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.treiber@cncb.ox.ac.uk, 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 transposition in neural tissue could contribute to neuropathology and individuality, but its prevalence is debated. We used single-cell mRNA sequencing to map transposon expression in the *Drosophila* midbrain. We found that neural transposon expression is driven by cellular genes. Every expressed transposon is resident in at least one cellular gene with a matching expression pattern. A new long-read RNA sequencing approach revealed that coexpression is a physical link in the form of abundant chimeric transposon-gene mRNAs. We identified 148 genes where transposons introduce cryptic splice sites into the nascent transcript and thereby produce many additional mRNAs. Some genes exclusively produce chimeric mRNAs with transposon sequence and on average transposon-gene chimeras account for 20% of the mRNAs produced from a given gene. Transposons therefore significantly expand the neural transcriptome. We propose that chimeric mRNAs produced by splicing into polymorphic transposons may contribute to functional differences between individual cells and animals.

Introduction

Transposons comprise almost half of every eukaryote genome (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. Non-heritable *de novo* transposon activity in neural tissue has been proposed to 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 faithfully 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). A growing number of studies have therefore correlated the development of 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 expression as a proxy for mobility could be misleading because high-level transposon expression does not appear to result in elevated *de-novo* somatic transposition in the brain (Evrony et al., 2012, 2016; Treiber and Waddell, 2017). It is therefore important to understand what controls the expression of transposon-derived sequences in the brain and whether their elevated expression relates to neural function.

An early study of the human LINE-1 (L1) promoter demonstrated that its activity was heavily influenced by flanking cellular sequences, and concluded that expression of a given L1 depended on its location in the genome (Lavie et al., 2004). Such a locus specific model could more generally explain the apparent cell-type restricted nature of transposon expression and mobilization in the brain. So far, studies of somatic transposon expression have either focused on single transposon families or have been based on bulk sequencing of tissues or cultured cells (Chung et al., 2019; Faulkner et al., 2009; Li et al., 2013; Philippe et al., 2016; Rangwala et al., 2009). However, answering this question on a brain- and genome-wide scale requires a means to relate the cellular expression of each transposon in the genome to that of their neighboring genes. Recent technical developments in whole-

genome DNA sequencing and high-throughput single-cell transcriptomics of complex tissues now make this possible (Macosko et al., 2015).

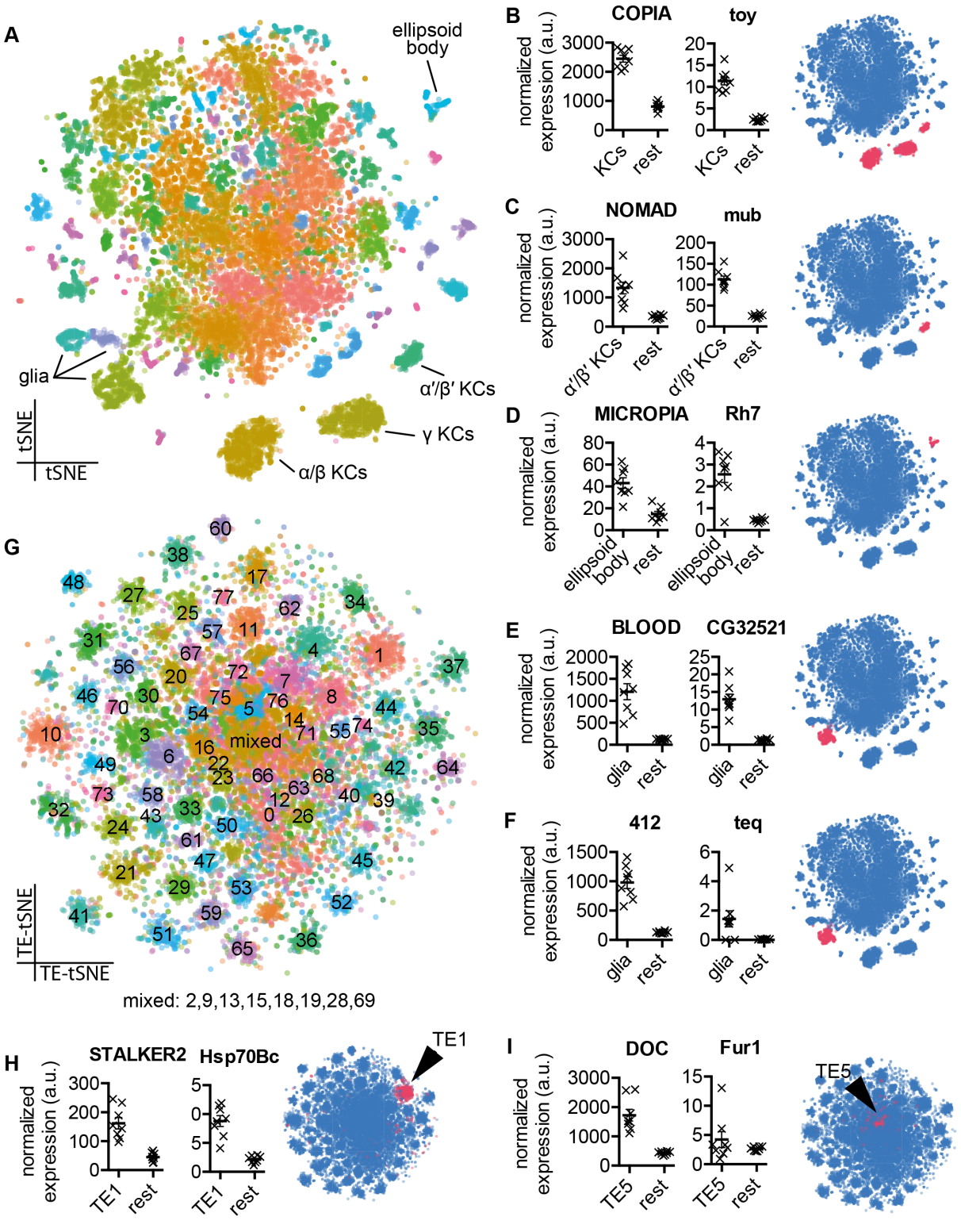
Here we used single-cell transcriptomics to map transposon expression to individual cells in the *Drosophila* midbrain. We found that many transposons are expressed with cell-specificity that is often highly correlated to that of a neighboring gene. A more detailed analysis revealed that >90% of transposon expression can be linked to co-expression with a host gene, indicating that these genes are the main driver of somatic transposon expression. Long-read sequencing showed that the transposon and neighboring gene are alternatively spliced together becoming part of the same chimeric mRNAs. Sometimes all mRNAs produced from a particular gene in which a transposon resides include transposon sequence. Therefore, transposons produce genome-wide diversification of cellular transcripts. Analysis of sequencing data produced from other fly strains demonstrated large differences in their chimeric transcriptomes. Inter-strain and individual differences in transposon complement therefore constrain the cellular specificity and likelihood of transposon-directed pathology.

Results

Single-cell transcriptomics reveals cell-type restricted transposon expression

The *Drosophila* genome contains 112 families of transposons and the number of an individual type varies from a few to hundreds of copies (Kaminker et al., 2002). Conventional single-cell RNAseq (scRNAseq) analysis pipelines typically discard sequencing reads that align to multiple genomic loci, and so they overlook transposon expression. We therefore devised an analysis pipeline to map the expression of all transposons within scRNAseq data. We masked all repetitive sequences in the reference genome and then added a single copy of the consensus sequence for every known transposon to the masked genome. In essence this produces a *Drosophila* reference genome with one copy of each type of transposon. We first used this modified reference genome to map transposon expression onto single cells of the midbrain prepared 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 both the sense and the antisense strand of most transposons, which comprised 76.2 and 23.8% (+/- 1.7% SD) of all transposon expression, respectively (Supplemental figure 1). We first performed principal component decomposition of cellular genes and clustered cells from the midbrain by constructing a k-Nearest-Neighbor graph on the Euclidean distances in the PCA space, optimizing the modularity using the Louvain algorithm (Butler et al., 2018). This analysis grouped cells into many discrete clusters. We next assigned many of these clusters to cell types in the midbrain using the expression patterns of known marker genes (Croset et al., 2018) (Figure 1a). Displaying the expression of individual types of transposons on the cluster plot revealed that some transposons are up-regulated in specific cell types. For example, the long-terminal repeat (LTR) retrotransposons COPIA and NOMAD showed elevated expression in the αβ, α´β´ and γ Kenyon Cells (KCs) classes (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


A
ellipsoid body
glia
α′/β′ KCs
γ KCs
α/β KCs
tSNE
tSNE
B
COPIA
toy
normalized expression (a.u.)
KCs
rest
C
NOMAD
mub
α′/β′ KCs
D
MICROPIA
Rh7
ellipsoid body
E
BLOOD
CG32521
glia
F
412
teq
G
mixed
TE-tSNE
mixed: 2,9,13,15,18,19,28,69
H
STALKER2
Hsp70Bc
TE1
I
DOC
Fur1
TE5

101

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

A Two-dimensional reduction (tSNE) 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 tSNE representation of cell-type restricted expression. **B** COPIA and *twin-of-eyeless* (*toy*) in all Kenyon Cell (KC) classes. **C** NOMAD and *mushroom-body expressed* (*mub*) in $\alpha'\beta'$ 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 that transposon- and gene levels were normalized separately. Blue schematic shows location of cell cluster (pink) in tSNE 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 *Hsp70Bc* and DOC and *Furin 1* (*Fur1*) in their relevant transposon clusters and the position of the cluster in the overall transposon-based tSNE (indicated in pink).

Transposon expression correlates with that of cellular genes they are inserted within. Transposons could be elevated in specific cell types because they are inserted in genes that are highly expressed in the same cells. To test this hypothesis, we re-used our previously published high-coverage gDNA sequence of αβCherry flies. We mapped all the germline transposon insertions in these flies using TEchim, a new custom-built transposon analysis program. TEchim first generates long nucleotide contigs from either gDNA or cDNA sequencing reads, then creates in-silico paired-end reads and screens them for cases where one in-silico end maps to a cellular gene and the mate read maps to a specific transposon. Since these paired-end reads are derived from contiguous sequences, TEchim permits one to subsequently refer back to the original long reads to determine the precise nucleotide sequence of the transposon-gene breakpoints. Using TEchim, we found copies of COPIA, NOMAD, MICROPIA, BLOOD and 412 inside the genes *twin-of-eyeless* (*toy*), *mushroom-body expressed* (*mub*), *Rhodopsin-7* (*Rh7*), *CG32521* and *tequila* (*teq*), respectively. The expression of each of these genes mirrored the expression pattern of the transposon they harbored (Figure 1b-f, second graphs). The expression of these transposons in the brain of αβCherry flies therefore appears to be driven by their relevant host genes.

