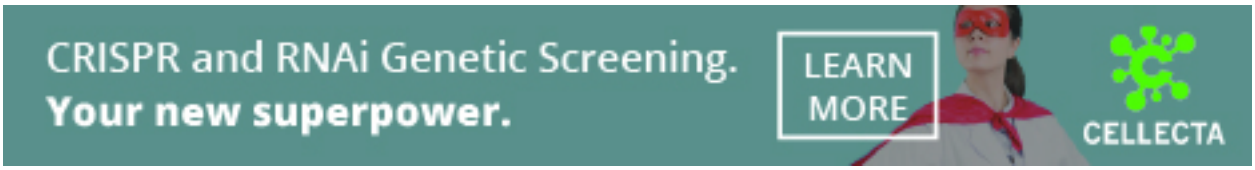
# Transposon expression in the *Drosophila* brain is driven by neighboring genes and diversifies the neural transcriptome

Christoph D Treiber and Scott Waddell

*Genome Res.* published online September 24, 2020
Access the most recent version at doi:10.1101/gr.259200.119





# Transposon expression in the *Drosophila* brain is driven by neighboring genes and diversifies the neural transcriptome

Christoph D. Treiber* & Scott Waddell*

Centre for Neural Circuits and Behaviour, University of Oxford, Tinsley Building, Mansfield Road, Oxford OX1 3SR, UK

*Correspondence.

Email: christoph.d.treiber@gmail.com, scott.waddell@cncb.ox.ac.uk

Running title: Transposons diversify the neural transcriptome

Keywords: Transposon expression; alternative splicing; transcriptional heterogeneity; single-cell transcriptomics.

## Abstract

Somatic transposon expression in neural tissue is commonly considered as a measure of mobilization and has therefore been linked to neuropathology and organismal individuality. We combined genome sequencing data with single-cell mRNA sequencing of the same, inbred fly strain to map transposon expression in the *Drosophila* midbrain and found that transposon expression patterns are highly stereotyped. Every detected transposon is resident in at least one cellular gene with a matching expression pattern. Bulk RNA sequencing from fly heads of the same strain revealed that coexpression is a physical link in the form of abundant chimeric transposon-gene mRNAs. We identified 264 genes where transposons introduce cryptic splice sites into the nascent transcript and thereby significantly expand the neural transcript repertoire. Some genes exclusively produce chimeric mRNAs with transposon sequence and on average 11.6% of the mRNAs produced from a given gene are chimeric. Conversely, most transposon-containing transcripts are chimeric, which suggests that somatic expression of these transposons is largely driven by cellular genes. We propose that chimeric mRNAs produced by alternative splicing into polymorphic transposons, rather than transposon mobilization, may contribute to functional differences between individual cells and animals.

## Introduction

Transposons comprise up to ~50% of eukaryotic genomes (Britten and Kohne, 1968; International Human Genome Sequencing Consortium et al., 2001; Ketchum et al., 2000) and their mobilization in the germline contributes to chromosome evolution. Transposon activity comprises a wide array of molecular functions (Bourque et al., 2018; Sienski et al., 2012). Non-heritable *de novo* transposition in neural tissue may contribute to functional heterogeneity in the brain and to neurological disease (Baillie et al., 2011; Coufal et al., 2009; Evrony et al., 2012; Kazazian, 2011; Kazazian and Moran, 2017; Muotri et al., 2005; Schauer et al., 2018). However, it is difficult to map rare *de novo* transposon insertions using whole-genome DNA sequencing (Baillie et al., 2011; Evrony et al., 2012, 2016; Perrat et al., 2013; Treiber and Waddell, 2017; Upton et al., 2015). Some studies therefore correlate neurodegeneration in animal models with changes in transposon expression (Guo et al., 2018; Krug et al., 2017; Li et al., 2013; Li et al., 2012; Sun et al., 2018). Using elevated expression as a proxy for mobility could be misleading because it does not always result in *de-novo* somatic transposition (Evrony et al., 2012, 2016; Treiber and Waddell, 2017). It is therefore important to understand what controls neural expression of transposon-derived sequences.

Transposons often reside in introns where they can introduce splice sites producing chimeric mRNAs between the transposon and the relevant gene (Deininger, 2011; Makałowski et al., 1994). Around 4% of human genes incorporate transposon sequences as novel exons (Nekrutenko and Li, 2001) and 75% of human lncRNAs contain segments of transposon origin (Kapusta et al., 2013). However, it is unclear how chimeric transcripts from these loci contribute to the overall pool of transposon mRNAs in somatic cells. Reliable measurement of autonomous- and non-autonomous transposon expression in somatic tissue, is hampered by repetitive sequences being difficult to map and germline transposons being polymorphic (Lanciano and Cristofari, 2020). Hence, many somatic transposon expression studies have analyzed single transposon families or have used bulk sequencing of tissues or cultured

cells (Babaian et al., 2019; Chung et al., 2019; Faulkner et al., 2009; Li et al., 2013; Philippe et al., 2016; Pinson et al., 2018; Rangwala et al., 2009; Wang et al., 2016). A genome-wide assessment of the prevalence of chimeric transcripts, requires that cellular expression of each transposon in the genome can be related to that of their surrounding genes. Technical developments in high-throughput single-cell transcriptomics of complex tissues, such as the fly brain, make this possible (Croset et al., 2018; Macosko et al., 2015).

Here we used single-cell RNA-seq (scRNA-seq) to map transposon expression to individual cells in the *Drosophila* midbrain. Combining these data with high-coverage genomic DNA (gDNA) sequencing of the same inbred fly strain permitted neural transposon expression to be correlated with that of genes within which they are inserted. We confirmed these transposon-gene interactions by extracting mRNA from heads of the same strain and performing high-coverage bulk-mRNA sequencing. Breakpoint-spanning sequences identified genome-wide splicing of host genes to transposons that generates a considerable diversity of mature chimeric mRNAs. We also present a quality-control approach using ‘immobile genetic elements’ (IGEs) to quantify rates of amplification artifacts in bulk mRNA sequencing data. Finally, we analyze mRNA sequencing data from other fly strains to assess how chimeric transcripts vary between strains.

## Results

### Single-cell transcriptomics reveals cell-type restricted transposon expression

The *Drosophila* genome contains at least 112 transposon sub-families, with copy number of an individual type ranging from a few to hundreds (Kaminker et al., 2002). Conventional scRNA-seq analyses typically discard sequencing reads that align to multiple genomic loci, and therefore underestimate transposon expression. Multiply aligned scRNA-seq reads can be kept and their counts divided by the number of copies in the genome. However, germline insertions in the reference genome are likely to differ substantially from insertions in our tested fly strain, making quantification of their expression inaccurate. We therefore devised an alternative analysis pipeline to map expression of all transposons within scRNA-seq data (scTE-seq). scTE-seq masks repetitive sequences in the reference genome and adds a single copy of the consensus sequence for every known transposon to this masked genome. This produces a *Drosophila* reference genome with one copy of each transposon sub-family. We used this modified reference genome to map transposon and gene reads onto individual midbrain cells from a fly strain expressing mCherry in αβ Kenyon cells (KCs) of the mushroom body (MB); from here called αβCherry flies. We found evidence for expression of the sense and antisense strands of most transposons, which comprised 75.5 and 24.5% (+/- 1.9% SD) of all transposon expression, respectively (Supplemental Figure S1, Supplemental Table S1). We verified our mapping approach reliably captured transposon reads by comparing our results from scTE-seq to those obtained using RepEnrich2 (Criscione et al., 2014). Counts computed by RepEnrich2 were strongly correlated with the number of uniquely mapping reads identified by scTE-seq ($R^2$=0.661, Supplemental Figure S2). Therefore, mapping to consensus sequences did not bias transposon expression levels. We clustered cells from the midbrain and assigned many clusters to known cell types using marker gene expression (Croset et al., 2018) (Figure 1A). Displaying transposons on the cluster plot revealed some to be up-regulated in specific cell types. For example, the long-terminal repeat (LTR) retrotransposons *copia* and *opus* were elevated in the αβ, α´β´ and γ KCs (Figure 1B, first graph) and α´β´ KCs (Figure 1C, first graph), respectively. Other LTR

retrotransposons such as *micropia* were upregulated in the ellipsoid body (Figure 1D, first graph) whereas *blood* and *412* were higher in glia (Figure 1E, F, first graphs).

