Nucleic Acids Research, 2025, **53**, gkaf485
https://doi.org/10.1093/nar/gkaf485
Gene regulation, Chromatin and Epigenetics

An endogenous retroviral element co-opts an upstream regulatory sequence to achieve somatic expression and mobility

Natalia Rubanova[1,2], Darshika Singh[3,†], Louis Barolle[3,†], Fabienne Chalvet[3], Sophie Netter[3,4], Mickaël Poidevin[3], Nicolas Servant[2], Allison J. Bardin[1,*], Katarzyna Siudeja[3,*]

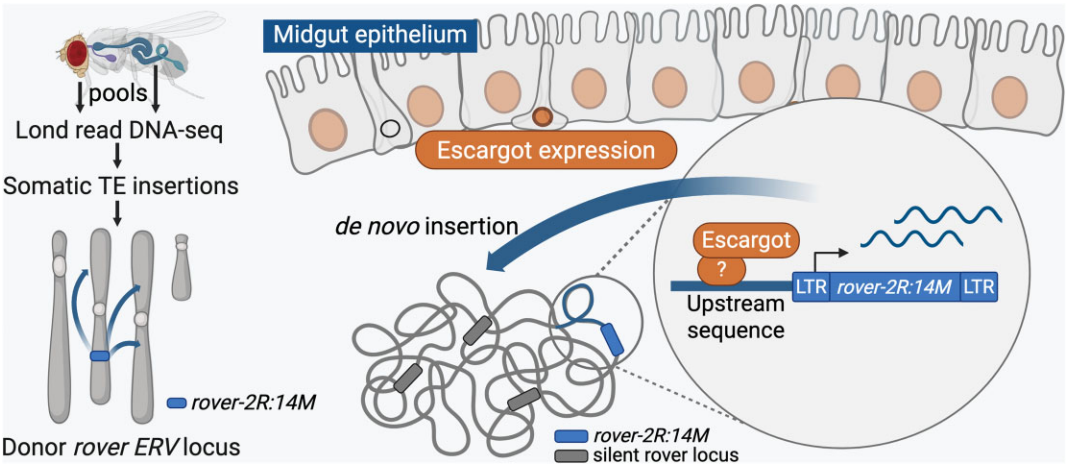
[1]Institut Curie, PSL Research University, CNRS UMR 3215, INSERM U934, Stem Cells and Tissue Homeostasis Group, Paris 75005, France
[2]Institut Curie Bioinformatics Core Facility, PSL Research University, INSERM U900, MINES ParisTech, Paris 75005, France
[3]Institute for Integrative Biology of the Cell (I2BC), INSERM U1280, CEA, CNRS, Université Paris-Saclay, Gif-sur-Yvette 91198, France
[4]Department of Biology, University of Versailles St-Quentin, Versailles 78035, France

[*]To whom correspondence should be addressed. Email: katarzyna-anna.siudeja@inserm.fr
Correspondence may also be addressed to Allison J. Bardin. Email: allison.bardin@curie.fr
[†]These authors contributed equally to this work.

Abstract

Retrotransposons, multi-copy sequences that propagate via copy-and-paste mechanisms, occupy large portions of eukaryotic genomes. A great majority of their manifold copies remain silenced in somatic cells; nevertheless, some are transcribed, often in a tissue-specific manner, and a small fraction retains its ability to mobilize. While it is well characterized that retrotransposon sequences may provide *cis*-regulatory elements for neighboring genes, how their own expression and mobility are achieved is not well understood. Here, using long-read DNA sequencing, we characterize somatic retrotransposition in the *Drosophila* intestine. We show that retroelement mobility does not change significantly upon aging and is limited to very few active sub-families. Importantly, we identify a donor locus of an endogenous LTR (long terminal repeat) retroviral element *rover*, active in the intestinal tissue. We reveal that gut activity of the *rover* donor copy depends on its genomic environment. Without affecting local gene expression, the copy co-opts its upstream genomic sequence, rich in transcription factor binding sites, for somatic expression. Further, we show that *escargot,* a snail-type transcription factor, can drive transcriptional activity of the active *rover* copy. These data provide new insights into how locus-specific features allow active retrotransposons to produce functional transcripts and mobilize in a somatic lineage.

Graphical abstract



Introduction

Transposable elements (TEs), repetitive genetic elements capable of self-replicating and moving from one position in the genome to another, represent a large part of eukaryotic genomes [1]. Class I long terminal repeat (LTR) and long interspersed nuclear element-1 (L1) retroelements mobilize by a copy-and-paste mechanism, through an RNA intermediate. They are often present in thousands of copies,

Received: January 17, 2025. **Revised:** April 24, 2025. **Editorial Decision:** April 27, 2025. **Accepted:** May 21, 2025

following what are believed to be waves of mobilization in genomes that date back millions of years (e.g HERV-K elements in human [2]). Many copies subsequently lose DNA sequence integrity over time by acquiring mutations that inactivate their mobilization machinery or are eliminated from genomes through recombination events. Thus, only a small number of copies remain intact and transposition-competent. Additionally, to protect genome integrity, host organisms have developed multiple pre- and post-transcriptional mechanisms of repression operating in the germline and somatic tissues [3, 4].

The impact of TEs on evolution and their vast contribution to genome regulation and cell biology are now widely recognized. TE sequences can be repurposed for the benefit of the host, can introduce heritable mutations that cause diseases, contribute to species variation, and regulate cellular transcriptional programs (e.g. development and neural progenitor cells) (reviewed in [3]). In addition, through the somatic mobility, they are a source of genetic mosaicism that may facilitate tumorigenesis and is of yet unclear role in healthy tissues. Moreover, apart from their mobility, TE RNA or protein products are believed to contribute to important biological processes, such as normal development [5, 6] or immune response [7–9]. Indeed, TE transcripts, including those with intact open reading frames (ORFs), are repeatedly detected in many tissue types, often with tissue- or cell type-specific expression patterns [10–19]. Nevertheless, how TE expression and mobility are regulated in diverse somatic contexts is not well understood.