We next tested whether all transposons exhibit patterned expression throughout the midbrain. We re-clustered the single-cell data of the fly using only transposon expression. This generated 78 cell clusters that mostly contained cells from all 8 biological replicates in the data (Figure 1g, Supplemental figure 2). This result suggests that transposon expression is stereotyped across samples derived from different flies collected from the same strain. We then analyzed the expression of cellular genes across the transposon clusters and found many clusters also preferentially expressed certain genes. For example, the cluster of cells expressing the LTR of STALKER2 was enriched for cells that also expressed the *Hsp70Bc* gene (Figure 1h), and cells in the DOC-positive cluster showed increased expression of *Furin1* (*Fur1*). By referring back to the gDNA, we found that αβCherry flies harbor a copy of STALKER2 within *Hsp70Bc* and a copy of the LINE-like DOC element inside *Fur1*. Again,

these data suggest that expression of STALKER2 and DOC is driven by a neighbouring gene.

Quantitative analysis reveals high fidelity transposon-gene co-expression

Our gDNA analysis also revealed many transposon insertions inside genes that were more broadly expressed across the brain. In total, we identified 1952 germline transposon insertions within genes. Of these, 881 cases were inserted in the sense direction to the open reading frame of the gene and 1071 were in the antisense orientation (Supplemental Table 1). To quantify the correlated expression of transposons and cellular genes we devised a method based on the established Hardy-Weinberg principle for quantifying linkage equilibrium of two alleles in population genetics (Lewontin and Kojima, 1960) (Figure 2a). We first binarized our scRNAseq data to generate the equivalent of bi-allelic traits in a population. We then calculated the proportion of cells expressing a specific transposon, multiplied it by the proportion of cells expressing a certain gene, and then subtracted this value from the proportion of cells that expressed both the transposon and the gene. We termed this value the Coexpression Disequilibrium, CD. We also normalized these CD values to account for the variable abundance of each transposon and gene in every transposon-gene pair and repeated the analysis for all transposon-transposon and gene-gene pairs. These normalized values were then ranked within each of the 8 biological replicates. P-values were corrected for multiple comparisons and describe the probability that a transposon-gene pair would have such a highly ranked CD value across multiple replicates if they were expressed independently.

Figure 2

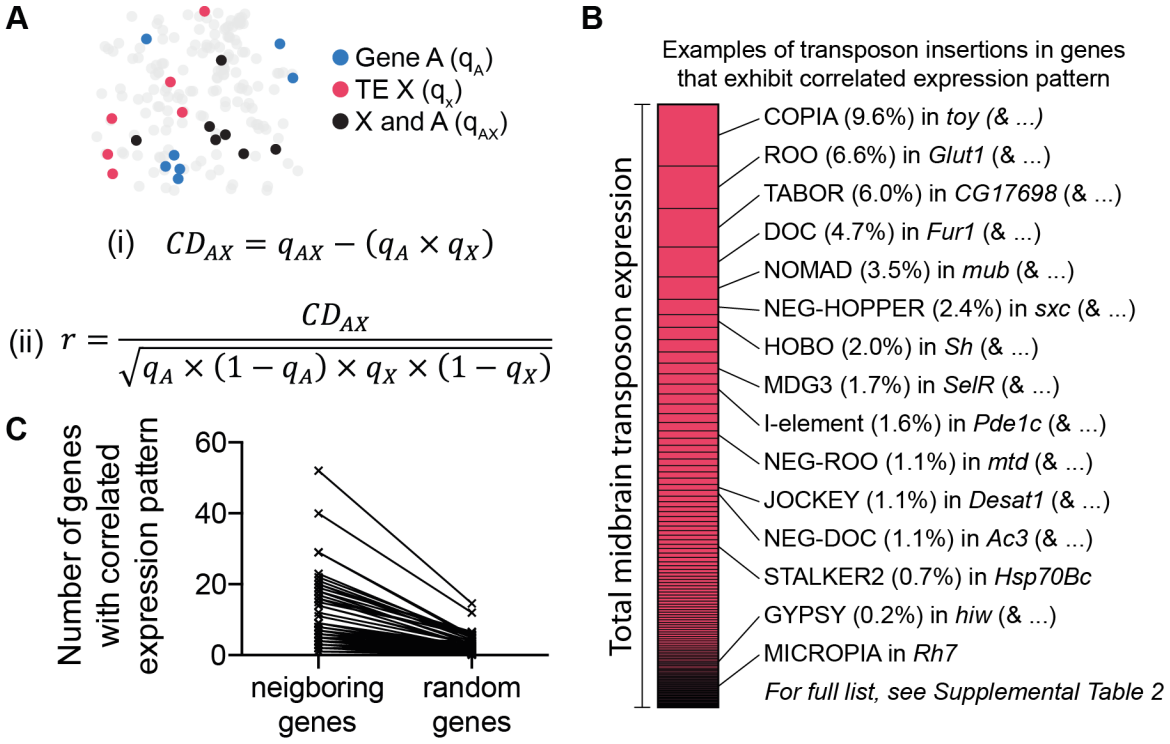


Figure 2. Transposons are co-expressed with neighboring genes.

A Schematic and formulae describing the calculation of Co-expression Disequilibrium values. **B** Examples of transposon-gene pairs that are neighboring in the genome and co-expressed across the midbrain. Pink bar represents the total transposon expression. Examples are selected from the most highly expressing transposons. See Supplemental Table 2 for entire list of correlated transposon-gene pairs. **C** Graph illustrating that more neighboring genes are correlated with their resident transposons than if transposons are randomly assigned to genes.

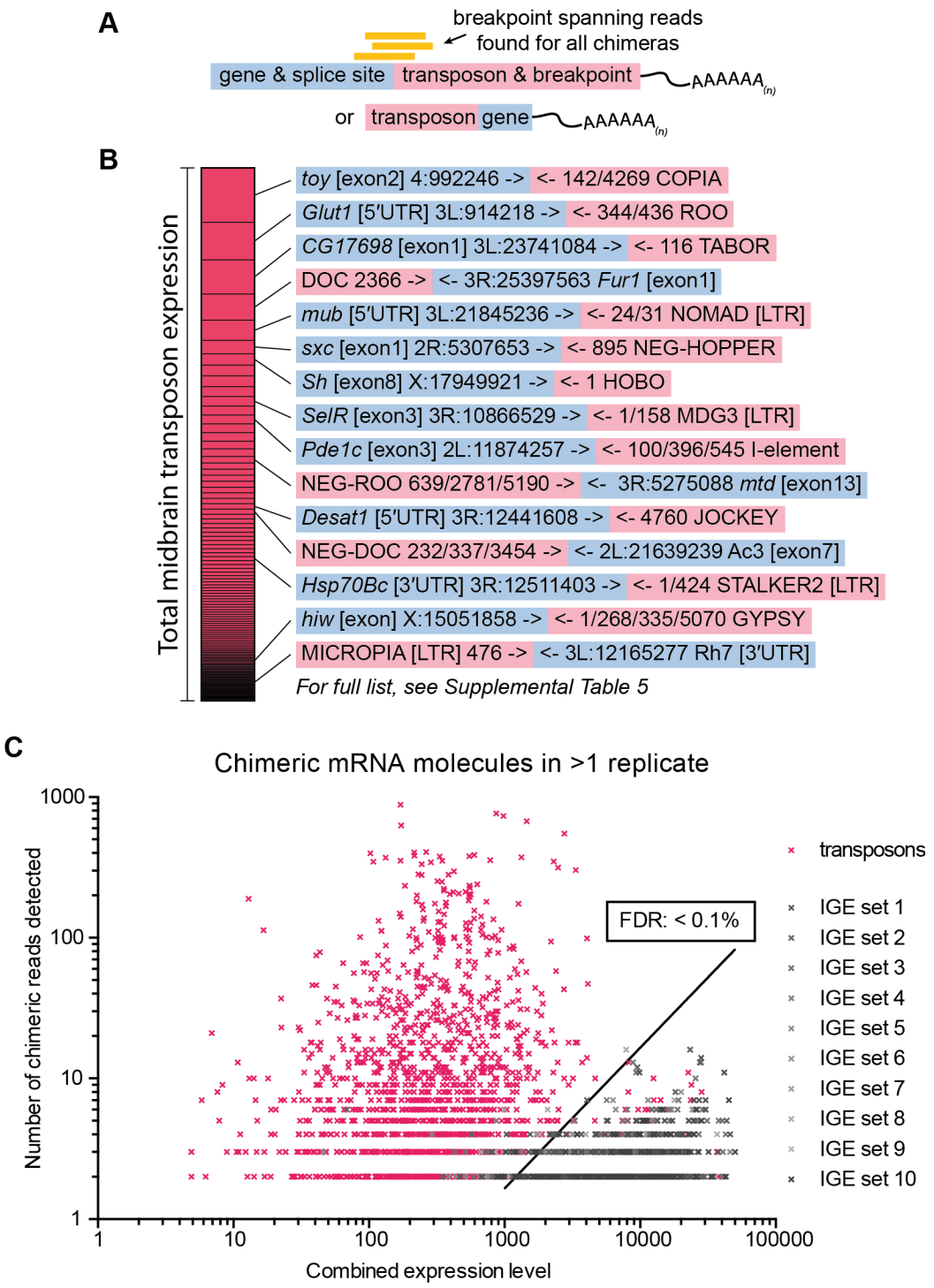
179 We combined the list of all germline transposon insertions in αβCherry flies with the
180 scRNAseq data generated from the same population of flies and calculated the CD values
181 between every transposon and the gene in which it was inserted. We tested every
182 transposon that contributed at least 0.1% of the overall transposon expression in our
183 midbrain αβCherry fly samples. This cut-off left 59 different transposons, 34 of which were
184 expressed in both the sense- and the antisense direction, 22 only in sense, and 3 only in
185 antisense (Figure 2b, Supplemental Table 2). For 56 of these transposons we found at least
186 one copy inside a gene that exhibited a correlated expression pattern (Benjamini-Hochberg
187 corrected p-val >0.05). For those cases where the transposon was inserted in the same
188 orientation as the transcription unit of the gene the expression of the sense strand of the
189 transposon correlated to that of the gene. In contrast, the antisense strand of reverse
190 orientation transposons was correlated with the host gene. Importantly, the average number
191 of correlated genes for each transposon was significantly less if equivalent values were
192 calculated using random assignment of transposons to host genes (Figure 2c). These
193 analyses therefore demonstrate that the genomic locus strongly influences the expression
194 patterns of almost all transposons in the fly brain. We did not identify a neighboring gene
195 with a correlated expression pattern for the transposons TART-A, P-element and
196 HMSBEAGLE. This is expected for TART-A, which is a telomeric retrotransposon, and for P-
197 element, which is a remnant of transgenic intervention in αβCherry flies. It is conceivable
198 that the HMSBEAGLE retrotransposon is the only element that is expressed independently
199 of a cellular gene. However, despite our high coverage sequence of αβCherry flies we may
200 have missed a germline HMSBEAGLE insertion that sits near a gene that is driving its
201 expression.
202

203 **Transposons are exonized into cellular mRNAs**

204 Recent work has shown that mRNAs from the *Arc* gene contain transposon-like sequence in
205 the coding sequence and 3′ UTR (Pastuzyn et al., 2018; Zhang et al., 2015). We therefore
206 tested whether chimeric mRNAs might occur more broadly and extend to all transposons.