# Figure 1

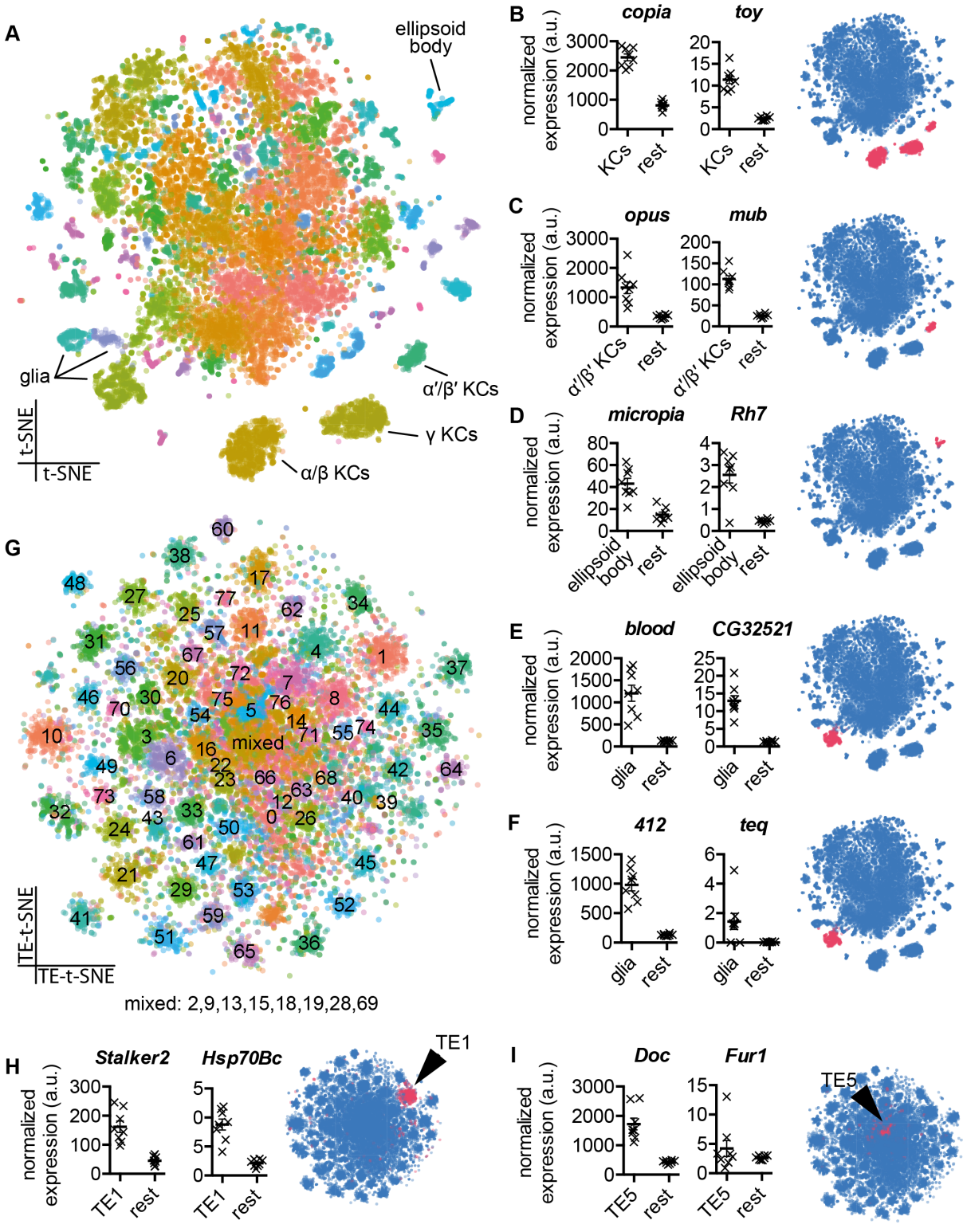


**Figure 1. Single-cell transcriptomics reveals patterned transposon expression in the *Drosophila* midbrain.**

**A** Two-dimensional reduction (t-SNE) of 14,804 *Drosophila* midbrain cells, based on gene expression levels. Colors represent cell clusters (at SNN resolution of 3.5). **B-F** Mean expression of transposons and neighboring cellular genes in the relevant cell groups in 8 biological replicates and t-SNE representation of cell-type restricted expression. **B** *copia* and *twin of eyeless* (*toy*) in all Kenyon Cell (KC) classes. **C** *opus* and *mushroom-body expressed* (*mub*) in α´β´ KCs. **D** *micropia* and *Rhodopsin 7* (*Rh7*) in the ellipsoid body **E and F** *blood* and *CG32521*, and *412* and *Tequila* (*teq*) in glia. Values represent the mean normalized number of unique molecular identifiers (UMI's) in an average cell from each cell type, and from the rest of the midbrain. Error bar indicates standard error of mean (SEM). Note; transposon- and gene levels were normalized separately. Blue schematic shows location of cell cluster (pink) in t-SNE plot. **G** Two-dimensional reduction of 14,804 *Drosophila* midbrain cells, based exclusively on transposon expression levels. Colors represent cell clusters (at SNN resolution of 3.5). **H and I** Mean expression of *Stalker2* and *Heat-shock-protein-70Bc* (*Hsp70Bc*), and *Doc* and *Furin 1* (*Fur1*) in their relevant transposon clusters and the position of the cluster in the overall transposon-based t-SNE (indicated in pink).

**Transposon expression correlates with that of cellular genes they are inserted within**

We reasoned that transposon expression might be elevated in specific cells because a copy of that transposon is inserted in a gene that is highly expressed in the same cells. To test this hypothesis, we took our previously published high-coverage gDNA sequence of αβCherry flies and mapped the germline transposon insertions in these flies using TEchim, a custom-built transposon analysis program. TEchim uses STAR aligner (Dobin et al., 2013) to screen sequencing data for reads that span the junction between a genomic locus and a consensus transposon sequence, and BLAST (Altschul et al., 1990) to extract information about the transposon insertion site at single-nucleotide resolution. The aim of TEchim is to extract high-fidelity contiguous breakpoint-spanning reads, which distinguishes it from other approaches such as those combined in the integrated analysis pipeline “McClintock” (Nelson et al., 2017). TEchim generates nucleotide contigs from gDNA or cDNA sequencing reads, then creates *in-silico* paired-end reads and screens them for cases where one end maps to a gene and the mate read maps to a transposon. Since these *in-silico* reads are derived from contiguous sequences, one can refer back to the original reads to determine transposon-gene breakpoint sequence. TEchim also generates sequencing coverage around insertion sites, which permits estimation of the population frequency of germline insertions. Our gDNA data from 10 individual flies revealed a range of population frequencies for transposons in inter- and intragenic regions (Supplemental Table S2). In the subsequent analyses we focus on insertions detected in at least 50% of flies tested. We found highly penetrant *copia*, *opus*, *micropia*, *blood* and *412* insertions in *twin of eyeless* (*toy*), *mushroom-body expressed* (*mub*), *Rhodopsin 7* (*Rh7*), *CG32521* and *Tequila* (*teq*), respectively. Expression of these genes mirrored the pattern of the transposon they harbored (Figure 1B-F, second graphs). Neural expression of these transposons in αβCherry flies therefore appears to be driven by these nearby genes.

We next assessed whether all our annotated transposons exhibited patterned midbrain expression. Re-clustering the scRNA-seq data using transposon expression generated 78

clusters that mostly contained cells from all 8 biological replicates (Figure 1G, Supplemental Figure S3), indicative of stereotyped transposon expression between different flies from the same strain. Analysis of cellular gene expression across the transposon clusters showed that many clusters preferentially expressed certain genes. For example, the cluster expressing *Stalker2* LTR was enriched for cells also expressing *Heat-shock-protein-70Bc* (*Hsp70Bc*) (Figure 1H), and cells in the *Doc*-cluster had high *Furin 1* (*Fur1*) (Figure 1I). Referring back to the gDNA revealed that αβCherry flies harbor a *Stalker2* copy within *Hsp70Bc* and a LINE-like *Doc* element inside *Fur1*. Again, these data suggest expression of *Stalker2* and *Doc* is driven by a neighboring gene.

**Quantitative analysis reveals high fidelity transposon-gene coexpression**

Our gDNA analysis also revealed many transposons inside genes that were more broadly expressed across the brain. In total, we identified 4306 germline transposons (Supplemental Table S2 displays all sites where the upstream breakpoint was detected in at least 50% of flies tested); 2163 of these lie outside, and 2143 sit within, a gene – from here-on denoted the neighboring gene. Of these, 910 cases were inserted in the same direction as the gene, 1175 in antisense orientation and 58 in loci within genes in both orientations. To quantify the correlated expression of transposons and cellular genes we devised a method based on the Hardy-Weinberg principle for quantifying linkage equilibrium of two alleles in population genetics (Lewontin and Kojima, 1960) (Figure 2A). We binarized our scRNA-seq data to generate the equivalent of bi-allelic traits in a population (see methods for details). The proportion of cells expressing a specific transposon was calculated, multiplied by the proportion of cells expressing a certain gene, and then this value was subtracted from the proportion of cells that expressed both the transposon and gene. We termed this value the Coexpression Disequilibrium, CD. These CD values were normalized to account for variable abundance of each transposon and gene in every transposon-gene pair and the analysis was repeated for all transposon-transposon and gene-gene pairs. Normalized values were then ranked within each of the 8 biological replicates and *p*-values calculated and corrected

for multiple comparisons (Benjamini-Hochberg). These values describe the probability that a transposon-gene pair would have such a highly ranked CD across multiple replicates if they were expressed independently.