Addressing TE regulation in the soma is complex due to the highly repetitive nature of TE sequences, which complicates bioinformatics analysis [20, 21]. Indeed, TE transcripts are derived only from a small subset of active, de-repressed loci [16–18, 22–26]. Additionally, a further challenge is the difficulty in detecting somatic TE mobility, which occurs in a small proportion of cells and thus represents only rare events. To date, studies on somatically active TE loci have largely focused on mammalian L1 elements, for which mobility-competent loci were identified [27–34]. These donor L1 loci were shown to carry unmethylated promoters, lineage-specific transcription factor (TF) motifs, or deletions of binding sites for repressive factors, all contributing to their activity [22–24, 35, 36].

In contrast to L1 elements, identification of somatically active loci of LTR retrotransposons and an understanding of their regulation have been lagging behind. This is in part because of the fact that no replication-competent LTR elements have been identified in the human genome. However, their transcription in different tissues is substantial [16–18], and they continue to mobilize in other species, including model organisms such as mouse [37] or *Drosophila* [38, 39]. It is generally believed that expression of LTR retrotransposons is determined by the 5′ untranslated region (5’ UTR), including the LTR sequence itself [40–45]. Indeed, sequence variations of these regulatory regions were proposed to explain transcriptional activity and tissue- or cell type-specific patterns of expression of LTR retroelements [16–19]. However, other levels of regulation are likely to play a role, and they may be uncovered through identification and careful analysis of somatically active LTR-element loci.

In our previous work, we showed that retroelements are expressed and mobile in *Drosophila melanogaster* intestinal tissue and that somatic transposition can lead to gene inactivation via LTR-element insertion [39]. Here, using long-read DNA sequencing, we further characterize this mobility in healthy somatic tissues isolated from flies of different ages. Importantly, we uncover a donor locus of a *rover* sub-family of endogenous retroviruses (ERVs). ERVs constitute a subclass of LTR retrotransposons, which acquired an *envelope* (*env*) ORF, in addition to the *gag* (capsid) and *pol* ORFs carried by all LTR retroelements [46, 47]. We demonstrate that the expression of the somatically active *rover* retroviral locus is driven by its genomic *cis*-regulatory elements, providing new insight into regulation of LTR/ERV elements by their genomic environment.

Materials and methods

Experimental techniques

Drosophila stocks

Unless stated otherwise, *Pros>2xGFP* (*ProsGFP*) stock was used for most sequencing experiments, the same as in [39]. This genotype is obtained by crossing ;;*Pros*^{V1}*GAL4/TM6BTbSb* females (J. de Navascués) with ;*UAS-2xGFP;* males (Bloomington, #6874). The characterized *rover-2R:14M* insertion is present in the ;;*Pros*^{V1}*GAL4/TM6BTbSb* stock, but not in the *;UAS-2xGFP;*. Thus, it is heterozygous in the *ProsGFP* animals. w^{1118} flies (gift from M. McVey, Tufts University, USA) were used for RNA-seq as a control stock not carrying the *rover-2R:14M* insertion. To generate flies carrying the *rover-2R:14M* insertion in an Ago2 mutant background, we isolated the second chromosome carrying *rover-2R:14M* insertion from the ;;*Pros*^{V1}*GAL4/TM6BTbSb* stock and backcrossed the *rover-2R:14M* insertion as well as the $Ago2^{414}$ mutant allele (from C. Saleh, Institut Pasteur, Paris, France) to the same w^{1118} background for seven generations. We then combined the two stocks using standard crosses with balancer lines to obtain flies homozygous for the *rover-2R:14M* insertion and the $Ago2^{414}$ mutation. *Rover-2R:14M-lacZ* reporter lines were generated in this study (see below). Other stocks used included: UAS-*white RNAi* (BL, #33762); *UAS-spn-E RNAi* (BL, #34808); *UAS-zuc RNAi* (E. Brasset, iGReD, Clermont-Ferrand, France); *UAS-esg RNAi* (BL, #34063); *tj-Gal4* and *nos-Gal4* (L. Teysset, C. Carré, IBPS, Paris, France); and *esgGAL4 UAS-GFP GAL80ts* [48]. Full genotypes corresponding to main and supplementary figures are listed in the supplementary information file.

Fly husbandry

Flies were maintained on a standard medium at 25°C with a day/night light cycle. For crosses, 10–15 females were mixed with males in standard vials. For larval experiments, third instar wandering larvae were selected. Adult progeny was collected over 2–4 days after eclosion and kept at a density of 25–30 flies/tube (mixed sexes). Flies were flipped to fresh tubes every 2–3 days until needed. For aging experiments, three time points were analyzed: young 5–7-day-old, mid-age 20–25-day-old, and aged 50–60-day-old. For temporal induction of *white* or *esg* RNAi in adult gut progenitor cells (using *esgGAL4 UAS-GFP GAL80ts* [48]), crosses were maintained at 18°C, and 5–10-day-old adult flies were switched to 29°C for 2 days. Females were used for most experiments, unless stated otherwise.