207 We extracted mRNA from αβCherry fly heads and generated 250 basepair long reads which
208 were screened using TEchim for chimeric reads. We also incorporated a function in TEchim
209 that maintains strand-specificity of the input reads which enabled us to unambiguously
210 assign chimera to cellular genes. This analysis revealed that a large number of transposons
211 inside introns lead to the formation of chimeric mRNAs. In total, we found chimeric mRNA
212 from 887 transposon insertions (Figure 3a, Supplemental Table 3). Chimera included
213 sequences from LTR, LINE-like and DNA transposons attached to mRNAs from genes
214 involved in a broad range of biological processes. For example, we found sequence from the
215 LTR-retrotransposon GYPSY in transcripts of the ubiquitin gene *Ubi-P5E* and of the neuron
216 specific ubiquitin ligase *highwire* (*hiw*), the non-LTR element DOC in *Fur1*, encoding a
217 synaptic membrane bound protease, and the TIR element HOBO attached to transcripts
218 from *Shaker*, which encodes a voltage-gated potassium channel (Izquierdo, 1994; Kaplan
219 and Trout, 1969; Roebroek et al., 1991; Wan et al., 2000).

Figure 3


A
breakpoint spanning reads
found for all chimeras
gene & splice site
transposon & breakpoint
AAAAAA(n)
or
transposon
gene
AAAAAA(n)
B
Total midbrain transposon expression
toy [exon2] 4:992246 -> <- 142/4269 COPIA
Glut1 [5'UTR] 3L:914218 -> <- 344/436 ROO
CG17698 [exon1] 3L:23741084 -> <- 116 TABOR
DOC 2366 -> <- 3R:25397563 Fur1 [exon1]
mub [5'UTR] 3L:21845236 -> <- 24/31 NOMAD [LTR]
sxc [exon1] 2R:5307653 -> <- 895 NEG-HOPPER
Sh [exon8] X:17949921 -> <- 1 HOBO
SelR [exon3] 3R:10866529 -> <- 1/158 MDG3 [LTR]
Pde1c [exon3] 2L:11874257 -> <- 100/396/545 I-element
NEG-ROO 639/2781/5190 -> <- 3R:5275088 mtd [exon13]
Desat1 [5'UTR] 3R:12441608 -> <- 4760 JOCKEY
NEG-DOC 232/337/3454 -> <- 2L:21639239 Ac3 [exon7]
Hsp70Bc [3'UTR] 3R:12511403 -> <- 1/424 STALKER2 [LTR]
hiw [exon] X:15051858 -> <- 1/268/335/5070 GYPSY
MICROPIA [LTR] 476 -> <- 3L:12165277 Rh7 [3'UTR]
For full list, see Supplemental Table 5
C
Chimeric mRNA molecules in >1 replicate
Number of chimeric reads detected
1000
100
10
1
1
10
100
1000
10000
100000
Combined expression level
FDR: < 0.1%
transposons
IGE set 1
IGE set 2
IGE set 3
IGE set 4
IGE set 5
IGE set 6
IGE set 7
IGE set 8
IGE set 9
IGE set 10

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

222 **A** Schematic of the structure of chimeric mRNA molecules, and illustrating the data
223 representation in **B**. **B** Examples of transposon-gene pairs that form chimeric mRNA. The
224 blue box contains the gene name, the section of the gene that forms a chimera and the
225 breakpoint, which is always the endogenous exon-intron junction. The red box contains the
226 transposon name, and the breakpoint(s) on the transposon. Note that for consistency,
227 transposon breakpoints are always taken from the sense orientation, starting from the core
228 transposon sequence, unless specifically indicated as [LTR]. Examples shown here are the
229 same transposon-gene pairs as in Figure 2b. For the entire list of chimera, see
230 Supplemental Table 5. **C** Graph showing the number of chimeric reads, and the combined
231 expression levels of each transposon-gene pair (pink), as well as for all 10 sets of IGE-gene
232 pairs (grey). Combined expression levels are the square root of the product of reads in both
233 transcripts of a transposon/IGE-gene pair. IGEs were used to calculate a threshold for a
234 False Discovery Rate that is less than 0.1%.

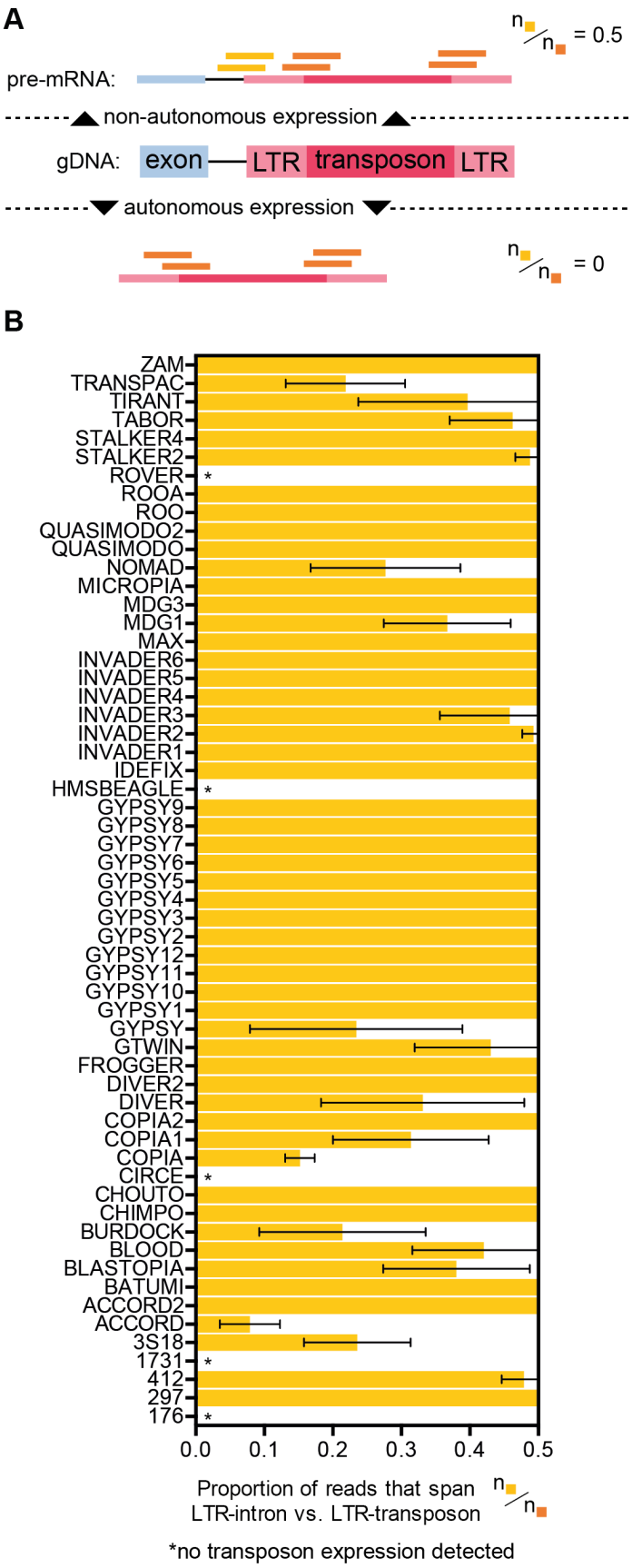
235 Previous studies of chimeric sequencing reads have established that *in vitro* amplification of
236 genetic material often leads to chimeric amplification artefacts (Evrony et al., 2016; Treiber
237 and Waddell, 2017). It was thus important to account for similar errors in our data. We
238 therefore calculated the rate of these artefacts in our mRNA data by selecting 10 sets of 167
239 exons (with each set providing a number of sequences corresponding to the 112 different
240 transposon types and 55 LTRs) with matching expression levels in the brain. These exons
241 lack the ability to relocate in gDNA so we refer to them as immobile genetic elements (IGEs).
242 Since IGEs should only occur as single copies in the gDNA from αβCherry, chimeric reads
243 between IGEs and other genes most likely represent amplification artefacts. We found the
244 rate of generating IGE chimeras was directly correlated to the expression level of the IGE
245 and the gene that it formed a chimeric molecule with. Critically, the IGE chimera rate was
246 substantially lower than that of chimera formed between genes and transposons. We
247 therefore used the rate of IGE chimera to filter the transposon chimera detected using
248 TEchim, defining a false discovery rate (FDR) of 0.1% (Figure 3b, Supplemental Table 3). All
249 examples of chimeric transcripts that are presented in detail in this study have supporting
250 evidence that exceeds this FDR.
251

252 **LTR retrotransposon expression is predominantly non-autonomously**

253 Given that transposon expression was highly correlated with at least one neighboring gene,
254 we hypothesized that all transposon expression in the brain might occur as co-expression
255 with cellular genes. To test this, we focused on LTR retrotransposons. We quantified the
256 number of reads that spanned an LTR-gene breakpoint, and compared it to the number of
257 reads that crossed the LTR-transposon breakpoint within the transposon (Figure 4a).
258 Autonomously expressed, full-length transposons only generate the latter type of read, whilst
259 non-autonomous expression should generate both kinds of reads, at varying proportions.
260 We found that around 95% of the highly expressed LTR transposons produced a roughly
261 equivalent number of gene-transposon and LTR-transposon reads, suggesting that LTR

262 transposons are expressed as chimeras with cellular genes, rather than being autonomously
263 expressed (Figure 4b, Supplemental Table 4).

Figure 4


A
n / n = 0.5
pre-mRNA:
non-autonomous expression
gDNA:
exon
LTR
transposon
LTR
autonomous expression
n / n = 0
B
ZAM
TRANSPAC
TIRANT
TABOR
STALKER4
STALKER2
ROVER
ROOA
ROO
QUASIMODO2
QUASIMODO
NOMAD
MICROPIA
MDG3
MDG1
MAX
INVADER6
INVADER5
INVADER4
INVADER3
INVADER2
INVADER1
IDEFIX
HMSBEAGLE
GYPSY9
GYPSY8
GYPSY7
GYPSY6
GYPSY5
GYPSY4
GYPSY3
GYPSY2
GYPSY12
GYPSY11
GYPSY10
GYPSY1
GYPSY
GTWIN
FROGGER
DIVER2
DIVER
COPIA2
COPIA1
COPIA
CIRCE
CHOUTO
CHIMPO
BURDOCK
BLOOD
BLASTOPIA
BATUMI
ACCORD2
ACCORD
3S18
1731
412
297
176
0.0
0.1
0.2
0.3
0.4
0.5
Proportion of reads that span LTR-intron vs. LTR-transposon
n / n
*no transposon expression detected

264

265 **Figure 4. LTR retrotransposon expression is predominantly non-autonomous**

266 **A** Illustration showing method of calculating the percentage of chimeric transcripts vs.
267 autonomously expressed transposon transcripts. Non-autonomous expression should result
268 in an approximate value of 0.5 for reads spanning the LTR section of transposons and the
269 intron of a neighboring gene over the number of reads spanning the LTR and the core
270 section of the transposon. In contrast, autonomous expression would not result in LTR-gene
271 spanning reads. **B** List of all LTR transposons analyzed in our mRNA data. We identified
272 LTR-gene spanning reads for every LTR transposon that is expressed in the midbrain. Error
273 bars represent standard deviation. Values have been capped at 0.5 in this graph. However,
274 some transposons produced a much higher number of LTR-gene reads (see Supplemental
275 Table 4).