# Figure 2

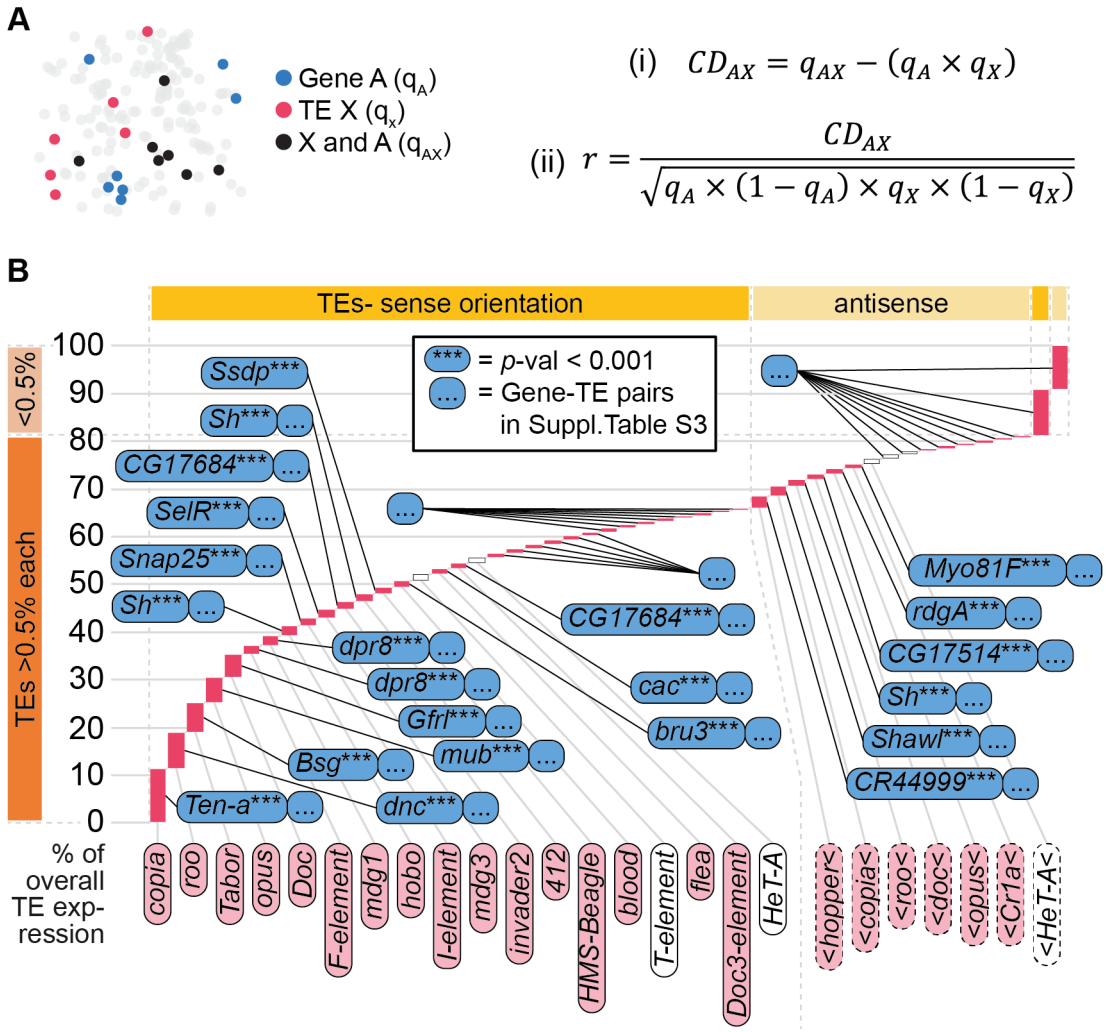



**Figure 2. Most transposons are coexpressed with neighboring genes.**

**A** Schematic and formulae describing the calculation of Coexpression Disequilibrium ($CD_{AX}$) values. **B** Examples of transposon-gene pairs that are neighboring in the genome and coexpressed across the midbrain. Height of pink bars shows relative transposon expression levels in scRNA-seq data. Transposons contributing to >0.5% of overall transposon expression, indicated by dark orange bar on bottom left, are individually displayed and the associated gene with the lowest corrected *p*-value is indicated for each one. Transposons contributing <0.5%, indicated by light orange bar on top left, are pooled into sense- and antisense expression. Transposons are also organized horizontally into sense (left side of plot marked with dark yellow bar on top) and antisense expressing elements (right side of

plot, light yellow). See Supplemental Table S3 for entire list of correlated transposon-gene pairs.

We combined the list of all detected germline transposon insertions in αβCherry flies with the scRNA-seq data generated from these flies and calculated CD values between every transposon and its neighboring gene (Supplemental Table S3). For all transposons that contributed to ≥ 0.5% of overall expression we found at least one copy inside a gene that exhibited a correlated expression pattern (Benjamini-Hochberg corrected p <0.05, Figure 2B). Exceptions were the telomeric *TART-element*, *TART-A* and *HeT-A* which are likely to be autonomously expressed. Transposons inserted in the same orientation as the gene's transcription unit had correlated expression of the sense strand of the transposon with that of the gene. In contrast, the antisense strand was correlated for reverse orientation transposons. Since the number of transposon copies, and therefore the number of potentially correlated neighboring genes, varied between 1 (for e.g. *accord*, *1731*, *Tirant*, etc.) and 91 for *roo* in an antisense orientation, we tested whether the same number of randomly chosen (not neighboring) genes would exhibit similar coexpression patterns with transposons. We randomly selected 10 sets of 2143 genes and counted the number of transposon-gene pairs with correlated expression (below the *p*-value threshold of 0.05) in each gene set. We then performed a chi-square test using the mean number of randomly correlated pairs as the expected frequency if there was no interaction between transposons and neighboring genes (Supplemental Table S4). These analyses demonstrated that a neighboring gene significantly influences the expression pattern of almost all transposons in the fly brain.

**Transposons become part of chimeric transcripts with cellular mRNAs**

We next tested whether observed coexpression of transposons and neighboring genes might result from chimeric mRNAs formed from the transposon-gene pairs. We extracted mRNA from αβCherry fly heads and generated 250 basepair long reads which were screened for chimera using TEchim. Incorporating a function in TEchim that maintains strand-specificity of input reads enabled unambiguous assignment of chimera to cellular genes. We found that a large number of intronic transposons give rise to chimeric pre-

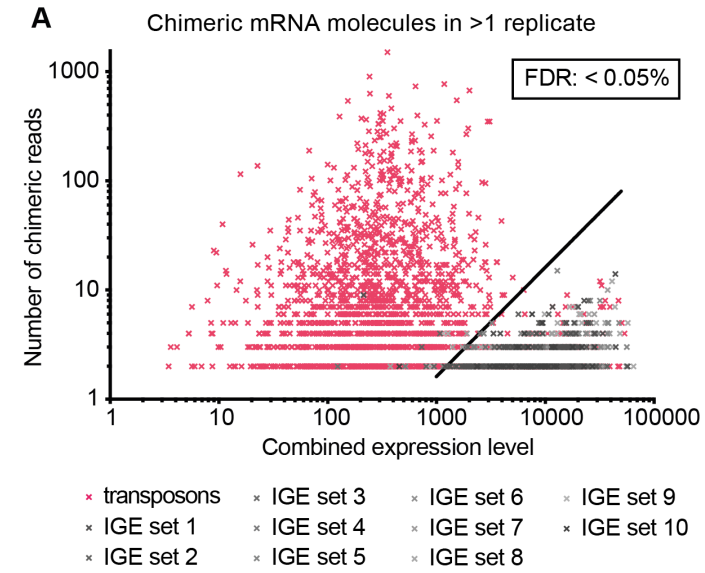
mRNAs. In total, we retrieved chimeric mRNA segments from 4732 different genomic loci, with 2430 spanning a gene-to-transposon (5´ to 3´) and 2302 a transposon-to-gene junction (Supplemental Table S5). These pre-mRNAs were poly-adenylated and frequently contained intron and transposon sequences. Importantly, qPCR-, bulk- and scRNA-seq analyses would count these transposon-containing pre-mRNAs as evidence for transposon expression. Chimera included sequences from LTR, LINE-like and DNA transposons attached to mRNAs from genes involved in many biological processes. For example, we found sequence from the LTR-retrotransposon *gypsy* in transcripts of the ubiquitin gene *Ubi-P5E* and of *highwire* (*hiw*), encoding a neuron specific ubiquitin ligase, the non-LTR element *Doc* in *Fur1*, encoding a synaptic membrane bound protease, and the TIR element *hobo* attached to transcripts from *Shaker*, which encodes a voltage-gated potassium channel (Izquierdo, 1994; Kaplan and Trout, 1969; Roebroek et al., 1991; Wan et al., 2000).

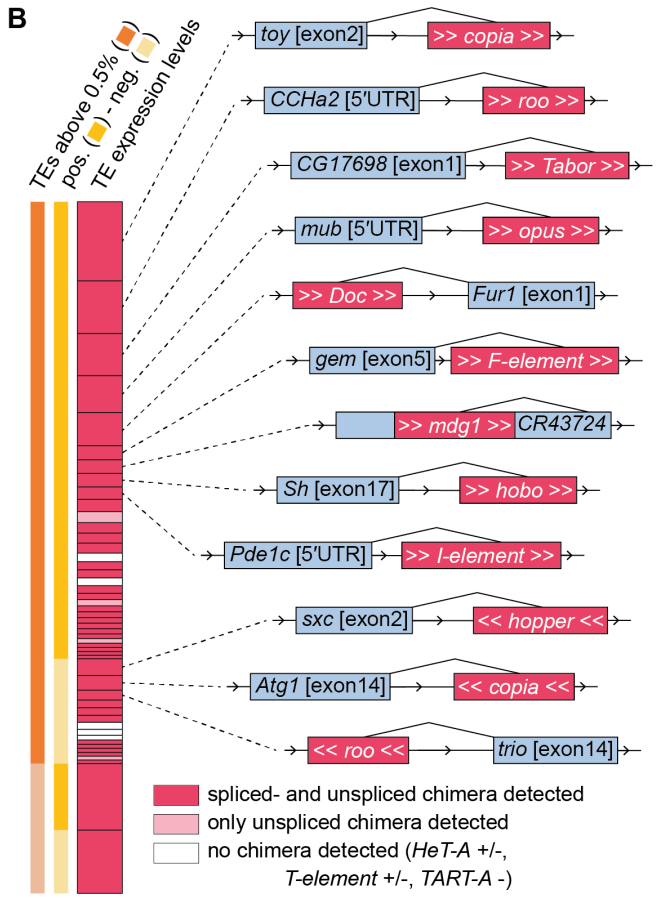
**“Immobile genetic elements” generate a threshold to exclude amplification artifacts**

Previous studies of transposon mapping have established that *in-vitro* amplification of DNA often leads to chimeric amplification artifacts (Evrony et al., 2016; Treiber and Waddell, 2017). We therefore accounted for similar errors in our mRNA data by calculating the rate of amplification artifacts with 10 sets of 167 exons that were expressed at the same level as each transposon. These exons cannot relocate in gDNA so we name them “immobile genetic elements” (IGEs). Since IGEs should only occur in one location in gDNA from αβCherry flies. Chimeric reads between IGEs and other genes most likely represent amplification artifacts. As expected, the rate of generating IGE chimeras was correlated to the expression level of the IGE and the gene it formed a chimeric molecule with. Critically, the IGE chimera rate was substantially lower than that formed between genes and transposons (Figure 3A). We therefore used prevalence of IGE chimera to define a false discovery rate (FDR) of 0.05%. The FDR was calculated by dividing the number of IGE chimera per total chimera (i.e. including transposon chimera). This 0.05% threshold resulted in an average of 1.9 IGE hits per 2165 total chimera (Supplemental Figure S4, Supplemental