Genomic DNA isolation

Tissues were dissected in ice-cold, nuclease-free phosphate buffered saline (PBS) and snap-frozen in liquid nitrogen before DNA isolation. High molecular weight genomic DNA was isolated from pools of 60 guts or 60 heads with the MagAttract HMW DNA Kit (Qiagen, #67563) or *Quick*-DNA HMW MagBead Kit (Zymo Research, #D6060) according to manufacturers' instructions, with tissue lysis performed overnight at 55°C. Genomic DNA (gDNA) was eluted with nuclease-free water. DNA integrity was verified on a 0.6% agarose gel and concentrations were measured with Qubit dsDNA Broad Range Assay Kit. All samples had A260/280 ratios above 1.8 and A260/230 ratios above 2.0.

DNA sequencing

Whole genome long-read DNA sequencing libraries were prepared with 500–800 ng of genomic DNA following the 1D Genomic DNA Ligation Protocol (SQK-LSK109, Oxford Nanopore Technologies). Sequencing was performed on MinION or GridION using R9.4.1 flow cells (Oxford Nanopore Technologies) and 48-h-long sequencing runs. Supplementary Table S1 provides basic sequencing statistics for all samples.

RNA isolation

For RNA isolation, gut and head tissues were dissected in cold, RNase-free PBS, transferred to 100 µl of TRIzol reagent (Thermo Fisher Scientific), homogenized with a plastic pestle, and snap-frozen in liquid nitrogen for storage at −80°C. Upon thawing, samples were further processed according to the TRIzol reagent manufacturer's protocol. Purified RNA was treated with DNase (Ambion) for 1 h at 37°C, further purified with phenol–chloroform extraction and isopropanol precipitation, and resuspended in RNase-free water. All samples had A260/280 ratios above 1.9 and A260/230 ratios above 2.0. RNA integrity was checked on Bioanalyzer (Agilent) using the Agilent RNA 6000 Nano Kit and concentrations were assayed with the Qubit RNA Broad Range Assay Kit (Thermo Fisher Scientific).

RNA sequencing

For the short-read transcriptome analysis, 700 ng of total RNA was used to prepare libraries according to the TruSeq Stranded mRNA protocol (Illumina). Samples were processed in biological triplicates. 2 × 100 bp paired-end sequencing was performed on NovaSeq (Illumina). For long-read Oxford Nanopore (ONT) complementary DNA (cDNA) sequencing, we first prepared messenger RNA (mRNA) with Dynabeads mRNA Purification Kit (Invitrogen, #61006) starting from 100 µg of DNase-digested total RNA, according to the manufacturer's protocol. mRNA concentration was quantified with the Qubit RNA Broad Range Assay Kit (Thermo Fisher Scientific) and sample quality was checked on Bioanalyzer (Agilent). Samples were then prepared according to the protocol for ONT direct cDNA sequencing (SQK-DCS109, Oxford Nanopore Technologies), using 150 ng of purified mRNA as input. Samples were run on MinION using R9.4.1 flow cells (Oxford Nanopore Technologies) and 48-h-long sequencing runs. Supplementary Table S1 provides basic sequencing statistics for all samples.

Generation of *lacZ* and *luciferase* reporters

To construct reporter plasmids, we amplified the *rover-2R:14M* 5′UTR region alone or with its upstream genomic sequence of 2 or 5 kb from *ProsGFP* genomic DNA. These sequences were then cloned into appropriate vectors using NEBuilder® HiFi DNA Assembly protocol (New England Biolabs, #E5520). lacZ reporter plasmids were obtained by replacing the *hsp70* promoter sequence from the placZ-attB [49] vector upstream of the *lacZ* gene with the regulatory sequences of interest. The plasmids were then injected by Bestgene (Chino Hills, CA, USA) for integration into two different landing sites: VK38 (Ch X, Bloomington stock #9753) and VK05 (Ch 3L, Bloomington stock #9725). For S2 cell reporter plasmids, we used the same cloning strategy to place the *rover-2R:14M* regulatory sequences into the pAct-GL3 vector (Promega) upstream of the luciferase gene. pAct-Renilla vector was used to correct for transfection efficiency. All final constructs were checked by ONT whole plasmid sequencing (Eurofins Genomics).

Immunolabeling

Midguts were dissected and fixed in 4% paraformaldehyde (PFA, Fisher Scientific S.A.S., 15828264) and 1× PBS for 3 h. Fixed tissues were washed with PBT (1× PBS, 0.1% Triton X-100) and incubated for at least 30 min in 50% glycerol and 1× PBS, followed by placing back into PBT, to allow the waste to exit by osmotic pressure. The tissues were then incubated with primary antibodies at 4°C overnight, washed three times with PBT (20 min each), and incubated with the secondary antibodies for 3 h at room temperature. Next, the guts were washed three times in PBT and stained with DAPI at the last wash. Stained tissues were equilibrated in 50% glycerol and 1× PBS for at least 1 h before mounting on microscopy slides in mounting medium. Ovaries were dissected in 1× PBS, fixed in 4% PFA for 20 min, and washed 3 × 20 min in PBT (PBS, 0.3% Triton X-100). Ovaries were then incubated in a blocking solution (PBTA: PBT, 2% bovine serum albumin; Sigma, A3059) for a minimum of 30 min. Primary antibodies were diluted in PBTA and ovaries were incubated in this solution overnight at 4°C. Tissues were then washed in PBT 3 × 10 min and incubated in PBTA for at least 30 min before incubation in secondary antibodies in PBTA for 3 h at room temperature. Finally, tissues were washed 3 × 10 min in PBT, and mounted in DABCO (D27602, Sigma) with 70% glycerol.

The following primary antibodies were used: chicken anti-β-Gal (Abcam, #9361) and mouse anti-Prospero (DSHB, #MR1A-c).