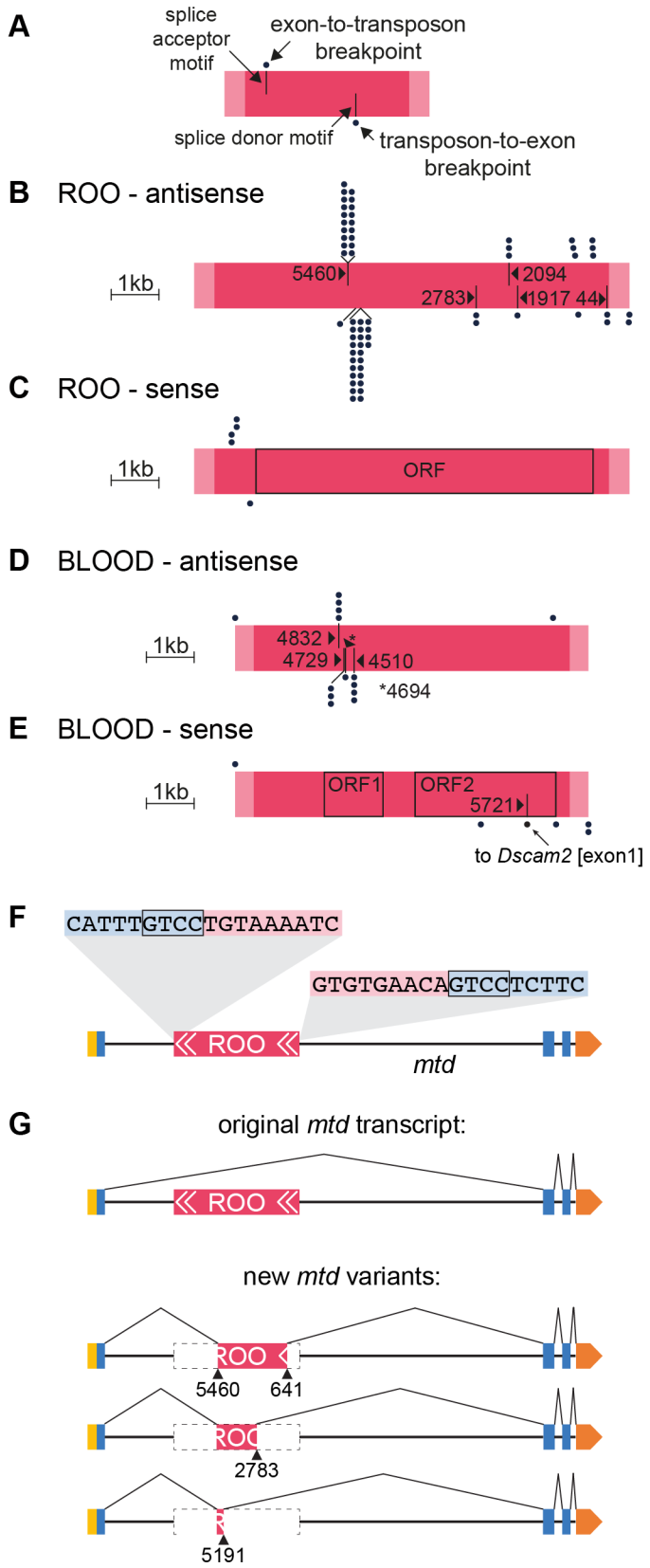
276 **Many transposons introduce cryptic alternative splice-sites into cellular genes**

277 Given the abundance of reads that span a gene-transposon breakpoint, we next investigated
278 the structure of these transcripts in more detail. Our transposon mapping identified 887 loci
279 with a germline transposon insertion in the intron of a gene. For each of these we found
280 mRNA molecules where one section mapped to the beginning or end of the transposon and
281 the other section corresponded to the flanking intronic sequence. These reads could
282 represent nascent unspliced chimeric pre-mRNAs from which transposon-derived sequence
283 could be removed by splicing to yield intact host mRNAs, and full-length transposon
284 sequences. However, we also found 148 examples where breakpoint-spanning reads
285 indicate that specific sections of transposon sequence are spliced into host-gene transcripts
286 (Supplemental Table 5).

287

288 Analysis of the breakpoints inside transposons at these 148 sites revealed that chimera are
289 formed at conserved locations in each type of transposon. For example, in cases where an
290 antisense ROO resided within an intron, we found transcripts where the 3´-end of an
291 upstream exon had formed a new phosphate bond to a section of ROO at positions 5460
292 and 2094 at 19 and 3 different genomic loci, respectively, and also at several additional
293 breakpoints with lower frequency (Figure 5a,b). In addition, we identified transcripts where
294 sections of ROO were bound to the 5´-end of a downstream exon. We found breakpoints at
295 position 5191 from 24 genes, two at 2783, and several others at unique positions (note that
296 the numbering runs backward because it relates to the forward orientation of ROO).
297 Whereas intronic antisense ROO provides gene-transposon breakpoints for 28 exons, and
298 transposon-gene breakpoints for 33, intronic sense ROO only introduced 4 and 1 (Figure
299 5c). Similarly, the LTR BLOOD also introduced more breakpoints when it was inserted in the
300 antisense orientation relative to the host gene (14 vs. 6, Figure 5d,e).

Figure 5


A
splice acceptor motif
exon-to-transposon breakpoint
splice donor motif
transposon-to-exon breakpoint
B ROO - antisense
1kb
5460
2094
2783
1917 44
C ROO - sense
1kb
ORF
D BLOOD - antisense
1kb
4832
4729
4510
*4694
E BLOOD - sense
1kb
ORF1
ORF2
5721
to Dscam2 [exon1]
F
CATTTGTCCTGTAAAATC
GTGTGAACAGTCCTCTTC
ROO
mtd
G
original mtd transcript:
ROO
new mtd variants:
ROO
5460
641
ROO
2783
5191

301

302 **Figure 5. Transposons introduce splice sites at conserved locations.**

303 **A** Illustration of the labelling scheme in panels **B-E**. The pink bar represents the transposon;
304 light pink ends indicate the LTRs and the dark pink the core sequence. The positions of the
305 dots above the bar represent the site on the transposon where an upstream exon splice
306 donor site has merged. Every dot represents a different gene. Black lines in the top half of
307 the pink bar represent splice acceptor (SA) motifs in the transposon. Dots below the pink bar
308 indicate the location of breakpoints on the transposon that are spliced to upstream exonic
309 SA sites of different genes. Bars in the lower half indicate splice donor (SD) motifs. **B-E**
310 Representations of sense- and antisense ROO and BLOOD (to scale), with all breakpoints
311 to SA and SD sites of neighboring genes. Note that the frequently used site on antisense
312 ROO at position 5191 is a non-consensus SD site, which lacks the expected GT motif at the
313 immediate breakpoint. The sequence around 5191 resembles the consensus SD motif,
314 although the GT is a GC. Compare TTTGGCAAGTT to motif in Supplemental figure 3a. **F**
315 Illustration of antisense ROO insertion in the *mustard* (*mtd*) gene. Only one isoform of *mtd* is
316 shown. Yellow box represents the 5´UTR, blue boxes are exons, orange box the 3´UTR,
317 pink represents ROO transposon with white arrows indicating the LTRs. Breakpoint-
318 spanning gDNA reads reveal Target Site Duplication (TSD, inset). **G** Schematic of original
319 *mtd* transcript, and of three new splice isoforms.

320 We screened the transposon sections around breakpoints for consensus splice- acceptor
321 (SA) and donor (SD) sequence motifs and found many cases where the gene-to-transposon
322 chimera had been formed at SA consensus motifs, and transposon-to-gene chimera at SD
323 motifs (Stephens and Schneider, 1992) (Supplemental figure 3, Supplemental Table 6). For
324 example, all breakpoints in antisense BLOOD that were formed with more than one exon
325 were precisely located at the predicted SA and SD splice site (Figure 5d). Interestingly, we
326 did not find a consensus SD motif at the transposon-gene breakpoint at position 5191 of
327 antisense ROO, although it frequently provided 5´- sequence to transposon-gene chimeric
328 RNAs. However, the sequence around position 5191 matched the consensus motif, with the
329 exception of a GT-to-GC conversion (see Supplemental figure 3). Taken together, our
330 analysis revealed that transposons introduce many alternative splice sites, which are
331 recognized by the host cell spliceosome to combine cellular exonic sequences with sections
332 of transposon sequence.
333
334 We also identified cases of alternative splicing to different sites within the same transposon
335 insertion. Again using ROO as an example, αβCherry flies harbor a reverse orientation ROO
336 in the intron between exons 10 and 11 of the pan-neurally expressed *mustard* (*mtd*) gene,
337 which to date has only been implicated in fly innate immunity (Wang et al., 2012) (Figure 5f).
338 RNAseq revealed a complex collection of *mtd* splice variants that incorporated different
339 fragments of ROO (Figure 5g). SD sites upstream of this ROO came from the end of either
340 *mtd* exon 11 or 13. and these spliced in to the corresponding SA at position 5462 within
341 ROO (Figure 5g). We also found three different SD sites (at positions 641, 2784 and 5191)
342 within ROO, which spliced out to the closest downstream SA (exon 6) of *mtd*. Therefore, this
343 ROO element substantially increases the *mtd* mRNA isoform repertoire. Without ROO the
344 *mtd* locus can express 23 isoforms whereas with ROO it can generate 68 differentially
345 spliced mRNAs.
346

347 We identified 147 other genes whose transcript diversity was similarly increased by
348 transposon insertions. These alternative transcripts incorporate 43 different transposon
349 families which each introduce cryptic SA and/or SD sites into host genes (see Supplemental
350 Table 5). For example, we found a sense insertion of BLOOD inside the *Dscam2* gene which
351 encodes the transmembrane Down Syndrome cell adhesion molecule 2. Chimeric reads
352 indicate that transcription of *Dscam2* is frequently initiated in BLOOD, which is then spliced
353 into exon 33 (the second exon) of the gene. This splicing event combines ORF2 of BLOOD
354 with the remaining exons of *Dscam2* and correctly aligns the reading frames of the two
355 transcripts, generating a novel N-terminus (Supplemental figure 4). We also observed cases
356 where transposon chimera resulted in exon skipping (e.g. the above mentioned ROO in *mtd*
357 and 412 inside *tequila*, Supplemental figure 5). Most transposon chimera resulted from
358 intronic insertions, but we also detected one case where a HOBO insertion resided in the
359 exon of the *CG31705* gene. This HOBO introduced a cryptic SA site which was spliced to
360 the upstream SD from the first exon, creating a truncated *CG31705* transcript (Supplemental
361 figure 6). Together these data show that a broad range of *Drosophila* transposons are
362 alternatively spliced into mRNAs producing many more isoforms of a large number of
363 neurally expressed genes.
364