254 Table S6). All chimeric transcripts presented in this study were detected with an FDR below
255 0.05%.

# Figure 3




256

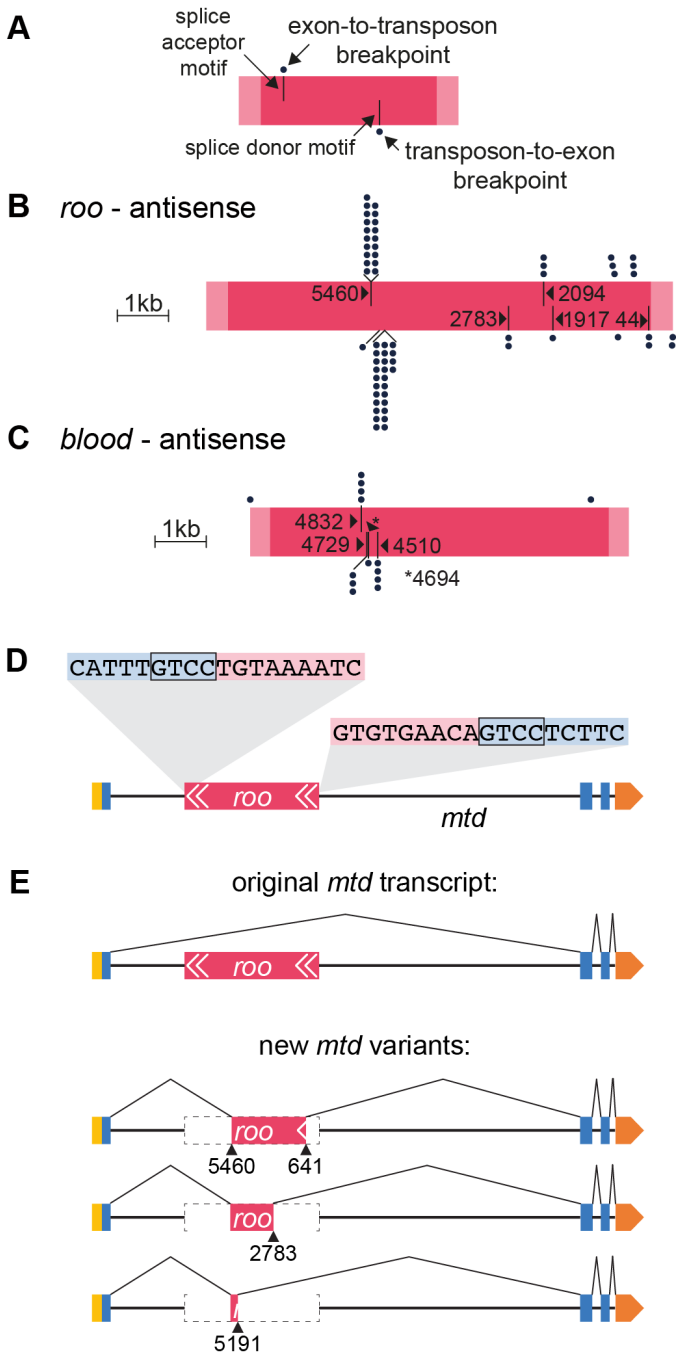
**Figure 3. Chimeric transposon-gene mRNA is abundant in the midbrain.**

**A** Graph showing number of chimeric reads, and combined expression levels of each transposon-gene pair (pink), and in 10 sets of IGE-gene pairs (grey). Combined expression levels are the square root of the product of reads in our bulk RNA data for both transcripts of a transposon/IGE-gene pair. IGEs were used to calculate a FDR<0.05%. **B** Examples of transposon-gene pairs for which chimeric mRNAs were detected. Pink bar represents total transposon expression in scRNA-seq data (as in Figure 2B), grouped into sense- and antisense, as well as contributing to more- and less than 0.5% of total transposon expression. Dark pink bars indicate that both pre-mRNA and mature spliced mRNA chimeric fragments were detected. Light pink indicates only unspliced chimera were found. Schematics show splice sites between transposon and the neighboring gene (grey and pink boxes are not to scale). For list of all chimera see Supplemental Table S5.

**Many transposons introduce cryptic alternative splice-sites into cellular genes**

Transposon sequences could be removed from the unspliced chimeric pre-mRNAs to yield intact host mRNAs, and full-length transposon sequences. However, for most transposon sub-families we found at least one neural gene where breakpoint-spanning reads indicate that specific sections of a transposon are spliced into host-gene transcripts (Figure 3B, Supplemental Table S5 – spliced insertions are labeled in column 2). Analysis of the breakpoints inside transposons at these 264 sites revealed that chimera were formed at conserved locations in each transposon type. For example, where antisense *roo* resided within an intron, we found transcripts where the 3´-end of an upstream exon was fused to either a section of *roo* beginning at position 5460 (for 19 different loci) or 2094 (3 loci), and also at several additional breakpoints with lower frequency (Figure 4A,B). In addition, we identified transcripts where sections of *roo* were bound to the 5´-end of a downstream exon. 3´ breakpoints at position 5191 of *roo* spliced into transcripts of 24 genes, two genes from position 2783 of *roo*, and several others from unique positions in *roo* (note numbering runs backward because it relates to forward orientation of *roo*). Whereas intronic antisense *roo* provided gene-transposon breakpoints for 28 exons, and transposon-gene breakpoints for 33, intronic sense *roo* only introduced 4 and 1, respectively (Supplemental Table S5). Similarly, the LTR *blood* contributed more breakpoints when inserted in antisense orientation relative to the host gene (14 vs. 6, Figure 4C and Supplemental Table S5).

# Figure 4




288

**Figure 4. Transposons introduce splice sites at conserved locations.**

**A** Key for labelling scheme in panels B and C. Pink bar represents the transposon; light pink ends indicate LTRs and dark pink the core sequence. Positions of dots above the bar represent sites on the transposon where an upstream exon splice donor (SD) has merged. Every dot represents a different gene. Black lines in the top half of pink bar represent splice acceptor (SA) motifs in the transposon. Dots below the pink bar indicate location of breakpoints on the transposon that splice to upstream exonic SA sites of different genes. Bars in the lower half indicate SD motifs. **B-C** Representations of antisense *roo* and *blood* (to scale), with all breakpoints to SA and SD sites of neighboring genes. Note, the frequently used site on antisense *roo* at position 5191 is a non-consensus SD site, lacking the expected GT motif at the immediate breakpoint. The sequence around 5191 resembles a consensus SD motif, although the GT is a GC. Compare TTTGGCAAGTT to motif in Supplemental Figure S5A. **D** Illustration of antisense *roo* insertion in the *mustard* (*mtd*) gene. Only one isoform of *mtd* is shown. Yellow box represents 5´UTR, blue boxes are exons, orange box 3´UTR, pink represents *roo* transposon with white arrows indicating LTRs. Breakpoint-spanning gDNA reads reveal Target Site Duplication (TSD, inset). **E** Schematic of original *mtd* transcript, and of three new splice isoforms.

We screened transposon sections around breakpoints for consensus splice-acceptor (SA) and donor (SD) sequence motifs. Often gene-to-transposon chimera formed at SA consensus motifs, and transposon-to-gene chimera at SD motifs (Stephens and Schneider, 1992) (Supplemental Figure S5, Supplemental Table S7). For example, all breakpoints in antisense *blood* formed with more than one exon were precisely located at predicted SA and SD splice sites (see vertical lines in Figure 4C). A consensus SD motif was not evident at position 5191 of antisense *roo*, although it frequently provided 5´-sequence to transposon-gene chimeric RNAs (Figure 4B). However, sequence around position 5191 resembles the consensus, with exception of a GT-to-GC conversion (see Supplemental Figure S5). Taken together, our analysis revealed that transposons introduce many alternative splice sites, which are recognized by the host cell spliceosome to join cellular exons to sections of transposon.

We also identified alternative splicing to different sites within the same transposon insertion. Again using *roo* as an example, αβCherry flies harbor an intronic reverse orientation *roo* in the pan-neurally expressed *mustard* (*mtd*) gene, which to date has only been implicated in innate immunity (Wang et al., 2012) (Figure 4D). The wildtype *mtd* locus produces many splice variants and RNA-seq revealed a complex collection of additional *mtd* splice variants that incorporated different *roo* fragments (Figure 4E). SD sites upstream of this *roo* came from either *mtd* exon 11 or 13 (annotated exons are numbered backwards) and these spliced to the SA at position 5462 within *roo* (Figure 4E). Three different SD sites (at positions 641, 2784 and 5191) within *roo* spliced out to the closest downstream SA (exon 6) of *mtd*. This *roo* substantially increases the *mtd* mRNA isoform repertoire; without *roo* the locus can express 23 *mtd* isoforms, with *roo* it can generate 68 differentially spliced mRNAs.

The transcript diversity of 263 other genes was similarly increased by a transposon. These transcripts incorporate 66 different transposon families with each introducing cryptic SA and/or SD sites into host genes (see Supplemental Table S5). For example, chimeric reads

indicate that transcription of *Dscam2,* which encodes the transmembrane Down Syndrome cell adhesion molecule 2, is frequently initiated inside a sense insertion of *blood* which spliced into exon 33 (the second exon) of *Dscam2*. This splicing combines ORF2 of *blood* with the remaining *Dscam2* exons and aligns the reading frames, generating a novel N-terminus (Supplemental Figure S6). We also found evidence of transposons resulting in exon skipping (e.g. *412* inside *Tequila*, Supplemental Figure S7). Most transposon chimera resulted from intronic insertions. However, an exonic *hobo* in the *CG31705* gene introduced a cryptic SA spliced to the upstream SD from the first *CG31705* exon, creating a truncated mRNA (Supplemental Figure S8). These data show that many *Drosophila* transposons are alternatively spliced into cellular mRNAs increasing the isoforms of a large number of neurally expressed genes.

**Alternative splicing into and out of transposons can be highly penetrant**

Chimeric transcripts could be inconsequential if they only constitute a small percentage of the overall transcript repertoire of a gene. To quantify how frequently a transposon harboring gene produces chimeric mRNAs we analyzed loci where a transposon splices into an exon-intron junction. For each gene we counted the number of reads spanning the transposon-exon boundary, and the number spanning the exon immediately up- and downstream of the transposon. For some genes, most mRNAs contained transposon sequences. For example, 95.3% of all *Rhodopsin 7* transcripts included *micropia* in the 3´ UTR (Figure 5A), and all mRNAs of the *Allatostatin C receptor 1* (*AstC-R1*) contained a section of *F-element*, spliced into one of two different SA sites in the gene (Figure 5B). In addition to the *blood* insertion in *Dscam2* mentioned above, we also found a *Doc* insertion in *Dscam2*, which contributed to around a third of all transcripts initiated at the *Dscam2* transcription start site (Figure 5C). We also found a sense-orientation *flea* in the X-linked *cacophony* (*cac*), which encodes a voltage-gated calcium channel (Smith et al., 1996). This *flea* insertion truncated 12.4% of *cac* transcripts, potentially deleting the last 8-11 coding exons and suggesting that many

αβCherry males are likely mutant for the *cac* gene (Supplemental Figure S9). Another interesting example on the X Chromosome of αβCherry flies is a sense *opus* insertion in *Beadex* (*Bx*), which encodes a long-term memory relevant LIM-type transcription factor (Hirano et al., 2016). This *opus* produces at least two new *Bx* mRNAs (Supplemental Figure S10), which constitute 4.9% of all *Bx* transcripts. On average, transposons contributed 11.6% of transcripts derived from a gene (Supplemental Table S5).