Microscopy

Images were obtained using an upright confocal laser scanning microscope TCS SP8, Leica (Leica Microsystems, Germany), using an HC PL APO CS2 63×/1.4 oil immersion objective lens. Twelve-bit numerical images were acquired with the Leica Application Suite X software (LAS version 3.5.6; Leica, Germany) and processed using Fiji (ImageJ [50]) version 1.53c. Adult and larval guts were imaged in the posterior R4 region (according to [51]). To quantify β-Gal signal (Fig. 6F), we used a custom macro available upon request. Briefly, a maximum intensity projection of the z-stacks was generated and a binary mask was created on the green channel. Then, Analyze Particles function was used to delineate individual cells (criteria size, larger than 10 µm^2) and mean fluorescence intensities were measured for each cell, with measurements restricted to the defined regions of interest corresponding to the GFP-positive cells.

S2 cell culture

Drosophila Schneider's S2 cells (DGRC, #181) were maintained at 25°C in Schneider's *Drosophila* medium (Invitrogen) supplemented with 10% heat inactivated fetal calf serum (Gibco) and antibiotics (penicillin/streptomycin, Invitrogen). Cells in the exponential phase of growth were used for all the experiments.

Luciferase assays

S2 cells were transiently transfected with Effectene Transfection Reagent (Qiagen, #301425) according to the manufacturer's protocols, to deliver the following vectors: (i) luciferase reporter plasmid, (ii) Esg overexpression plasmid under constitutive promoter (pAc5.1), and (iii) Renilla construct for normalization. The ratio of luciferase:renilla plasmids was kept at 10:1. The total amount of transfected DNA was kept constant by adjusting with an empty vector. Levels of luciferase and Renilla were measured 48 h after transfection using Dual-Glo Luciferase Assay System (Promega, #E2920) and Spectramaax M5 (Molecular Devices). For each transfection, at least three biological replicates were performed, each done in triplicate.

cDNA for Esg was amplified from the DGRC stock number 1645028 (RRID:DGRC_1645028).

Computational analysis

Processing of long-read DNA-seq samples

Raw sequencing reads from the ONT DNA sequencing libraries generated in this study and in the ONT DNA sequencing libraries from [39] were basecalled using guppy v6.0.1, a basecaller developed by ONT, with the dna_r9.4.1_450bps_hac model. Basecalled reads were merged into a single FASTQ file for each library. Sequencing adapters were trimmed using Porechop v0.2.4 (https://github.com/rrwick/Porechop). The reads were filtered using NanoFilt v2.8.0 [52] with filters for a minimum average read quality score (-q 10) and a minimum read length (-l 500). The reads were aligned to the FlyBase dm6.48 [53] reference genome using minimap2 [54] with -x map-ont preset and the parameter to retain the MD tag (–MD). The resulting alignments were sorted, filtered for a minimum mapping quality of 5, and indexed using samtools v1.13 [55]. Sequencing and alignment quality were assessed with NanoPlot v1.36.2 [56] and pycoQC v2.5.2 [57].

Calling non-reference TE insertions in long-read DNA-seq samples

tldr v1.2.2 [58] was used to call non-reference insertions with the following parameters: –color_consensus –trdcol –detail_output –minreads 1 –min_te_len 500 –max_cluster_size 100. The TE consensus sequences for *D. melanogaster* species made available by the Bergman's lab (https://github.com/bergmanlab/drosophila-transposons) were used as the reference TE library. The following criteria were used to filter raw calls:

1. the fraction of the inserted sequence covered by the TE sequence >0.5;
2. the length of the inserted sequence >500 bp;
3. the median mapping quality score of the reads supporting the insertion ≥30;
4. the chromosome of the insertion was the autosomes or chrX;
5. the coverage at the integration site ≥10.

Calls meeting the following criterion were classified as full-length insertions: (total length of the insertion) × (fraction of the inserted sequence covered by the TE sequence)/(consensus TE length) ≥ 0.7. All other calls were classified as truncated insertions. Next, calls from all samples were clustered based on sub-family, genomic breakpoint coordinates (allowing 100 bp margin), and DNA strand. Clusters were reviewed and filtered based on the presence of full-length calls, target site duplications (TSDs), the minimum coverage at breakpoints, and the status of the "remappable" filter in the tldr output. Clusters were retained if they satisfied the following conditions:

1. at least one call in the cluster was for a full-length insertion;
2. the breakpoint region had coverage <200;
3. coverage at the integration site in at least one sample was ≥15;
4. at least one call in the cluster passed "remappable" filter in the tldr ouput.

The calls in each retained cluster were collapsed, and

1. left genomic breakpoint of the cluster was defined as the median of all left genomic breakpoints in the cluster;
2. right genomic breakpoint of the cluster was defined as the median of all right genomic breakpoints in the cluster;
3. ONT read ratio of the cluster was defined as the median of all ONT read ratios in the cluster;
4. TSDs as well as samples and tissues that supported each cluster were recorded.

Clusters within 1 kb of a reference insertion of the same sub-family were removed, as they likely represented misalignments or DNA-based events. This was supported by the fact that most such clusters included calls with TSDs >100 bp and/or long non-TE flanking regions.

The remaining clusters were further iterated over, and clusters within a 2 kb margin of genomic breakpoints for the same sub-family were extracted. The following rules were applied:

1. if no cluster in the subsection had the "PASS" filter flag in the tldr output, only the cluster with the highest number of supporting samples was kept;
2. if only one cluster in the subsection had the "PASS" filter flag in the tldr output, only this cluster was kept;
3. if multiple clusters in the subsection had the "PASS" filter flag in the tldr output, the cluster supported by both tissues was kept (if such a cluster was present in the subsection); otherwise, all clusters in the subsection were kept.