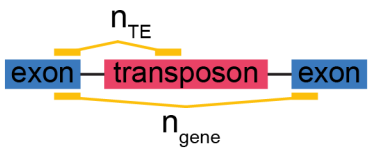
365 **Alternative splicing into and out of transposons can be highly penetrant**

366 Chimeric transcripts could be inconsequential to a cell if they only constitute a small
367 percentage of the overall transcript repertoire of a given gene. We therefore quantified the
368 percentage of mRNAs produced from a transposon harboring gene that include transposon
369 sequence. To do this we analyzed sites where transposons were spliced into an exon-intron
370 junction of a gene. For each gene we counted the number of reads spanning transposon-
371 exon boundaries, and the number of reads that spanned the exon immediately up- and
372 downstream of the transposon insertion (Figure 6a). This analysis showed that for some
373 genes, all derived mRNAs contained transposon sequences. For example, all spliced copies
374 of the isoform B of the Allatostatin A receptor 1 (*AstA-R1*) contained a section of ROO, and

375 every transcript of *Piezo-like* (*Pzl*), a gene encoding a predicted mechanosensitive ion
376 channel, ended in the I-element instead of exon 1 (Hu et al., 2019; Larsen et al., 2001). All
377 mRNAs retrieved from *Mec2* contained TRANSPAC as the most 5´ sequence, suggesting
378 that transcripts might initiate within the TRANSPAC transposon and spliced to the SD of
379 exon 2 of *Mec2*. On average, transposons contributed to 14% of transcripts derived from a
380 specific gene (Figure 6b, Supplemental Table 7). Insertions in genes on the X chromosome,
381 resulted in a higher average percentage of chimeric transcripts (27% vs. 13% for the rest of
382 the genome) and a larger number of cases where 75% or more of transcripts were chimeric.
383 Insertions on the X chromosome are hemizygous in male flies and our samples were
384 generated by pooling an equal number of male and female flies. The increased chimera rate
385 on the X might therefore reflect the smaller number of different X chromosomes when
386 compared to the rest of the genome. For example, we found that the X-linked *cacophony*
387 (*cac*) gene, which encodes a voltage-gated calcium channel, harbored a sense-orientation
388 BLASTOPIA (Smith et al., 1996). This transposon insertion resulted in 54.7% of *cac*
389 transcripts being truncated in αβCherry fly samples, potentially missing the last 8-11 coding
390 exons, suggesting that many flies in this strain are likely mutant for the *cac* gene
391 (Supplemental figure 7). Another interesting example on the X chromosome of αβCherry
392 flies is *Beadex* (*Bx*) which encodes a long-term memory relevant LIM-type transcription
393 factor (Hirano et al., 2016). A sense NOMAD insertion gives rise to at least two new *Bx*
394 transcript isoforms (Supplemental figure 8), which constitute 9.8% of all *Bx* transcripts.

Figure 6

A



B

Gene	Transposon	$\frac{n_{TE}}{n_{TE}+n_{gene}}$
AstA-R1	NEG-ROO	100.0%
Pzl	I	100.0%
nAChRalpha1	NEG-DOC6	100.0%
Mec2	TRANSPAC	100.0%
Thd1	MAX	100.0%
CR45197	NEG-GYPSY	100.0%
CG45263	NEG-POGO	100.0%
Mocs1	P-ELEMENT	100.0%
AstC-R1	NEG-FW	94.1%
teq	412	83.9%
Cyp4p2	COPIA	80.4%

(for full list, see Supplementary Table 7)

Figure 6. High penetrance of transposon-containing splice isoforms.

A Schematic showing method used to calculate the frequency of transposon chimera produced from a given gene. The number of reads that span the exon-transposon junction were divided by the total number of reads that map beyond the exon-intron junction (exon-transposon reads plus reads that partially map to the exon immediately downstream of the transposon insertion. **B** Examples of chimeric transposon-gene transcripts detected in fly heads. Many genes are almost exclusively expressed as chimeric transcripts with transposon sequence. See Supplemental Table 7 for complete list.

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

406 The transposon complement is highly variable between fly strains. We therefore tested
407 whether other fly strains express chimeric gene:transposon mRNAs by analyzing three
408 previously published mRNA sequencing data sets (Croset et al., 2018; Daines et al., 2011;
409 MacKay et al., 2012). Although these prior studies generated shorter paired-end RNAseq
410 reads, than those collected here, we were still able to find chimeric mRNAs in all three data
411 sets (Supplemental Table 8). Some transposon:gene chimera were conserved across all
412 strains, whilst others appeared to be strain-specific. 318 of the 887 chimera identified in our
413 αβCherry flies were present in at least one of the three other data sets, whereas 120 of
414 those occurred in at least three of the four strains. Chimera that were not detected in data
415 from other strains could either reflect genomic heterogeneity between the tested fly strains,
416 or the absence of evidence could result from lower sequencing coverage. Nevertheless,
417 these results demonstrate the prevalence of cellular mRNAs containing transposon
418 sequence.

419 **Discussion**

420 Somatic transposition in neurons has been proposed to contribute to age-dependent
421 neuronal decline in wildtype and disease models of *Drosophila* and (Li, e al., 2013; Guo et
422 al., 2018; Sun et al., 2018). Although the frequency of neural transposition is debated,
423 expression is a prerequisite for movement. Therefore, transposons can only mobilize in cells
424 that express full-length elements, or transposon mRNAs that encode enzymes permitting
425 other elements to move in *trans*. It is therefore important to understand how transposon
426 expression is controlled in the brain. Here, we combined single-cell expression data from the
427 *Drosophila* midbrain with high-coverage gDNA sequence data of the same fly strain and find
428 that the majority, if not all, expressed transposon sequences are parts of chimeric mRNAs
429 with cellular genes.

430

431 Transposons residing in introns have previously been shown to contribute novel exons to
432 several genes (Nekrutenko and Li, 2001). In addition long non-coding (lnc) RNAs frequently
433 contain transposon sequences (Kapusta et al., 2013). We found transposon exonization to
434 be highly prevalent in the *Drosophila* brain. Transposons within the coding regions of genes
435 and those in introns dramatically increase the transcript repertoire by introducing new splice
436 variants. At this stage it is difficult to definitively determine the whole-genome functional
437 consequences of splicing into transposons because we most often only retrieve the
438 sequence across the splice junctions. Furthermore, although each transposon has a known
439 consensus sequence individual copies are highly polymorphic. Nevertheless, our
440 sequencing shows that transposon exonization often truncates and/or changes the amino
441 acid sequence of the encoded gene products, potentially changing protein structure and
442 function. We did also identify several examples where the inclusion of transposon sequence
443 conserved the reading frame of the host gene, likely generating a novel chimeric protein.
444 Amongst the 148 transposon harboring genes identified in this study, there are several that
445 we have described in detail for which the observed locus disruption and altered expression
446 would be expected to have significant consequences for neural function. Flies harboring

447 HOBO in *Sh* and BLASTOPIA in *cac* might exhibit altered voltage-gated currents, whereas
448 those with ROO in *AstA-R1* will respond differently to the modulatory Allatostatin A
449 neuropeptide (Larsen et al., 2001; Smith et al., 1996). We also described insertions of 412 in
450 *teq* and NOMAD in *Bx*, two genes which have been implicated in long-term memory
451 formation (Didelot et al., 2006; Hirano et al., 2016).
452
453 Some transposons provide more splice sites than others. Our analysis of the abundant ROO
454 retrotransposon (found in sense orientation inside 48 genes, and in antisense orientation
455 inside 59 genes in our strain) provides a good example, and the insertion in *mtd* exemplifies
456 the extended variance transposons can introduce to cellular transcription units. Interestingly,
457 antisense insertions of ROO provide more potential SA and SD sites than sense orientation
458 ROO. Insertions of ROO that are sense to the expression of the host gene were found to be
459 mainly part of pre-mRNA, from which functional reverse transcriptase could potentially be
460 translated. In contrast, antisense ROO introduces over 10 times more cryptic splice sites
461 (compare Figure 5b and c).
462
463 mRNA molecules that contain transposon sequences might be susceptible to short-
464 nucleotide mediated gene-silencing (Malone and Hannon, 2009), as recently reported for
465 LINE-2 containing mRNAs and LINE-2 derived microRNAs in humans (Petri et al., 2019).
466 This would provide a challenge for the cells expressing transposon-targeting piRNAs in
467 differentiating between chimeric transcripts and full-length transposon sequences. It is worth
468 noting that transposon-directed piRNAs have been identified in the fly brain (Li et al., 2009).
469 It is therefore conceivable that transposon sequence permits cellular mRNAs to be
470 selectively regulated. In addition, transposon sequence might confer the capacity to be
471 specifically trafficked within the cell, and even between cells (Ashley et al., 2018; Pastuzyn
472 et al., 2018).
473

474 The process of transposable elements acquiring new cellular functions that benefit the host
475 cell is called transposon exaptation (Gould and Vrba, 2013). Several examples of exaptation
476 events during the evolution of eukaryotes have been reported. How these functional
477 transitions occur is, however, not fully understood (Joly-Lopez and Bureau, 2018). Stress-
478 induced transposon mobilization has been observed in many species, and it has been
479 hypothesized that these mobilization events trigger the formation of new transposon-gene
480 chimera. Our results reveal a new mechanism by which transposons participate in the
481 generation of new gene variants. We show that transposons do not need to mobilize in order
482 to form chimeric mRNA molecules. Instead, the cellular gene splicing machinery frequently
483 uses intronic transposon insertions to increase transcriptional diversity.
484
485 Transposon sequences in the gDNA can deliver enhancer elements to cellular genes and
486 thereby influence their expression. However, our results do not support the idea that
487 transposons provide enhancer sequences contributing to neural expression of neighboring
488 genes. If this was the case, we would expect to have found that cellular genes that harbor
489 the same type of transposon are more likely to be expressed in the same cells. Our data
490 instead show that transposon expression is dictated by the genes they are inserted within.
491 This is further supported by the fact that many genes that share transposon insertions with
492 neurally expressed genes were not expressed in the brain.
493
494 Our studies also introduce important practical concerns for the analysis of transposon
495 expression in somatic tissue and disease models. Several previous studies have used
496 Quantitative PCR (qPCR) to measure levels of transposon expression in mutant flies and
497 disease models . Since qPCR probes are usually not designed to cross transposon-gene
498 breakpoints, they cannot distinguish autonomous transposon expression from chimeric
499 mRNA (Guo et al., 2018; Li et al., 2013; Sun et al., 2018). Similarly, standard RNA
500 sequencing protocols rely on the alignment of cDNA fragments to the reference genome and
501 would not identify chimeric reads as non-autonomous transposon expression (e.g. De Cecco

et al., 2013). Baseline and changing cell-specific expression of host genes that form chimeric transcripts with transposons can therefore be misinterpreted as cell-restricted autonomous transposon expression and mobilization.