# Figure 5

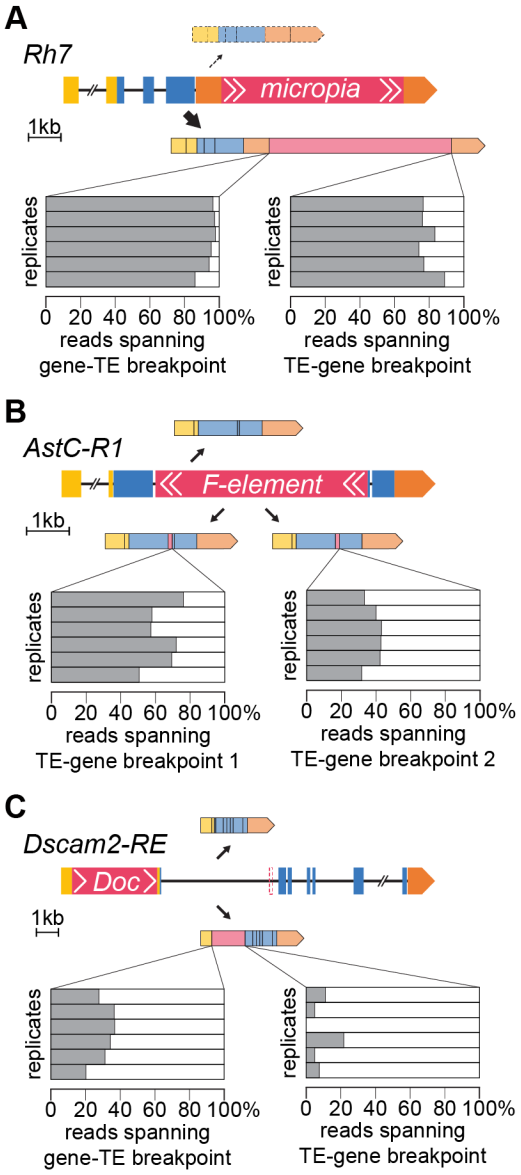



**Figure 5. High penetrance of transposon-containing splice isoforms.**

**A** Schematic showing *Rhodopsin 7* locus harboring a sense *micropia* in the 3´-UTR, and two splice isoforms. Grey bars show percentage of reads spanning the gene-TE (left) and TE-gene (right) breakpoint in each of the 6 tested replicates. **B** The *AstC-R1* gene harbors an antisense *F-element* immediately upstream of the second exon which introduces cryptic splice-sites. Three spliced isoforms are shown. **C** *Dscam2* harbors a sense *Doc* in its 5´-UTR (in addition to a *blood* insertion in its first intron, see Supplemental Figure S6). For ease of visualization, only the shortest *Dscam2* isoform -RE is depicted. The *blood* insertion is indicated with a dashed box. See Supplemental Table S5 for complete list.

**Splicing into transposons is common and varies between strains**

Transposons are highly variable between fly strains. We therefore analyzed three previously published mRNA sequencing data sets from other fly strains for chimeric transposon-gene mRNAs (Croset et al., 2018; Hemphill et al., 2018; MacKay et al., 2012). Although these prior studies generated shorter paired-end RNA-seq reads, we still found chimeric mRNAs in all three data sets (Supplemental Table S8). Some chimera were conserved across all strains, whilst others appeared to be strain-specific. 466 of the 1332 chimera identified in at least 2 samples of αβCherry flies were present in at least one of the three other strains, whereas 92 of those occurred in all four strains. Chimera that were not detected in other strains could indicate genomic heterogeneity between strains, or absence of evidence resulting from lower sequencing coverage. Nevertheless, these results demonstrate the prevalence of cellular mRNAs containing transposon sequence.

**Transposon expression is predominantly non-autonomous**

Finding that neural expression of consensus transposon sequences is highly correlated with at least one neighboring gene, and that most transposon sequence is part of spliced chimeric mRNAs, implies that expression is largely driven by neighboring genes. Testing this hypothesis further requires comparing the number of reads mapping to a specific transposon with the abundance of breakpoint-spanning reads for that same transposon. However, most transposons are multi-copy (6 is the median copy number in αβCherry flies (Supplemental Table S3)) so a read mapping inside a transposon cannot be assigned to a specific copy. To overcome these challenges, we quantified the average number of reads that only map to a given transposon consensus sequence (TE-only) per nucleotide for each transposon across our 6 biological replicates. Next, we counted the number of locus-specific reads that span each transposon and a genomic region (TE-gene). We reasoned that autonomous transposon expression should exclusively generate TE-only reads, while non-autonomous expression driven by a neighboring gene, should generate similar numbers of TE-only and TE-gene reads. The number of TE-gene reads was higher than the average number of TE-

406 only reads (normalized to transposon length) for every transposon tested (Figure 6A,
407 Supplemental Table S9), suggesting expression is non-autonomous. We also tested
408 autonomous- vs. non-autonomous expression by analyzing transposons with LTRs at both
409 ends. Autonomous expression of LTR elements should not result in reads upstream of the
410 element's 5´ LTR (5´-gene-LTR-3´ reads). Quantifying the ratio of 5´-gene-LTR-3´ reads and
411 5´-LTR-TE-3´ reads revealed that most LTR retrotransposons expressed in the head
412 generate roughly equivalent numbers of each fragment. Breakpoint-spanning reads at the 3´
413 ends of LTR retrotransposons revealed a similar situation (Figure 6B, Supplemental Table
414 S10). These analyses provide further evidence that LTR transposons are predominantly
415 expressed as chimeric mRNAs with cellular genes, rather than as autonomous elements.

# Figure 6

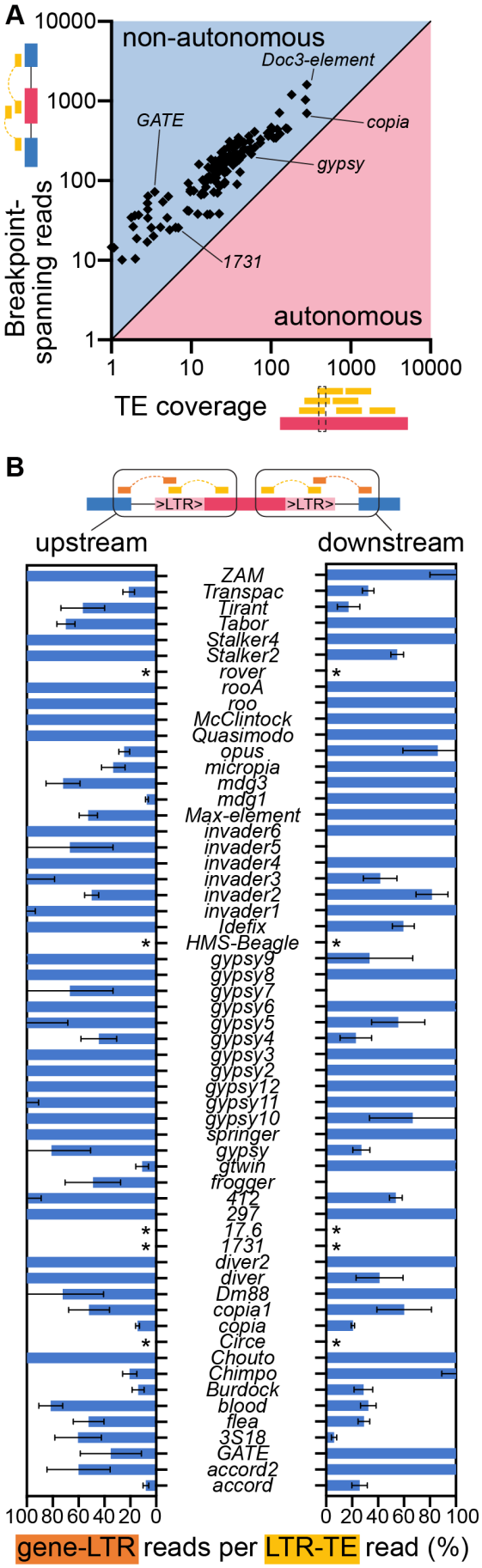


**Figure 6. LTR retrotransposon expression is predominantly non-autonomous**

**A** Plot showing the average number of reads per nucleotide (x-axis), and total number of transposon-gene spanning reads (y-axis) for every tested transposon. Number of spanning reads is higher for every transposon. **B** List of all LTR retrotransposons analyzed in mRNA data. LTR-gene spanning reads were identified for every LTR transposon expressed in the midbrain. Numbers represent percentage of reads spanning LTR-gene versus LTR-TE breakpoints. Values are capped at 100%, but some transposons produced more LTR-gene than LTR-TE reads (see Supplemental Table S10). Error bars represent standard deviation. Please note, transposon reference sequences did not contain LTR sections for the 5 transposons, indicated with *.

## Discussion

Combining single-cell expression data from the *Drosophila* midbrain with high-coverage gDNA sequence of the same fly strain revealed that most transposons are expressed as parts of chimeric mRNAs with cellular genes.

Several prior studies have documented that transposons are transcriptionally active in somatic tissue. These reports employed methods that either generate cDNA fragments (RNA-seq, e.g. De Cecco et al., 2013)) or amplify short sections of transposon mRNAs (RT-qPCR, e.g. in Guo et al., 2018; Li et al., 2013; Sun et al., 2018). However, these approaches cannot distinguish between autonomous transposon expression and chimeric transposon-gene mRNAs investigated in this study. Baseline and changing cell-specific expression of host genes that produce chimeric transcripts with transposons could therefore be misinterpreted as cell-restricted autonomous transposon expression, with potential for mobilization.

