative to non-LTR elements (Pearson's χ^2 test, $P = 0.0338$, 2 d.f.). Combining exonic and intronic TE insertions, we find that the median age does not differ between genic and intergenic regions for LTR elements

(Wilcoxon test p=0.1019) whereas for non-LTR elements, we find that intergenic insertions are older than genic insertions (Wilcoxon test, *P* =0.0287). Despite this signal of older non-LTR elements in intergenic regions from univariate analysis, we find no significant main effect of whether a TE is in a genic region on its age in our linear mixed effects model ($F_{1,407}$ = 3.2, *P* = 0.08). However, we do find that genic TEs are younger than intergenic TEs in high recombination regions, but less so in low recombination regions leading to a significant interaction between recombination and transcription in our linear mixed effects model (likelihood ratio = 6.0, d.f. = 1, *P* = 0.015).

Consistent with no strong confounding effect of transcription on average retrotransposon age, our main result that LTR elements are significantly younger than non-LTR elements is true for both intronic (Wilcoxon test *P* = 0.04104) and intergenic (Wilcoxon test *P* < 2.2e-16) regions. This is confirmed by our linear mixed effects model ($F_{1,25}$ = 28.0, *P* < 0.0001, see above), which also shows that LTR elements are significantly younger than non-LTR elements regardless of whether TEs are in genic or intergenic regions. Thus, despite the deleterious effects of intronic insertions, the ages of both subclasses of retrotransposon are not strongly affected by the mode of selection that purges them from genic regions of the genome except in high recombination regions.

1. Shevelyov, Y. Y. (1993) *Mol Gen Genet* **239,** 205-8.
2. Bowen, N. J. & McDonald, J. F. (2001) *Genome Res* **11,** 1527-40.
3. Bergman, C. M., Quesneville, H., Anxolabehere, D. & Ashburner, M. (2006) *Genome Biol* **7,** R112.
4. Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M. & Anxolabehere, D. (2005) *PLoS Comput Biol* **1,** e22.
5. Blumenstiel, J. P., Hartl, D. L. & Lozovsky, E. R. (2002) *Mol Biol Evol* **19,** 2211-25.
6. Bartolome, C. & Maside, X. (2004) *Genet Res* **83,** 91-100.
7. Lerat, E., Rizzon, C. & Biemont, C. (2003) *Genome Res* **13,** 1889-96.
8. Charlesworth, B. & Langley, C. H. (1989) *Annu Rev Genet* **23,** 251-87.
9. Bartolome, C., Maside, X. & Charlesworth, B. (2002) *Mol Biol Evol* **19,** 926-37.
10. Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., Ashburner, M. & Celniker, S. E. (2002) *Genome Biol* **3,** RESEARCH0084.
11. Lipatov, M., Lenkov, K., Petrov, D. A. & Bergman, C. M. (2005) *BMC Biol* **3,** 24.