Our data also constrain somatic transposition. If all transposon expression depends on the neighboring host gene, only cells expressing that host gene can be susceptible to transposition of that element. This would mean that GYPSY could only be active in glia, if the fly strains being studied harbor a copy of GYPSY in a glial-expressed gene (Krug et al., 2017). Interestingly, although we found sense GYPSY sequences in mRNAs for 14 different genes, we did not detect glial GYPSY expression in our αβCherry flies.

Analysis of RNA-seq data from other fly strains suggests that more than half of the chimeric transposon transcripts we identified in αβCherry flies are unique to this strain. This finding alone demonstrates the incredible heterogeneity of the transposon complement between strains. In addition, our prior genome sequencing revealed large differences between individual αβCherry flies (Treiber and Waddell, 2017). Given the broad range of target genes that we have identified to form chimeric mRNA with transposable elements, it will be important to understand the functional consequences of fixed and variable transposons inside genes for the host organism. Nevertheless, it seems highly likely that polymorphism of transposon load and the distribution of transposons across the host genome could contribute towards heterogeneity of neural function, and neurological pathology, between individual animals.

524 **Methods**

525

526 **Fly strains**

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

530

531 **Long-read mRNA sequencing**

532 For RNA extraction, groups of ~50 flies were frozen in liquid nitrogen and vortexed for 6 x 30
533 s to separate heads from abdomens. Heads were isolated using a sieve. To avoid gDNA
534 contamination, mRNA was purified with a combination of protocols. Samples were first
535 processed with a column-based kit (RNeasy Mini kit, Qiagen, UK), including the optional on-
536 column DNAseI digestion. Next, mRNA was extracted from total RNA using oligo-dT
537 magnetic beads (NEB, Ipswich, MA) and the mRNA was purified again using the RNA
538 columns. Finally, sequencing libraries were generated using oligo-dT magnetic beads from a
539 strand-specific mRNA library preparation kit (TruSeq, Illumina, San Diego, CA), with 17
540 cycles of PCR amplification. Fragmentation was optimized to obtain ~350nt long fragments.
541 Whole-genome sequencing was performed on a HiSeq 2500, with 250nt paired-end reads.
542 We mapped the long reads using MRTemp, as previously described (Treiber and Waddell,
543 2017).

544

545 **Single-cell read alignments**

546 The *Drosophila melanogaster* reference genome release 6.25 was used for all sequence
547 alignments. Transposon reference sequences were taken from Repbase (Jurka, 2000;
548 Kaminker et al., 2002). Repetitive sequences in the *Drosophila* reference genome were
549 masked using Repeatmasker (Smit et al.). Single-cell sequencing data was processed with
550 a custom-built data processing pipeline, which is available on GitHub. The masked reference

551 genome, as well as a gene reference file (refFlat) with all genes and each unique reference
552 transposon sequence is provided as supplemental data.
553

554 **Single-cell data analysis**

555 Digital Gene Expression (DGE) matrices were generated as previously described (Butler et
556 al., 2018). Rscripts can be downloaded as supplemental file 1. In summary, DGE's were
557 filtered (≥ 800 UMIs, ≥ 400 features) and 8 replicates were merged. Gene and transposon
558 expressions levels were normalized separately. Marker genes were taken from Croset et al.
559 (2018).
560

561 **Co-expression analysis**

562 Co-expression was quantified by calculating the Co-expression Disequilibrium (CD, see
563 main text). The R-code snippet is available through GitHub. For the statistical analysis, a
564 non-parametric test was performed. CD values between every gene- and transposon
565 combination were ranked within each biological replicate, p-values were calculated using the
566 student t-test and corrected for multiple comparisons using the Benjamini-Hochberg
567 correction. In addition, CD values were calculated for every tested transposon with a set of
568 10 randomly assigned genes (or transposons).
569

570 **Mapping germline transposon insertions in the fly**

571 Germline transposon insertions were mapped with single-nucleotide resolution using
572 previously published gDNA data from the αβCherry fly strain and MRTemp (Treiber and
573 Waddell, 2017). A new, purpose-built, multi-functional sequence analysis pipeline called
574 TEchim was developed. TEchim has 5 key functions: 1. generation of support files, including
575 a masked reference genome and endogenous intron-exon junctions. (input files: reference
576 genome, list of genes, list of transposon sequences). 2. alignment of un-stranded genomic
577 DNA sequence data of multiple sequencing lanes and multiple biological replicates,
578 detection of chimeric sequence fragments with single-nucleotide resolution, 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 4). Detailed descriptions and manuals are available on Github.

Splice acceptor (SA) and donor (SD) motif analysis

SA and SD motifs in the *Drosophila melanogaster* reference genome were generated by randomly selecting 500 known SA and SD sites from exons and screening for motifs in these sequences using the MEME suite (Bailey and Elkan, 1994) (Supplemental figure 3). These motifs were then searched in sequence sections across transposon-gene breakpoints using FIMO (Grant et al., 2011).

Data from previously published studies

Raw single-cell sequencing reads from Croset et al. (2018) was obtained from the NCBI Short Read Archive (SRA https://www.ncbi.nlm.nih.gov/sra) with the accession number: PRJNA428955. 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 1-8. All raw sequencing data generated in this study have been submitted to the NCBI SRA with the BioProject ID PRJNA588978. Custom-built software packages can be accessed via GitHub (https://github.com/charliefornia/TEchim and https://github.com/charliefornia/scHardyWeinberg)

Acknowledgements

We are grateful to other members of the Waddell group for discussion. CT was supported by a Wellcome Trust DPhil studentship. SW is funded by a Wellcome Principal Researcher Fellowship (200846/Z/16/Z), Gatsby Charitable Foundation (GAT3237), ERC Advanced Grant (789274) and Bettencourt–Schueller Foundations.

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.

616 **References**

617

618 Ashley, J., Cordy, B., Lucia, D., Fradkin, L.G., Budnik, V., and Thomson, T. (2018).
619 Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. Cell *172*,
620 262-274.e11.

621 Aso, Y., Hattori, D., Yu, Y., Johnston, R.M., Iyer, N.A., Ngo, T.T.B., Dionne, H., Abbott, L.F.,
622 Axel, R., Tanimoto, H., et al. (2014). The neuronal architecture of the mushroom body
623 provides a logic for associative learning. Elife *3*, e04577.

624 Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to
625 discover motifs in biopolymers. Proceedings. Int. Conf. Intell. Syst. Mol. Biol. *2*, 28–36.

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

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

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

635 De Cecco, M., Criscione, S.W., Peterson, A.L., Neretti, N., Sedivy, J.M., and Kreiling, J.A.
636 (2013). Transposable elements become active and mobile in the genomes of aging
637 mammalian somatic tissues. Aging (Albany. NY). *5*, 867–883.

638 Chung, N., Jonaid, G.M., Quinton, S., Ross, A., Sexton, C.E., Alberto, A., Clymer, C.,
639 Churchill, D., Navarro Leija, O., and Han, M. V. (2019). Transcriptome analyses of tumor-
640 adjacent somatic tissues reveal genes co-expressed with transposable elements. Mob. DNA
641 *10*, 1–22.

642 Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M.,
643 O'Shea, K.S., Moran, J. V., and Gage, F.H. (2009). L1 retrotransposition in human neural

644 progenitor cells. Nature *460*, 1127–1131.
645 Croset, V., Treiber, C.D., and Waddell, S. (2018). Cellular diversity in the Drosophila
646 midbrain revealed by single-cell transcriptomics. eLife 7, e34550.
647 Daines, B., Wang, H., Wang, L., Li, Y., Han, Y., Emmert, D., Gelbart, W., Wang, X., Li, W.,
648 Gibbs, R., et al. (2011). The Drosophila melanogaster transcriptome by paired-end RNA
649 sequencing. Genome Res. *21*, 315–324.
650 Didelot, G., Molinari, F., Tche, P., Comas, D., Milhiet, E., Munnich, A., Colleaux, L., and
651 Preat, T. (2006). Regulates Long-Term Memory Formation in Drosophila. Science *313*, 851–
652 853.
653 Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J.,
654 Atabay, K.D., Gilmore, E.C., Poduri, A., et al. (2012). Single-neuron sequencing analysis of
655 l1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483–496.
656 Evrony, G.D., Lee, E., Park, P.J., and Walsh, C.A. (2016). Resolving rates of mutation in the
657 brain using single-neuron genomics. Elife *5, e12966*
658 Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K.,
659 Cloonan, N., Steptoe, A.L., Lassmann, T., et al. (2009). The regulated retrotransposon
660 transcriptome of mammalian cells. Nat. Genet. *41*, 563–571.
661 Gould, S.J., and Vrba, E.S. (2013). Exaptation-A Missing Term in the Science of Form
662 Exaptation-a missing term in the science of form. Paleobiology *8*, 4–15.
663 Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: Scanning for occurrences of a
664 given motif. Bioinformatics *27*, 1017–1018.
665 Guo, C., Jeong, H.H., Hsieh, Y.C., Klein, H.U., Bennett, D.A., De Jager, P.L., Liu, Z., and
666 Shulman, J.M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. Cell
667 Rep. *23*, 2874–2880.
668 Hirano, Y., Ihara, K., Masuda, T., Yamamoto, T., Iwata, I., Takahashi, A., Awata, H.,
669 Nakamura, N., Takakura, M., Suzuki, Y., et al. (2016). Shifting transcriptional machinery is
670 required for long-term memory maintenance and modification in Drosophila mushroom
671 bodies. Nat. Commun. *7*, 1–14.