Some studies of transposon expression use cap analysis gene expression (CAGE), to distinguish pre-mRNA from 5´-ends of mature mRNAs (Faulkner et al., 2009). Although, CAGE-reads mapped to transposons represent transcripts where transcription started within a transposon, we identified 243 chimera that initiated inside (or at the start of) a transposon and spliced into a downstream exon of a gene (Supplemental Table S5). Short 5´ ends CAGE reads would rarely identify such chimeric transcripts. In theory, a combination of long-read sequencing (e.g. Pacbio (Rhoads and Au, 2015) or nanopore (Deamer et al., 2016)) and ways to identify 5´ caps and 3´ poly(A) tails could discover full-length transposon mRNAs.

Our study illustrates the utility of the *Drosophila* brain to study genome-wide expression of transposons. The single-cell atlases of the entire brain allows transposon expression to be assigned to specific cell types (Allen et al., 2020; Croset et al., 2018; Davie et al., 2018;

Konstantinides et al., 2018). This is made easier by transposon sub-families in *Drosophila* being very discrete with even related elements having different sequence. In addition, some of these transposons are low copy and even detected within one gene (Supplemental Table S3). This makes it simple to map their expression to cells and to significantly correlate their expression to that of a neighboring gene. In contrast, for a high copy number transposon resident in >10,000 genes (cf. LINE-1 in most mammalian genomes), it becomes impossible to distinguish a correlation from chance, because the transposon expression would also be correlated with at least one of 10,000 randomly chosen genes.

We complemented scRNA-seq analyses of transposon coexpression with neighboring genes with discovery of >833 chimeric transposon-gene transcripts using bulk RNA-seq. Chimeric transposon-gene fragments were identified in previous studies, with some focusing on individual genes, and others analyzing exonized transposons genome-wide (Kapusta et al., 2013; Van De Lagemaat et al., 2003; Nekrutenko and Li, 2001). However, to our knowledge, no other study investigated the proportion of transposon expression in somatic cells that is comprised of exonized transposon fragments. We found that transposon exonization is highly prevalent in the *Drosophila* brain and is likely the main driver of somatic transposon expression. Since we mapped reads to consensus transposon sequences we may have missed exonization of older transposons that have accumulated many mutations.

We introduce three new pieces of software that should be helpful to other researchers in the field. Although they were developed to analyze *Drosophila* data, they can be readily adapted for sequence data from other species. The three main components are (1) scTE-seq, a tool to map scRNA-seq data onto a masked reference genome and consensus transposon sequences, (2) scRNA-seq-Hardy-Weinberg (scHW), which implements the new method presented here to analyze expression correlations, and (3) TEchim, which combines all analysis steps for identification, characterization and quantification of chimeric transcripts in

bulk mRNA sequencing data, and includes IGE analysis to determine the rate of amplification artifacts for each sample.

We found the expression of many transposons to be restricted to small groups of cells. For example, *blood* was highly expressed in most glia, but silent in neurons. In contrast, *gypsy* was detected in some neurons but was absent in glia. Somatic transposition in neurons and glia has been implicated in age-dependent neuronal decline in wildtype and disease models of *Drosophila* (Guo et al., 2018; Li et al., 2013; Sun et al., 2018). Our results constrain these models because mobilization can only occur in cells that express full-length elements, or transposon mRNAs that encode enzymes permitting other elements to move in *trans*. Therefore, the *gypsy* retrotransposon is only likely to mobilize in glia, if the fly strain studied harbors a copy of *gypsy* in a glial-expressed gene (Krug et al., 2017). Expression below that typically detectable using scRNA-seq could generate full-length transposon mRNAs that reintegrate in the genome. For example, two LINE-1 elements on human Chromosomes 8 and 13 were shown to mobilize in the human brain (Evrony et al., 2015; Sanchez-Luque et al., 2019). However, data in this study, which include higher coverage bulk sequencing data, and our earlier study of the rate of somatic transposition (Treiber and Waddell, 2017) indicate that transposon transcripts in the fly brain most frequently represent diversification of the neural transcriptome, rather than mobilization.

At this stage we are unable to conclusively demonstrate the biological impact of transposon-gene chimera. The process of transposable elements acquiring new cellular functions that benefit the host cell has been coined transposon ‘exaptation’ (Gould and Vrba, 2013). A striking example of this is the neuronally expressed *Drosophila* and rodent Arc proteins, which resemble Ty3/*gypsy* retrotransposon-encoded *gag*. Arc also forms virus-like capsids and binds sequences in the 3′ UTR of *Arc* mRNAs, which enables their intercellular transport (Ashley et al., 2018; Pastuzyn et al., 2018; Zhang et al., 2015). We found a broad range of neural genes for which a substantial proportion of their mature mRNA transcript pool

contained transposon sequences. Sometimes transposon sequence is within the open reading frame, and other times it is positioned in 5´ or 3´ UTRs where it could alter traffic and/or translation. However, it is difficult to determine the whole-genome functional consequence of splicing into transposons, because we often only retrieve the sequence across the splice junctions. Furthermore, although each transposon has a known consensus sequence, individual copies are polymorphic. Nevertheless, our sequencing shows that transposon exonization often truncates and/or changes the amino acid sequence of the encoded proteins, potentially changing structure and function. We also identified several examples where inclusion of transposon sequence conserved the reading frame of the host gene and may generate a novel chimeric protein. Amongst the 264 transposon harboring genes identified in this study, there are several that we have described in detail for which disruption and altered expression of the locus would be expected to have significant consequences for neural function. Flies harboring *hobo* in *Sh* and *flea* in *cac* might exhibit altered voltage-gated currents, whereas those with *roo* in *AstA-R1* will respond differently to the modulatory Allatostatin A neuropeptide (Larsen et al., 2001; Smith et al., 1996). We also described insertions of *412* in *teq* and *opus* in *Bx*, two genes which have been implicated in long-term memory formation (Didelot et al., 2006; Hirano et al., 2016). The *412* insertion in *teq* is particularly interesting in light of several behavioral studies that have used a mutant fly strain where *teq* function is apparently impaired by a piggyBac transposon in the 3´-UTR (Didelot et al., 2006; Thibault et al., 2004). It seems likely that a *412* in the coding region will have at least as disruptive an effect on *teq* function as a 3´-UTR insertion.

We also discovered many cases where a single intronic transposon introduced several cryptic splice sites, and thereby increased the transcript repertoire of the host gene. For example, the antisense *roo* inside the innate-immunity gene *mtd* resulted in many new predicted protein isoforms. This *roo* insertion could increase allele diversity and enable the innate immune system to broaden its effectiveness against a wider range of pathogens.

RNA-seq data from other fly strains suggests that more than half of the chimeric transposon transcripts identified in αβCherry flies are unique to this strain. This finding alone demonstrates the incredible heterogeneity of transposons between strains. In addition, our prior genome sequencing revealed large differences between individual αβCherry flies (Treiber and Waddell, 2017). It seems likely that polymorphic transposons and differential distribution across the genome could contribute towards heterogeneity of neural function, and neurological pathology, between individual animals.

## Methods

### Fly strains

All experiments used αβCherry flies, which were generated by crossing MB008b females (Aso et al., 2014) with w-; +; UAS-mCherry males. Flies were raised on standard molasses food at 25°C, 40-50% humidity and 12 h:12 h light-dark cycles.

### Bulk mRNA sequencing

For RNA extraction, groups of ~50 flies were frozen in liquid nitrogen and vortexed for 6 x 30s to separate body segments. Heads were isolated using a sieve. To avoid gDNA contamination, mRNA was purified with a combination of protocols. Samples were first processed with a column-based kit (RNeasy Mini kit, Qiagen, UK), including on-column DNase I digestion. Next, mRNA was extracted from total RNA using oligo(dT) magnetic beads (NEB, Ipswich, MA) and mRNA was purified again using RNA columns. Finally, sequencing libraries were generated using oligo(dT) magnetic beads from a strand-specific mRNA library preparation kit (TruSeq, Illumina, San Diego, CA), with 17 cycles of PCR amplification. Fragmentation was optimized to obtain ~350nt long fragments. Whole-genome sequencing was performed on a HiSeq 2500, with 250nt paired-end reads.

### Single-cell read alignments

The *Drosophila melanogaster* reference genome release 6.25 was used for all sequence alignments (Hoskins et al., 2015). Transposon reference sequences were from Repbase (Jurka, 2000; Kaminker et al., 2002). Repetitive sequences in the *Drosophila* reference genome were masked using RepeatMasker (Smit et al., 2015), and a single consensus sequence copy of each transposon was added to the reference genome. Consequently, each transposon was treated as a separate “chromosome” by the down-stream analysis software. Single-cell sequencing data was processed with the DropSeq pipeline, as described (Croset et al., 2018; Macosko et al., 2015) and Digital Gene Expression (DGE)

matrices were processed using using Seurat in R (R Core Team, www.R-project.org, Vienna, Austria) (Butler et al., 2018).. A detailed protocol is provided in the Supplemental Methods. The modified reference genome and refFlat file are provided as Supplemental Files 1 and 2. Mapping efficiency was assessed by comparing the number of reads mapped to consensus transposon sequences with fractional read counts estimated by RepEnrich2. Consensus reads were quantified using SAMtools idxstats on the sorted and indexed output BAM files following STAR alignment in the scTE-seq pipeline. Fractional read counts were computed using standard RepEnrich2 parameters and the most recent transposon insertion library downloaded from RepeatMasker (db20140131) for each of the 8 biological replicates. Least-square linear regression was computed using GraphPad Prism (version 8, San Diego, California, USA) with default parameters.