Calling reference TE insertions

RepeatMasker v4.1.2 (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0) with CONS-Dfam_withRBRM_3.5 library [59] was used to mask dm6.48 reference genome. one_code_to_find_them_all v1.0 [60] was used to assemble the masked sequences into complete TE copies. Insertions that were longer than 70% of the respective consensus sequences and with ≤30% sequence divergence relative to the respective consensus sequences were called full-length. To detect reference insertions present in the genome of *ProsGFP* strain, reads with primary alignments with mapping quality >5 span-

ning regions upstream and downstream of a reference insertion were extracted from the ONT bam files using custom Python scripts and the pysam module (https://github.com/pysam-developers/pysam). Reads with alignments of at least 100 bp within reference TE coordinates were classified as supporting the insertion, while all other reads were classified as opposing the insertion. A reference insertion was considered present in the genome if it was supported by at least two reads in at least two ONT samples.

Genotyping TE calls in ONT samples

The clustered non-reference and reference calls were genotyped in two steps. Initial genotype assignments were made based on the following criteria:

- tldr calls supported by a single read and detected only in one ONT sample were assigned "singleton" genotype;
- tldr calls with ONT read ratio <0.1 that were detected in both tissues were assigned "rare"genotype;
- tldr calls with ONT read ratio ≥0.1 were assigned "non-reference" genotype;
- RepeatMasker calls were assigned "reference" genotype;
- all other calls were considered as "ungenotyped."

Calls detected in only one tissue were further examined as follows. For each such call, the clipped parts of reads longer than 100 bp were extracted from all samples, except the sample where the call was detected, within a 200 bp margin of the left genomic breakpoint, using pysam Python module. The extracted parts were mapped to the consensus TE sequences with mappy Python module. Primary alignments with a mapping quality higher than 15 were retained. The genotype, sample, and tissue supporting a call were updated if an alignment hit was found for the same sub-family. This procedure updated ∼5% singleton calls, 30% of ungenotyped calls, and 3% of non-reference calls. The updated calls were supplemented with the Illumina variant allele frequency (VAF), the number of Illumina sample pairs supporting a call, and the genotype inferred from the Illumina samples if a call matched a call in the Illumina samples for the same sub-family, on the same chromosome, and within a 100 bp margin of the genomic breakpoints. Singletons without a TSD footprint and those belonging to germline-active sub-families were considered as "ungenotyped".

Among the ungenotyped calls, 73 calls with supporting reads that poorly mapped to one TE consensus sequence and 4 calls with supporting reads that mapped to several TE consensus sequences were removed. To achieve this, mapping quality, the number of reads, and TE subfamily information for each call were obtained from the tldr detail.out output files. Among the remaining ungenotyped calls,

- 117 insertions were singletons without a TSD;
- 88 insertions were supported by an ambiguous set of samples (e.g. two samples from the same tissue);
- 87 insertions could potentially be considered somatic insertions of high clonality (i.e. insertions detected in one sample but supported by more than one read);
- 56 insertions were singletons from germline-active sub-families.
- 12 insertions were singletons with a TSD longer than 50 bp.

Calling non-reference germline TE insertions in short-read DNA-seq samples

The coordinates for somatic insertions detected in the short-read Illumina samples from [39] were obtained from Supplementary Table S2. The coordinates for the germline insertions detected in the short-read Illumina samples were obtained from Supplementary Table S7 of our previously published study [39]. The same Illumina samples (samples P7-P66, accession number PRJNA641572) were reanalyzed using a second independent bioinformatics approach. For this, ngs_te_mapper2 v1.0.2 [61] was used with default parameters, except for –min_af 0.01. Calls from ngs_te_mapper2 in all samples were collapsed based on sub-family, breakpoint coordinates, and strand, allowing a 50 bp margin for the coordinates. The resulting callset was merged with the somatic and germline callsets from [39], mentioned above, based on sub-family and breakpoint coordinates, also allowing a 50 bp margin for the coordinates. The samples and the tissue supporting each insertion were recorded. The VAF of an insertion was defined as the average of the VAFs in the supporting samples. An insertion was assigned a genotype based on the following criteria:

- "germline" if it was detected in at least three pairs of samples (i.e. gut and head samples from one fly);
- "private germline" if it was detected in exactly one sample pair;
- "rare germline" if it was detected in >1 and <3 sample pairs.

Definition of germline active sub-families

The sub-families were sorted in descending order based on the total sum of "rare germline" and "private germline" insertions detected in the Illumina samples. A cutoff of five insertions was set to define the germline-active sub-families.

Normalization of raw singleton counts

To normalize raw singleton counts in a sample, the raw singleton count of a sub-family was multiplied by 1000 and divided by the number of reads longer than 3.5 times the respective consensus TE length in the sample. This approach normalized raw counts for sequencing depth, considering only reads whose both flanks could potentially be aligned to the genome in the presence of a TE insertion, as this is a necessary condition for detecting a singleton insertion. The number of reads in a sample was calculated by extracting the reads with primary alignments from the respective BAM file and calculating their length.

A one-way ANOVA test and Holm's multiple test correction method were used to compare the mean normalized counts between three age groups.

Creating consensus sequences for *rover* reference and non-reference insertions

To create consensus sequences for reference insertions, the parts of the reads between reference TE coordinates that had alignments longer than 200 bp within the coordinates were extracted from the ONT BAM files with custom Python scripts. Multiple sequence alignments of the extracted sequences supporting each reference insertion were performed using MAFFT [62], and the consensus sequences were created using the cons tool from the EMBOSS package [63].