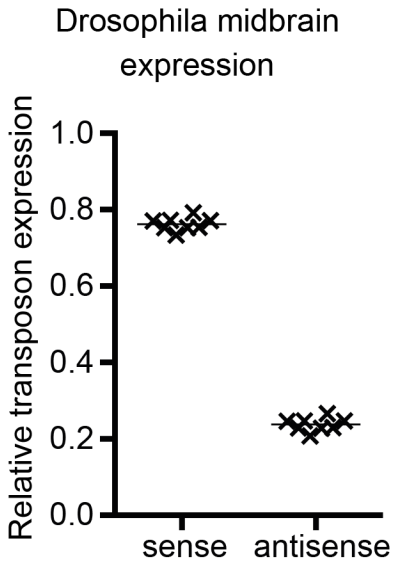
672 Hu, Y., Wang, Z., Liu, T., and Zhang, W. (2019). Piezo-like Gene Regulates Locomotion in
673 Drosophila Larvae. Cell Rep. *26*, 1369-1377.e4.
674 International Human Genome Sequencing Consortium, Eric S. Lander, Lauren M. Linton,
675 Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar,
676 and Michael Doyle (2001). Initial sequencing and analysis of the human genome. Nature
677 *409*, 860–921.
678 Izquierdo, M. (1994). Ubiquitin genes and ubiquitin protein location in polytene
679 chromosomes of Drosophila. Chromosoma *103*, 193–197.
680 Joly-Lopez, Z., and Bureau, T.E. (2018). Exaptation of transposable element coding
681 sequences. Curr. Opin. Genet. Dev. *49*, 34–42.
682 Jurka, J. (2000). Repbase Update: A database and an electronic journal of repetitive
683 elements. Trends Genet. *16*, 418–420.
684 Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E.,
685 Wheeler, D.A., Lewis, S.E., Rubin, G.M., et al. (2002). The transposable elements of the
686 Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol *3(12)*.
687 Kaplan, W.D., and Trout, W.E. (1969). The behavior of four neurological mutants of
688 Drosophila. Genetics *61*, 399–409.
689 Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L.A., Bourque, G., Yandell, M.,
690 and Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin,
691 Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet. *9*.
692 Kazazian, H.H. (2011). Mobile DNA transposition in somatic cells. BMC Biol. *9*, 2–5.
693 Kazazian, H.H., and Moran, J. V. (2017). Mobile DNA in health and disease. N. Engl. J.
694 Med. *377*, 361–370.
695 Ketchum, K., Ketchum, K., Hoskins, R., Hoskins, R., Wang, X., Wang, X., Smith, T., Smith,
696 T., Gocayne, J., Gocayne, J., et al. (2000). The genome sequence of Drosophila
697 melanogaster. Science *287*, 2185–2195.
698 Krug, L., Chatterjee, N., Borges-Monroy, R., Hearn, S., Liao, W.W., Morrill, K., Prazak, L.,
699 Rozhkov, N., Theodorou, D., Hammell, M., et al. (2017). Retrotransposon activation

700 contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. PLoS Genet.
701 13(3):e1006635
702 Larsen, M.J., Burton, K.J., Zantello, M.R., Smith, V.G., Lowery, D.L., and Kubiak, T.M.
703 (2001). Type A allatostatins from Drosophila melanogaster and Diplotera puncata activate
704 two Drosophila allatostatin receptors, DAR-1 and DAR-2, expressed in CHO cells. Biochem.
705 Biophys. Res. Commun. *286*, 895–901.
706 Lavie, L., Maldener, E., Brouha, B., Meese, E.U., and Mayer, J. (2004). The human L1
707 promoter: Variable transcription initiation sites and a major impact of upstream flanking
708 sequence on promoter activity. Genome Res. *14*, 2253–2260.
709 Lewontin, R.C., and Kojima, K. (1960). The Evolutionary Dynamics of Complex
710 Polymorphisms. Evolution (N. Y). *14*, 458–472.
711 Li, W., Prazak, L., Chatterjee N., Grüninger, S., Krug, L., Theodorou, D., Dubnau, J. (2013).
712 Activation of transposable elements during aging and neuronal decline in Drosophila.
713 Neurosci. Nat. *16*, 529–531.
714 Li, C., Vagin, V. V., Lee, S., Xu, J., Ma, S., Xi, H., Seitz, H., Horwich, M.D., Syrzycka, M.,
715 Honda, B.M., et al. (2009). Collapse of Germline piRNAs in the Absence of Argonaute3
716 Reveals Somatic piRNAs in Flies. Cell *137*, 509–521.
717 Li, W., Jin, Y., Prazak, L., Hammell, M., and Dubnau, J. (2012). Transposable Elements in
718 TDP-43-Mediated Neurodegenerative Disorders. PLoS One *7*, 1–10.
719 MacKay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S.,
720 Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic
721 Reference Panel. Nature *482*, 173–178.
722 Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I.,
723 Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide
724 expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.
725 Malone, C.D., and Hannon, G.J. (2009). Small RNAs as Guardians of the Genome. Cell *136*,
726 656–668.
727 Muotri, A.R., Chu, V.T., Marchetto, M.C.N., Deng, W., Moran, J. V., and Gage, F.H. (2005).

728 Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature
729 *435*, 903–910.
730 Nekrutenko, A., and Li, W.H. (2001). Transposable elements are found in a large number of
731 human protein-coding genes. Trends Genet. *17*, 619–621.
732 Pastuzyn, E.D., Day, C.E., Kearns, R.B., Kyrke-Smith, M., Taibi, A. V., McCormick, J.,
733 Yoder, N., Belnap, D.M., Erlendsson, S., Morado, D.R., et al. (2018). The Neuronal Gene
734 Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA
735 Transfer. Cell *172*, 275-288.e18.
736 Perrat, P.N., Dasgupta, S., Wang, J., Theurkauf, W., Weng, Z., Rosbash, M., and Waddell,
737 S. (2013). Transposition-driven heterogeneity in the Drosophila Brain. Science *340*, 91–96.
738 Petri, R., Brattås, P.L., Sharma, Y., Jönsson, M.E., Pircs, K., Bengzon, J., and Jakobsson, J.
739 (2019). LINE-2 transposable elements are a source of functional human microRNAs and
740 target sites. PLOS Genet. *15*, e1008036.
741 Philippe, C., Vargas-Landin, D.B., Doucet, A.J., Van Essen, D., Vera-Otarola, J., Kuciak, M.,
742 Corbin, A., Nigumann, P., and Cristofari, G. (2016). Activation of individual L1
743 retrotransposon instances is restricted to cell-type dependent permissive loci. Elife *5*,
744 e13926
745 Rangwala, S.H., Zhang, L., and Kazazian, H.H. (2009). Many LINE1 elements contribute to
746 the transcriptome of human somatic cells. Genome Biol. *10*, 1–18.
747 Roebroek, A.J.M., Pauli, I.G.L., Zhang, Y., and van de Ven, W.J.M. (1991). cDNA sequence
748 of a Drosophila melanogaster gene, Dfur1, encoding a protein structurally related to the
749 subtilisin-like proprotein processing enzyme furin. FEBS Lett. *289*, 133–137.
750 Schauer, S.N., Carreira, P.E., Shukla, R., Gerhardt, D.J., Gerdes, P., Sanchez-Luque, F.J.,
751 Nicoli, P., Kindlova, M., Ghisletti, S., Dos Santos, A.D., et al. (2018). L1 retrotransposition is
752 a common feature of mammalian hepatocarcinogenesis. Genome Res. *28*, 639–653.
753 Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker.
754 Smith, L.A., Wang, X.J., Peixoto, A.A., Neumann, E.K., Hall, L.M., and Hall, J.C. (1996). A
755 Drosophila calcium channel α1 subunit gene maps to a genetic locus associated with

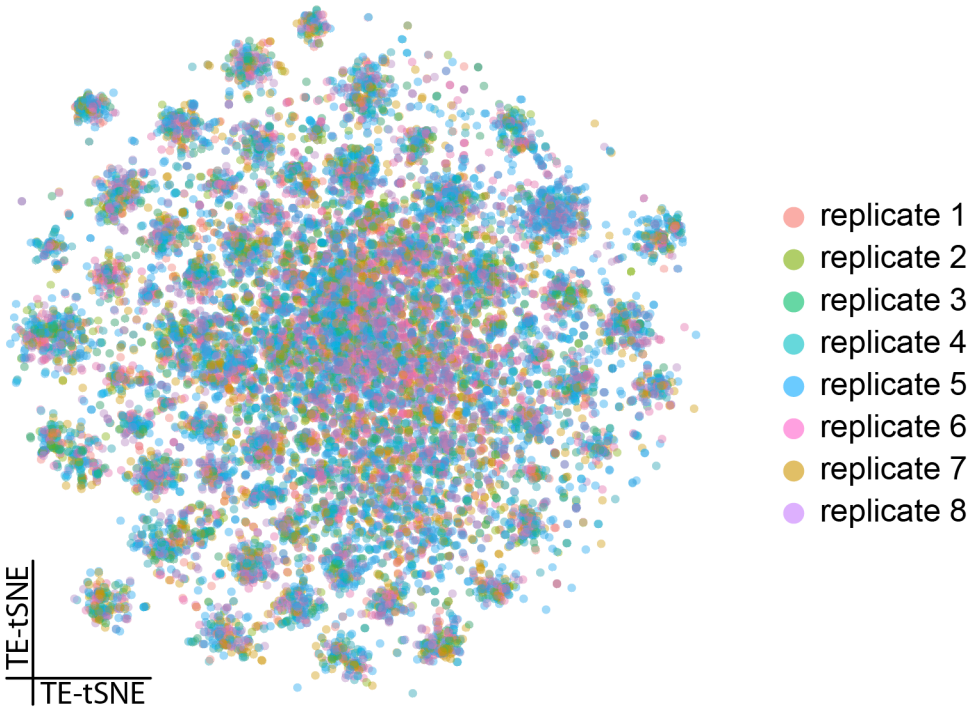
756 behavioral and visual defects. J. Neurosci. *16*, 7868–7879.
757 Stephens, R.M., and Schneider, T.D. (1992). Features of spliceosome evolution and function
758 inferred from an analysis of the information at human splice sites. J. Mol. Biol. *228*, 1124–
759 1136.
760 Sun, W., Samimi, H., Gamez, M., Zare, H., and Frost, B. (2018). Pathogenic tau-induced
761 piRNA depletion promotes neuronal death through transposable element dysregulation in
762 neurodegenerative tauopathies. Nat. Neurosci. *21*, 1038–1048.
763 Treiber, C.D., and Waddell, S. (2017). Resolving the prevalence of somatic transposition in
764 drosophila. Elife *6*, e28297
765 Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sánchez-Luque, F.J.,
766 Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., Van Der Knaap, M.S., Brennan, P.M.,
767 et al. (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell *161*, 228–239.
768 Wan, H.I., DiAntonio, A., Fetter, R.D., Bergstrom, K., Strauss, R., and Goodman, C.S.
769 (2000). Highwire regulates synaptic growth in Drosophila. Neuron *26*, 313–329.
770 Wang, Z., Berkey, C.D., and Watnick, P.I. (2012). The Drosophila Protein Mustard Tailors
771 the Innate Immune Response Activated by the Immune Deficiency Pathway. J. Immunol.
772 *188*, 3993–4000.
773 Zhang, W., Wu, J., Ward, M.D., Yang, S., Chuang, Y.A., Xiao, M., Li, R., Leahy, D.J., and
774 Worley, P.F. (2015). Structural basis of arc binding to synaptic proteins: Implications for
775 cognitive disease. Neuron *86*, 490–500.
776