**Coexpression analysis**

Expression levels of every annotated gene and transposon (i.e. feature) were binarized (expression ON/OFF) in the scRNA-seq data using a dynamic threshold for UMI counts. The threshold was chosen to separate the lower third of UMI counts (OFF) from the rest (ON). Next, the Coexpression Disequilibrium (CD) was calculated for each transposon-gene pair as described in the main text and Figure 2A, resulting in a CD-matrix. Normalized CD values of each transposon with every feature were ranked in each replicate. For coexpression analysis, the mean ranks across all 8 replicates of all features were first calculated. Next, a one-sample *t*-test was conducted with each CD value, and with the expected value μ set to the mean ranks. *P*-values were corrected for multiple comparisons using Benjamini-Hochberg correction. This process was repeated with a set of 10 randomly assigned features for each transposon. Finally, a chi-square test was performed, with the number of correlated features between each transposon and a randomly assigned feature as the expected value. Statistical analyses were performed in R.

**Mapping transposon insertions (gDNA and mRNA)**

Germline transposon insertions were mapped with single-nucleotide resolution using previously published gDNA data from αβCherry flies (Treiber and Waddell, 2017). Chimeric transcripts were detected by analyzing bulk mRNA data generated for this study. A new, purpose-built, multi-functional sequence analysis pipeline called TEchim was developed for both these tasks. TEchim has 6 key functions: 1. generation of support files, including a masked reference genome and endogenous intron-exon junctions. (input files: reference genome, list of genes, list of transposon sequences). 2. alignment of un-stranded genomic DNA sequence data of multiple sequencing lanes and multiple biological replicates, detection of chimeric sequence fragments with single-nucleotide resolution, the sequencing coverage around insertion sites, and the generation of summary output tables. 3. alignment of stranded cDNA data, detection of chimeric fragments, quantification of reads 4. generation of matching immobile genetic elements (IGE, see main text), analysis of these IGEs. These data are then used to determine sample-specific detection thresholds. 5. Quantification of LTR-gene and LTR-transposon reads (see Figure 6B). 6. Quantification of locus-specific breakpoint-spanning reads. For key function 1, the reference genome was first masked using RepeatMasker (Smit et al., 2015), [parameters: -no_is -s] using the same library of transposon consensus sequences as for mapping the scRNA-seq data (see above). In addition, several files were created that contain information about gene features and that were required for subsequent TEchim analysis steps. For key functions 2 and 3, paired-end sequencing reads were first merged using FLASH (Magoč and Salzberg, 2011) [parameters: -x 0.15 (maximum allowed ratio between the number of mismatched base pairs and the overlap length) -M 170 (maximum overlap)]. Next, *in-silico* paired-end reads were generated from contiguous sequences. For cDNA input, the strandedness was preserved throughout the analysis. *In-silico* reads were aligned using the STAR aligner (Dobin et al., 2013) and the masked genome (described above) [parameters: --chimSegmentMin 20 --chimOutType WithinBAM --outSAMtype BAM SortedByCoordinate]. For those *in-silico* read-pairs where one read mapped onto a transposon sequence, and their mate read mapped to a genomic locus in the masked reference genome, long-read contigs were taken and aligned

to (a) the masked reference genome and (b) to consensus transposon sequences using BLAST (Altschul et al., 1990). Reads for which BLAST successfully identified alignments for both the gene- and transposon breakpoint were further processed. For those cases where only the genomic locus could be mapped, the transposon breakpoint was computed from the STAR alignment and the size of the fragment. Pooled results were filtered to ensure that each read was only counted once. These steps were repeated for each sample and sequencing lane separately and individual results were combined by merging breakpoint-spanning reads based on the genomic locus with BEDTools (Quinlan and Hall, 2010), with a window of 20nt, and preserving single-nucleotide breakpoint information on the gene- and transposon sequence. For cDNA data, TE->gene and gene->TE reads and for gDNA data, up- and downstream reads were recorded separately. Pooled hits were intersected with annotated genes, gene features (5´- & 3´ UTRs, exons, introns) and splice sites. Finally, for cDNA data, gene- and transposon expression levels are added to each breakpoint, using SAMtools (Li et al., 2009). Key functions 4-6 are described in the Supplemental Methods. All step-by-step code and a more detailed manual are available on GitHub (https://github.com/charliefornia/TEchim). FlyBase was used for candidate-based gene searches (Thurmond et al., 2019).

**Data from previously published studies**

Raw single-cell sequencing reads from Croset et al., (2018) (PRJNA428955), Hemphill et al., (2018) (PRJNA412381) and Mackay et al., (2012) (PRJNA280097) were obtained from the NCBI Short Read Archive (SRA https://www.ncbi.nlm.nih.gov/sra). Genomic DNA data from Treiber and Waddell, (2017) was obtained from the Dryad Digital Repository (https://doi.org/10.5061/dryad.fd930).

## Data Access

All processed data is presented in Supplemental Tables S1-10. FASTQ files and wiggle tracks of the bulk RNA sequencing data have been submitted to the NCBI BioProject Database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA588978. Scripts are provided as Supplemental Code and can also be accessed via GitHub (https://github.com/charliefornia/TEchim and https://github.com/charliefornia/scHardyWeinberg)

## Acknowledgements

We thank other members of the Waddell group for discussion. CT was supported by a Wellcome Trust DPhil studentship. SW is funded by a Wellcome Principal Research Fellowship (200846/Z/16/Z), ERC Advanced Grant (789274) and the Bettencourt–Schueller Foundation.

## Author Contributions

C.D.T. and S.W. conceived the project and wrote the manuscript. C.D.T. performed and analyzed all experiments.

## Disclosure declaration

Both authors declare no financial and non-financial competing interests.

673 **References**

674

675 Allen, A.M., Neville, M.C., Birtles, S., Croset, V., Treiber, C.D., Waddell, S., and Goodwin,
676 S.F. (2020). A single-cell transcriptomic atlas of the adult Drosophila ventral nerve cord. Elife
677 *9*, 1–32.

678 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local
679 alignment search tool. J. Mol. Biol. *215*, 403–410.

680 Aso, Y., Hattori, D., Yu, Y., Johnston, R.M., Iyer, N.A., Ngo, T.T.B., Dionne, H., Abbott, L.F.,
681 Axel, R., Tanimoto, H., et al. (2014). The neuronal architecture of the mushroom body
682 provides a logic for associative learning. Elife *3*, 1–47.

683 Babaian, A., Thompson, I.R., Lever, J., Gagnier, L., Karimi, M.M., and Mager, D.L. (2019).
684 LIONS: Analysis suite for detecting and quantifying transposable element initiated
685 transcription from RNA-seq. Bioinformatics *35*, 3839–3841.

686 Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F.,
687 Brennan, P., Rizzu, P., Smith, S., Fell, M., et al. (2011). Somatic retrotransposition alters the
688 genetic landscape of the human brain. Nature *479*, 534–537.

689 Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault,
690 M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about
691 transposable elements. Genome Biol. *19*, 1–12.

692 Britten, R.J., and Kohne, D.E. (1968). Repeated sequences in DNA. Hundreds of thousands
693 of copies of DNA sequences have been incorporated into the genomes of higher organisms.
694 Science *161*, 529–540.

695 Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-
696 cell transcriptomic data across different conditions, technologies, and species. Nat.
697 Biotechnol. *36*, 411–420.

698 De Cecco, M., Criscione, S.W., Peterson, A.L., Neretti, N., Sedivy, J.M., and Kreiling, J.A.
699 (2013). Transposable elements become active and mobile in the genomes of aging
700 mammalian somatic tissues. Aging (Albany. NY). *5*, 867–883.

Chung, N., Jonaid, G.M., Quinton, S., Ross, A., Sexton, C.E., Alberto, A., Clymer, C., Churchill, D., Navarro Leija, O., and Han, M. V. (2019). Transcriptome analyses of tumor-adjacent somatic tissues reveal genes co-expressed with transposable elements. Mob. DNA *10*, 1–22.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J. V., and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127–1131.

Criscione, S.W., Zhang, Y., Thompson, W., Sedivy, J.M., and Neretti, N. (2014). Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics *15*, 1–17.

Croset, V., Treiber, C.D., and Waddell, S. (2018). Cellular diversity in the Drosophila midbrain revealed by single-cell transcriptomics. Elife *7*, 1–31.

Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, Ł., Aibar, S., Makhzami, S., Christiaens, V., Bravo González-Blas, C., et al. (2018). A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. Cell *174*, 982-998.e20.

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. Nat. Biotechnol. *34*, 518–524.

Deininger, P. (2011). Alu elements: Know the SINEs. Genome Biol. *12*, 1–12.

Didelot, G., Molinari, F., Tche, P., Comas, D., Milhiet, E., Munnich, A., Colleaux, L., and Preat, T. (2006). Tequila, a Neurotrypsin Ortholog, Regulates Long-Term Memory Formation in Drosophila. Science (80-. ). *313*, 851–853.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A., et al. (2012). Single-neuron sequencing analysis of l1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483–496.

Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley,

729 P., Lehmann, H.S., Park, P.J., et al. (2015). Cell Lineage Analysis in Human Brain Using
730 Endogenous Retroelements. Neuron *85*, 49–59.
731 Evrony, G.D., Lee, E., Park, P.J., and Walsh, C.A. (2016). Resolving rates of mutation in the
732 brain using single-neuron genomics. Elife *5*, 1–32.
733 Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K.,
734 Cloonan, N., Steptoe, A.L., Lassmann, T., et al. (2009). The regulated retrotransposon
735 transcriptome of mammalian cells. Nat. Genet. *41*, 563–571.
736 Gould, S.J., and Vrba, E.S. (2013). Exaptation-A Missing Term in the Science of Form
737 Exaptation-a missing term in the science of form. Paleobiology *8*, 4–15.
738 Guo, C., Jeong, H.H., Hsieh, Y.C., Klein, H.U., Bennett, D.A., De Jager, P.L., Liu, Z., and
739 Shulman, J.M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. Cell
740 Rep. *23*, 2874–2880.
741 Hemphill, W., Rivera, O., and Talbert, M. (2018). RNA-sequencing of Drosophila
742 melanogaster head tissue on high-sugar and high-fat diets. G3 Genes, Genomes, Genet. *8*,
743 279–290.
744 Hirano, Y., Ihara, K., Masuda, T., Yamamoto, T., Iwata, I., Takahashi, A., Awata, H.,
745 Nakamura, N., Takakura, M., Suzuki, Y., et al. (2016). Shifting transcriptional machinery is
746 required for long-term memory maintenance and modification in Drosophila mushroom
747 bodies. Nat. Commun. *7*, 1–14.
748 Hoskins, R.A., Carlson, J.W., Wan, K.H., Park, S., Mendez, I., Galle, S.E., Booth, B.W.,
749 Pfeiffer, B.D., George, R.A., Svirskas, R., et al. (2015). The Release 6 reference sequence
750 of the Drosophila melanogaster genome. Genome Res. *25*, 445–458.
751 International Human Genome Sequencing Consortium, Eric S. Lander, Lauren M. Linton,
752 Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar,
753 and Michael Doyle (2001). Initial sequencing and analysis of the human genome. Nature
754 *409*, 860–921.
755 Izquierdo, M. (1994). Ubiquitin genes and ubiquitin protein location in polytene
756 chromosomes of Drosophila. Chromosoma *103*, 193–197.