To create consensus sequences for non-reference insertions, the ONT BAM files were parsed with custom Python scripts, and inserted sequences ("I" in CIGAR string or "1" in CIGAR tuple in pysam) between genomic breakpoint coordinates were extracted. Extracted sequences supporting an insertion were aligned to the *rover* consensus sequence with minimap2. Alignments were sorted and indexed with samtools view and samtools index. A consensus sequence for an insertion was created with samtools consensus.

Identification of sequence variants supporting *rover* donor locus

Singletons in each ONT sample were aligned to the *rover* consensus sequence using minimap2, sorted and indexed using samtools, and visually inspected in Integrative Genomics Viewer (IGV) [64].

The consensus sequences of 15 reference and 5 non-reference *rover* insertions present in the genome of the *ProsGFP* strain were first visually compared to the *rover* consensus sequence by generating and visualizing multiple sequence alignments with Mauve [65]. Insertions with large structural variants (SVs) that were not present in singletons were excluded from the subsequent analysis. The sequences of the remaining four reference and four non-reference insertions were aligned to the *rover* consensus sequence using minimap2, sorted and indexed with samtools, and visually inspected in IGV. Visual inspection of 50–300-bp-long SVs present in singletons and germline insertions allowed to rule out additional six germline insertions. The non-reference *rover-2R:14M* and *rover-2L:18M* insertions differed only by a set of 10 SNVs. The fraction of singletons supporting each SNV present in the *rover-2R:14M* and not present in the *rover-2L:18M* was quantified using freebayes (https://doi.org/10.48550/arXiv.1207.3907).

Creating genome and transcriptome references for expression analysis

To quantify *rover* consensus expression, the masked dm6.48 genome was supplemented with an artificial chrTE that consisted of consensus sequences of all TEs present in the *D. melanogaster* TE library. To quantify *rover* per copy expression, the *rover* sequence in chrTE was masked, and an additional chrRS was added, which consisted of sequences of all reference and non-reference *rover* insertions present in the genome of the *ProsGFP* strain.

Illumina RNA sequencing analysis

To quantify *rover* consensus expression, Illumina RNA-seq libraries from [39] and Illumina RNA-seq libraries generated in this study were aligned to the masked reference genome plus chrTE using STAR 2.7.10a [66] with the following parameters: –sjdbOverhang 100 –outFilterMultimapNmax 1 –winAnchorMultimapNmax 10 –outMultimapperOrder Random –outFilterMismatchNoverLmax 0.3 –quantMode GeneCounts quantification mode. The expression count tables were grouped for all samples, and the counts were normalized using the Trimmed Mean of M-value (TMM) normalization method from the edgeR [67] R package. Differential expression analysis was performed using DESeq2 [68] R package, and the volcano plots were plotted using the ggplot2 package [69].

To quantify *rover* per copy expression, FASTQ files with the reads that had primary alignments within the *rover* consensus sequence region on the chrTE in the BAM files produced by STAR were created using view, collate, and fastq from samtools. These were then realigned to the masked chrTE plus chrRS genome using STAR 2.7.10a with the following parameters: –outFilterMultimapNmax 57 –winAnchorMultimapNmax 100 –outFilterMismatchNmax 999 –outFilterMismatchNoverLmax 0.01 –quantMode GeneCounts quantification mode. The expression count tables were grouped for all samples. A new count table was created that included gene and consensus TE counts (excluding *rover*) from the mapping to the masked reference genome plus chrTE, and *rover* fixed insertion counts from the mapping to masked chrTE plus chrRS. The counts were normalized using the TMM normalization method from the edgeR R package.

Analysis of ONT cDNA sample

Raw reads were basecalled with guppy v6.0.1 using the following parameters for the flow-cell and sequencing kit: –flowcell FLO-MIN106 –kit SQK-DCS109. Basecalled reads were trimmed and oriented with pychopper v2, developed by ONT.

To quantify *rover* consensus expression, basecalled reads were aligned to the masked reference genome plus chrTE using minimap2 with the -x splice preset and the –secondary=no parameter. Read counts were quantified using featureCounts v2.0.1 from the Subread package [70] with the following parameters: -Q 5 -L. Raw read counts are reported in Fig. 3.

To quantify *rover* per copy expression, FASTQ files with the reads that had primary alignments within the *rover* consensus sequence region on the chrTE in the BAM files produced by STAR were created using view, collate, and fastq from samtools. These were then realigned to the masked chrTE plus chrRS genome using minimap2 with the -x map-ont preset and the –secondary=no parameter. Reads were quantified with featureCounts as described above.

Motif analysis of the rover LTR region

Position weight matrices for the components of the polymerase II (PolII) transcription initiation complex [TATA-box, initiator element (Inr), downstream promoter element (DPE), and motif ten element (MTE)] were created from the frequency tables from [71] using TFBSTools [72] in R. The first 363 bases of the *rover-2R:14M*, *rover-2L:18M*, and *rover-2R:21M* insertions were scanned with parameters for a minimum motif score 80% and the "+" strand using the searchSeq function in TFBSTools. Motif scores were scaled to 1 separately for each element and plotted using custom Python scripts.

Motif analysis of the rover internal 2-kb region

Position weight matrices from the JASPAR database [73] for TF binding profiles were used to scan the internal 2-kb regions of the *rover-2R:14M*, *rover-2L:18M*, and *rover-2R:21M* insertions with parameters for a minimum motif score 80% and the "±" strands, in the same way as described above. Motif scores were scaled to 1 separately for each TF. Expression values for each TF in ISC, EB, EC, and EE cell types were taken from [74].

Motif analysis of the *rover-2R:14M* upstream region

ChIP-seq BED tracks with peaks for the *D. melanogaster* dm6 reference genome were downloaded from modENCODE [75].