Supplemental figures



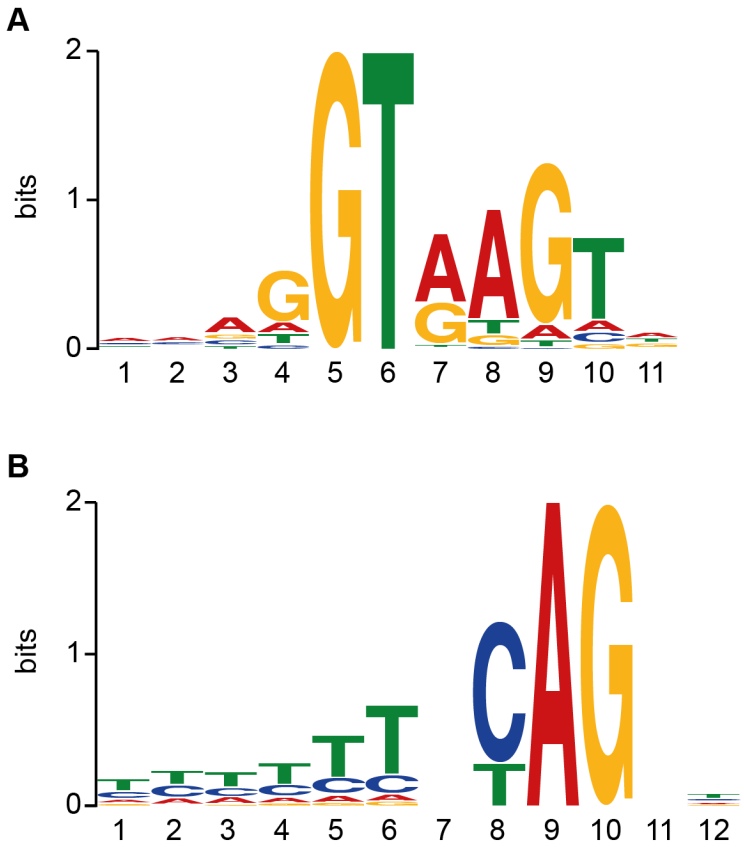
Supplemental figure 1. Sense strand transposon transcripts are twice as abundant as antisense strand transcripts in the *Drosophila* midbrain.

Graph showing mean expression levels across the entire midbrain of all sense and antisense transposon sequences. Each data point (cross) represents one biological replicate.



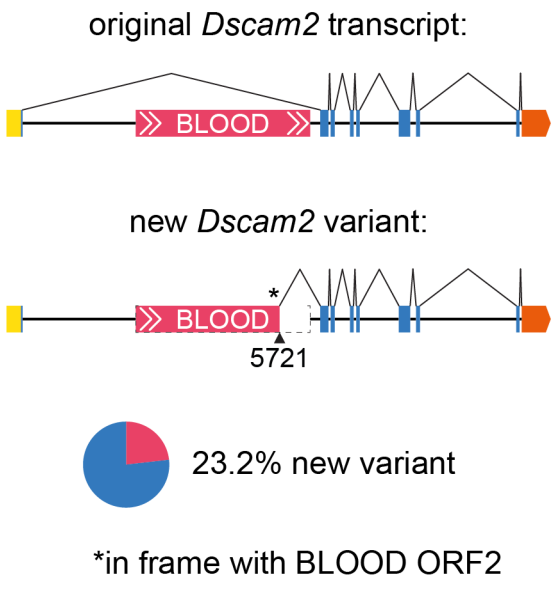
Supplemental figure 2. Transposon expression patterns are stereotyped across biological replicates.

tSNE based on transposon expression levels showing all 8 biological replicates. Each replicate contributes cells to each cluster.



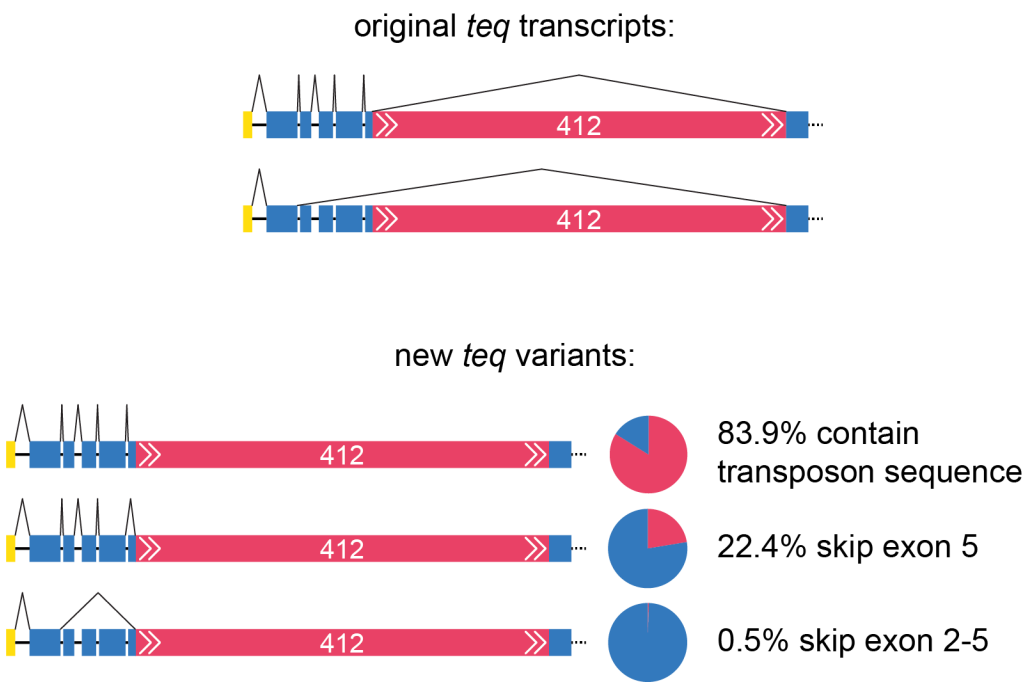
Supplemental figure 3. Splice acceptor and donor motifs.

A Splice acceptor motif, taken from 500 randomly chosen exon-intron junctions of *Drosophila* genes. **B** splice donor motif.



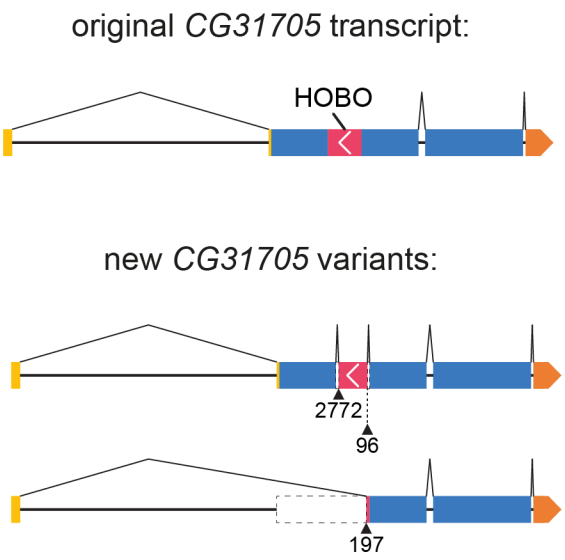
Supplemental figure 4. Sense BLOOD insertion in *Dscam2*.

Schematics of DScam2 mRNAs produced from locus containing BLOOD. Top shows the nascent transcript spliced around the intronic full-length sense BLOOD insertion. Bottom illustrates a new mRNA splice isoform, which reads through in frame from the ORF2 sequence of BLOOD into exon 2 of *DScam2*. The breakpoint in BLOOD is a consensus SD motif. 23.2% of all *Dscam2* transcripts in fly heads are chimeric with BLOOD.



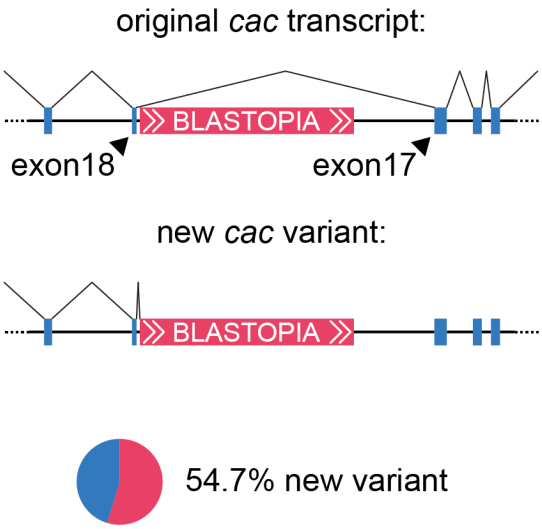
Supplemental figure 5. Sense 412 insertion in *tequila*.

Schematics of *teq* mRNAs produced from locus containing full-length sense orientation 412 insertion. Top, original transcripts of *teq* splicing around 412. Bottom, and new *teq* splice isoforms that include 412 sequence. 83.9% of *teq* transcripts contain 412 transposon sequence. In addition, 22.4 % of 412 containing mRNAs skip exon 5 of *teq*. In 0.5% of cases exons 2-5 are skipped.



Supplemental figure 6. Antisense HOBO insertion in *CG31705*.

Schematic of *CG31705* mRNAs produced from locus containing exonic antisense HOBO insertion. Transcripts containing unspliced HOBO and two additional new splice isoforms that are generated by alternative splicing into HOBO are shown.



818

819

820 **Supplemental figure 7. Sense BLASTOPIA insertion in *cacophony*.**

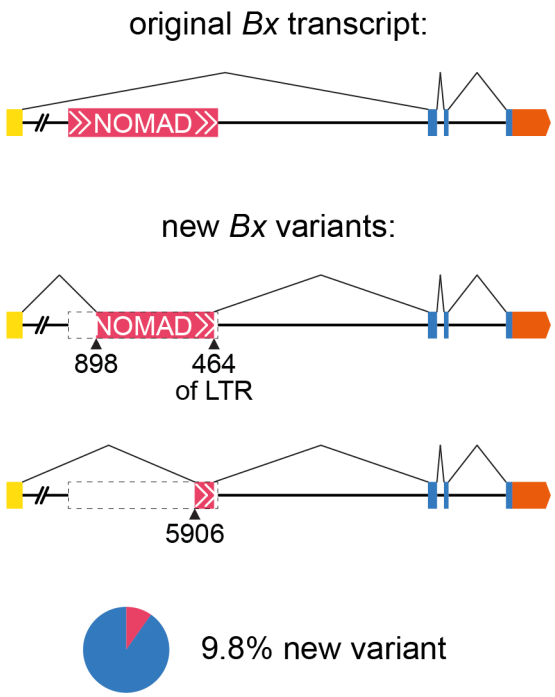
821 Schematic *cac* transcripts produced from locus containing a full-length intronic sense

822 BLASTOPIA insertion (the orientation is 5´ (left) to 3´ (right)). A regular cac transcript is

823 produced by splicing around the BLASTOPIA insertion and a new truncated *cac* isoform

824 results from splicing into BLAsTOPIA. The *cac* gene is on the X chromosome, and 54.7% of

825 *cac* transcripts in the fly head are truncated by splicing into BLASTOPIA.



826

827

828 **Supplemental figure 8. Sense NOMAD insertion in *Bx*.**

829 Schematic showing mRNAs produced from locus containing sense intronic NOMAD

830 insertion. Original transcript of *Bx* is generated by splicing around NOMAD. New splice

831 isoforms contain fragments of NOMAD. 9.8% of transcripts that start with the *Bx* 5´UTR are

832 spliced into NOMAD.