757 Jurka, J. (2000). Repbase Update: A database and an electronic journal of repetitive
758 elements. Trends Genet. *16*, 418–420.
759 Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E.,
760 Wheeler, D.A., Lewis, S.E., Rubin, G.M., et al. (2002). The transposable elements of the
761 Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol *3*, 1–20.
762 Kaplan, W.D., and Trout, W.E. (1969). The behavior of four neurological mutants of
763 Drosophila. Genetics *61*, 399–409.
764 Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L.A., Bourque, G., Yandell, M.,
765 and Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin,
766 Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet. *9*, 1–20.
767 Kazazian, H.H. (2011). Mobile DNA transposition in somatic cells. BMC Biol. *9*, 2–5.
768 Kazazian, H.H., and Moran, J. V. (2017). Mobile DNA in health and disease. N. Engl. J.
769 Med. *377*, 361–370.
770 Ketchum, K., Hoskins, R., Wang, X., Smith, T., Gocayne, J., Skupski, M., Wei, M., Smith, H.,
771 Kennison, J., Nixon, K., et al. (2000). The genome sequence of Drosophila melanogaster.
772 Science (80-. ). *287*, 2185–2195.
773 Konstantinides, N., Kapuralin, K., Fadil, C., Barboza, L., Satija, R., and Desplan, C. (2018).
774 Phenotypic Convergence: Distinct Transcription Factors Regulate Common Terminal
775 Features. Cell *174*, 622-635.e13.
776 Krug, L., Chatterjee, N., Borges-Monroy, R., Hearn, S., Liao, W.-W., Morrill, K., Prazak, L.,
777 Rozhkov, N., Theodorou, D., Hammell, M., et al. (2017). Retrotransposon activation
778 contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. PLOS Genet. *13*,
779 1–34.
780 Van De Lagemaat, L.N., Landry, J.R., Mager, D.L., and Medstrand, P. (2003). Transposable
781 elements in mammals promote regulatory variation and diversification of genes with
782 specialized functions. Trends Genet. *19*, 530–536.
783 Lanciano, S., and Cristofari, G. (2020). Measuring and interpreting transposable element
784 expression. Nat. Rev. Genet.

785 Larsen, M.J., Burton, K.J., Zantello, M.R., Smith, V.G., Lowery, D.L., and Kubiak, T.M.
786 (2001). Type A allatostatins from Drosophila melanogaster and Diplotera puncata activate
787 two Drosophila allatostatin receptors, DAR-1 and DAR-2, expressed in CHO cells. Biochem.
788 Biophys. Res. Commun. *286*, 895–901.
789 Lewontin, R.C., and Kojima, K. (1960). The Evolutionary Dynamics of Complex
790 Polymorphisms. Evolution (N. Y). *14*, 458–472.
791 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis,
792 G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools.
793 Bioinformatics *25*, 2078–2079.
794 Li, W., Jin, Y., Prazak, L., Hammell, M., and Dubnau, J. (2012). Transposable Elements in
795 TDP-43-Mediated Neurodegenerative Disorders. PLoS One *7*, 1–10.
796 Li, W., Prazak, L., Chatterjee, N., Gruninger, S., Krug, L., Theodorou, D., and Dubnau, J.
797 (2013). Activation of transposable elements during aging and neuronal decline in Drosophila.
798 Nat. Neurosci. *16*, 529–531.
799 MacKay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S.,
800 Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic
801 Reference Panel. Nature *482*, 173–178.
802 Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I.,
803 Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide
804 expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.
805 Magoč, T., and Salzberg, S.L. (2011). FLASH: Fast length adjustment of short reads to
806 improve genome assemblies. Bioinformatics *27*, 2957–2963.
807 Makałowski, W., Mitchell, G.A., and Labuda, D. (1994). Alu sequences in the coding regions
808 of mRNA: a source of protein variability. Trends Genet. *10*, 188–193.
809 Muotri, A.R., Chu, V.T., Marchetto, M.C.N., Deng, W., Moran, J. V., and Gage, F.H. (2005).
810 Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature
811 *435*, 903–910.
812 Nekrutenko, A., and Li, W.H. (2001). Transposable elements are found in a large number of

human protein-coding genes. Trends Genet. *17*, 619–621.

Nelson, M.G., Linheiro, R.S., and Bergman, C.M. (2017). McClintock: An integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. G3 Genes, Genomes, Genet. *7*, 2763–2778.

Pastuzyn, E.D., Day, C.E., Kearns, R.B., Kyrke-Smith, M., Taibi, A. V., McCormick, J., Yoder, N., Belnap, D.M., Erlendsson, S., Morado, D.R., et al. (2018). The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell *172*, 275-288.e18.

Perrat, P.N., DasGupta, S., Wang, J., Theurkauf, W., Weng, Z., Rosbash, M., and Waddell, S. (2013). Transposition-Driven Genomic Heterogeneity in the Drosophila Brain. Science (80-. ). *340*, 91–95.

Philippe, C., Vargas-Landin, D.B., Doucet, A.J., Van Essen, D., Vera-Otarola, J., Kuciak, M., Corbin, A., Nigumann, P., and Cristofari, G. (2016). Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. Elife *5*, 1–30.

Pinson, M.E., Pogorelcnik, R., Court, F., Arnaud, P., and Vaurs-Barrière, C. (2018). CLIFinder: Identification of LINE-1 chimeric transcripts in RNA-seq data. Bioinformatics *34*, 688–690.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rangwala, S.H., Zhang, L., and Kazazian, H.H. (2009). Many LINE1 elements contribute to the transcriptome of human somatic cells. Genome Biol. *10*, 1–18.

Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics, Proteomics Bioinforma. *13*, 278–289.

Roebroek, A.J.M., Pauli, I.G.L., Zhang, Y., and van de Ven, W.J.M. (1991). cDNA sequence of a Drosophila melanogaster gene, Dfur1, encoding a protein structurally related to the subtilisin-like proprotein processing enzyme furin. FEBS Lett. *289*, 133–137.

Sanchez-Luque, F.J., Kempen, M.J.H.C., Gerdes, P., Vargas-Landin, D.B., Richardson, S.R., Troskie, R.L., Jesuadian, J.S., Cheetham, S.W., Carreira, P.E., Salvador-Palomeque,

C., et al. (2019). LINE-1 Evasion of Epigenetic Repression in Humans. Mol. Cell *75*, 590-604.e12.

Schauer, S.N., Carreira, P.E., Shukla, R., Gerhardt, D.J., Gerdes, P., Sanchez-Luque, F.J., Nicoli, P., Kindlova, M., Ghisletti, S., Dos Santos, A.D., et al. (2018). L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. Genome Res. *28*, 639–653.

Sienski, G., Dönertas, D., and Brennecke, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell *151*, 964–980.

Smit, A.F.A., Hubley, R., and Green, P. (2015). RepeatMasker. RepeatMasker Open-4.0.

Smith, L.A., Wang, X.J., Peixoto, A.A., Neumann, E.K., Hall, L.M., and Hall, J.C. (1996). A Drosophila calcium channel α1 subunit gene maps to a genetic locus associated with behavioral and visual defects. J. Neurosci. *16*, 7868–7879.

Stephens, R.M., and Schneider, T.D. (1992). Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J. Mol. Biol. *228*, 1124–1136.

Sun, W., Samimi, H., Gamez, M., Zare, H., and Frost, B. (2018). Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. Nat. Neurosci. *21*, 1038–1048.

Thibault, S.T., Singer, M.A., Miyazaki, W.Y., Milash, B., Dompe, N.A., Singh, C.M., Buchholz, R., Demsky, M., Fawcett, R., Francis-Lang, H.L., et al. (2004). A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac. Nat. Genet. *36*, 283–287.

Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B., Millburn, G., Antonazzo, G., Trovisco, V., et al. (2019). FlyBase 2.0: The next generation. Nucleic Acids Res. *47*, D759–D765.

Treiber, C.D., and Waddell, S. (2017). Resolving the prevalence of somatic transposition in Drosophila. Elife *6*, 1–22.

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sánchez-Luque, F.J.,

Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., Van Der Knaap, M.S., Brennan, P.M., et al. (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell *161*, 228–239.

Wan, H.I., DiAntonio, A., Fetter, R.D., Bergstrom, K., Strauss, R., and Goodman, C.S. (2000). Highwire regulates synaptic growth in Drosophila. Neuron *26*, 313–329.

Wang, T., Santos, J.H., Feng, J., Fargo, D.C., Shen, L., Riadi, G., Keeley, E., Rosh, Z.S., Nestler, E.J., and Woychik, R.P. (2016). A novel analytical strategy to identify fusion transcripts between repetitive elements and protein coding-exons using RNA-Seq. PLoS One *11*, 1–20.

Wang, Z., Berkey, C.D., and Watnick, P.I. (2012). The Drosophila Protein Mustard Tailors the Innate Immune Response Activated by the Immune Deficiency Pathway. J. Immunol. *188*, 3993–4000.

Zhang, W., Wu, J., Ward, M.D., Yang, S., Chuang, Y.A., Xiao, M., Li, R., Leahy, D.J., and Worley, P.F. (2015). Structural basis of arc binding to synaptic proteins: Implications for cognitive disease. Neuron *86*, 490–500.