The tracks were intersected with the upstream genomic region for the *rover-2R:14M* locus, and TFs expressed in ISC, based on the expression data from [74], were plotted with custom R scripts.

Epigenetic data analysis

Heatmaps for DamID tracks for PolII, Brm, Pc, HP1, and H1 factors for ISC, EB, EC, EE, and ATAC-seq (assay for transposase-accessible chromatin) tracks for ISC from [76] (accession number PRJNA933194) were plotted for 10-kb upstream genomic regions of the full-length reference and non-reference *rover* insertions using computeMatrix and plotHeatmap from deepTools [77].

Results

The genomic landscape of reference, non-reference, rare, and somatic TE insertions

To gain new insight into somatic mobile element activity, we explored our previously published long-read DNA sequencing datasets [39] and generated new, complementary libraries (Supplementary Table S1). Using Oxford Nanopore Technology (ONT), we sequenced DNA isolated from pools of *Drosophila* heads or intestines, using one of our previously characterized genetic backgrounds with documented TE mobility in the gut (*ProsGAL4>UAS-GFP* [39, 78], hereafter abbreviated as *ProsGFP*). Of note, the *ProsGFP* flies are wild type for the known TE controlling pathways [39]. For each tissue type, we sequenced in two replicates DNA libraries from young (5–7-day-old), mid-aged (25–30-day-old), and old (55–60-day-old) female flies, obtaining on average a sequencing depth of 50× (Supplementary Table S1).

First, we aimed to precisely characterize the TE landscape of the studied genotype, focusing on full-length copies (that comprise more than 70% of the respective consensus sequence; see the "Materials and methods" section). To do this, we analyzed long-read DNA sequencing libraries for TE insertions annotated in the dm6 reference genome and non-reference TE insertions present in the genome of the *ProsGFP* strain. Among 4444 full-length TEs (Supplementary Table S2), 1685 (38%) were also found in the dm6 reference genome (Fig. 1A). The remaining 62% were non-reference. We genotyped all detected full-length TE insertions based on their (i) presence in the reference genome; (ii) ONT read ratio (number of supporting versus opposing reads); (iii) tissues in which the insertion was detected; and (iv) germline activity of their respective sub-families (details in the "Materials and methods" section). We further used our 31 previously published Illumina DNA sequencing libraries of head and intestine tissues coming from individual flies of the *ProsGFP* genetic background [39] to validate genotype assignment. As expected, reference TEs were mostly detected in each sample and had in general ONT read ratios close to 1 (Fig. 1B and Supplementary Table S2), suggesting that they represented homozygous insertions in the *ProsGFP strain*. We subdivided TE insertions not present in the reference genome into four categories: "non-reference," "rare," "singleton," and "ungenotyped," described further below.

We first characterized insertions that were detected in at least one ONT sample of each tissue with ONT read ratio greater than or equal than 0.1 (1227 insertions in total), which were assigned the "non-reference" genotype (Fig. 1A and B). A high number of non-reference insertions was consistent with previous reports on the high variation in TE composition in *D. melanogaster* laboratory strains [38, 79]. Only 60% (747 out of 1227) of "non-reference" insertions were detected in our previously published Illumina DNA sequencing libraries from the same genetic background [39], 98% of which were classified as germline TE insertions. ONT read ratios of the non-reference insertions peaked at 0.5 or 1 (Fig. 1B). The distribution of Illumina VAFs and population frequencies of these insertions showed that more than half of them were likely germline insertions nearly fixed in one or both parental strains, crossed to obtain the *ProsGFP* F1 animals (433 out of 747 insertions had population frequency $\geq$0.8 and Illumina VAF $\geq$0.3) (Supplementary Fig. S1A). In the *ProsGFP* genotype, these insertions were thus present as homozygous or heterozygous, the latter inherited from only one parental strain. The remaining non-reference calls likely represented polymorphic insertions.

One hundred seventy-one insertions (171) with ONT read ratio lower than 0.1 that were detected in at least one sample of each tissue were assigned a "rare" genotype (Fig. 1A and B). Low ONT read ratios suggested that they represented rare germline insertions, present only in some individuals (as in [80]). Only 25 of them (15%) were also detected in the short-read samples. Together, this analysis highlights the high variation in TE composition between the reference genome and the investigated genetic background and the advantage of long-read sequencing to comprehensively detect TE sequences.

We then focused on somatic insertions. We defined these as in our previous study [39], as insertions being supported by a single read present in only one sequencing library ("singleton") that had a TSD as a footprint of a true transposition event. In addition, we excluded sub-families active in the germline (Supplementary Fig. S1B). We detected 1001 of such "singleton" insertions (Fig. 1A and B). We have previously provided strong evidence that "singletons" can confidently be considered as true somatic insertions in this experimental and computational setup [39], and we further introduced additional filtering steps to minimize potential false positives (see the "Material and methods" section for details).

Finally, the insertions that did not meet the criteria for "non-reference," "rare," and "singleton" genotypes were categorized as "ungenotyped" insertions. We detected 360 such insertions (Fig. 1A). These are insertions with an ambiguous set of supporting samples, singletons lacking TSD footprint, or singletons from the sub-families active in the germline. They could represent artifacts, very rare germline insertions, somatic insertions including embryonic and developmental events, or a mixture of these three, which we cannot confidently distinguish.

Having described the TE landscape of the *ProsGFP* strain, we then aimed to investigate the genomic TE distribution (Fig. 1C). In agreement with previous reports [38, 81], reference insertions were found enriched in pericentromeric, gene-poor chromosome regions, and on the mostly heterochromatic chromosome 4. Conversely, non-reference insertions were present throughout all chromosome arms, without any significant "hot spots" similarly to the rare insertions. Accumulation of reference insertion in pericentromeric regions is likely an effect of a negative selection acting on these, likely evolutionarily "older," germline insertions. In contrast, non-reference and rare TEs possibly represent more recent germline insertions. Finally, consistent with our previous data [39], singleton (so-

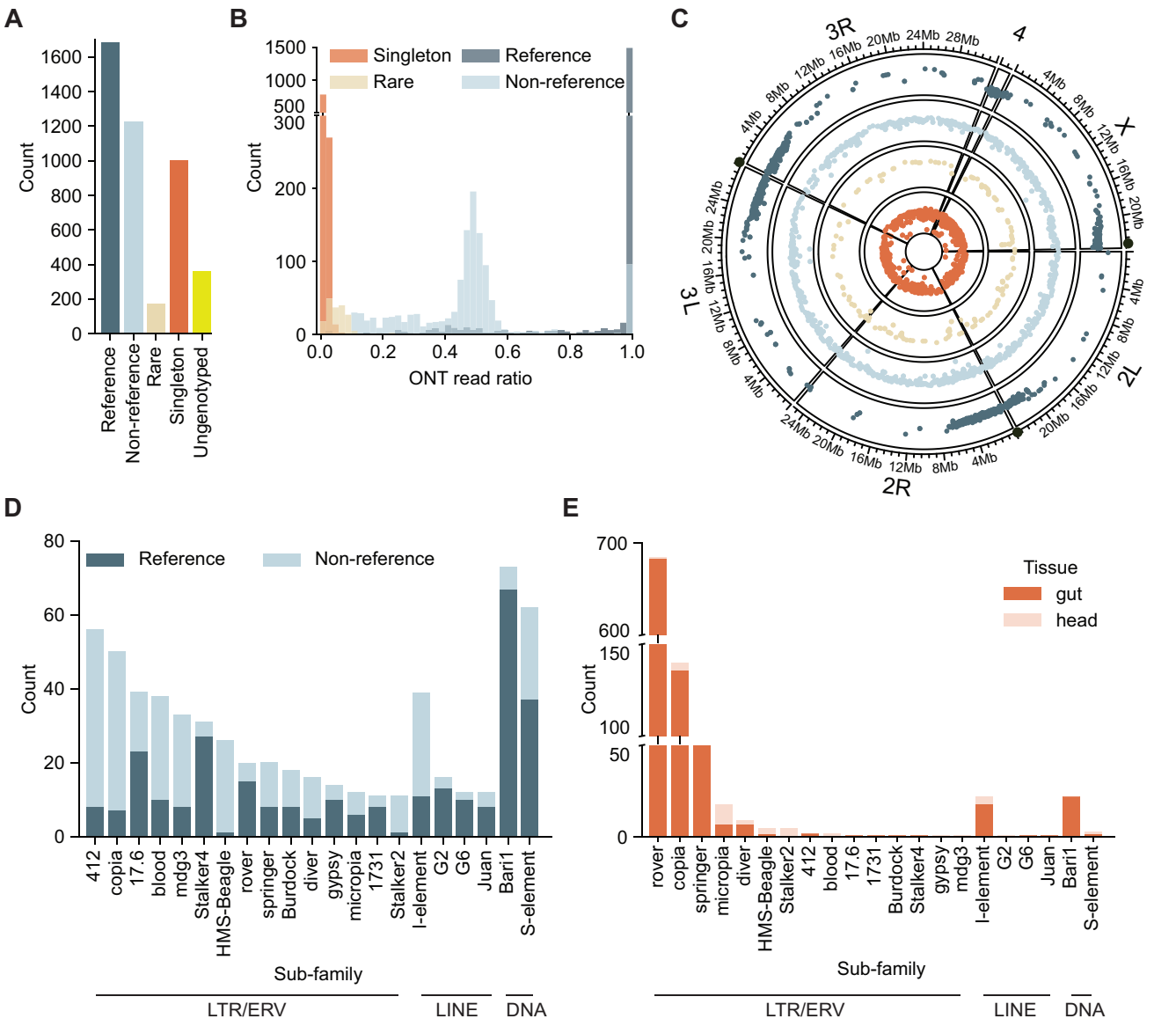


Figure 1. The landscape of full-length reference, non-reference, rare, and somatic TE insertions detected using long-read DNA sequencing. (**A**) Numbers of full-length TE insertions in the *ProsGFP* genome categorized in the "reference" (dark blue), "non-reference" (light blue), "rare" (light brown), and "singleton" (somatic, orange) genotypes, as well as the "ungenotyped" (yellow) insertions. (**B**) Distribution of ONT read ratios (number of supporting versus opposing reads) for TE insertions categorized as "reference" (dark blue), "non-reference" (light blue), "singleton" (somatic, orange), and "rare" (light brown). (**C**) Genome-wide distribution of the detected "reference" (dark blue), "non-reference" (light blue), "singleton" (orange), and "rare" (light brown) TE insertions on the *Drosophila* chromosome arms. Black circles indicate positions of centromeres. (**D**) Numbers of full-length "reference" (dark blue) and "non-reference" (light blue) insertions of the different TE sub-families. Only the sub-families that contributed somatic insertions are plotted. For all other TE sub-families, see Supplementary Fig. S1C. (**E**) Numbers of somatic insertions ("singletons") of different TE sub-families recovered from the gut (dark orange) or head (light orange) DNA libraries.

matic) TE insertions were distributed genome-wide, similarly to the non-reference and rare insertions.

Last but not least, we examined TE sub-families represented in each genotype. Full-length reference and non-reference insertions were recovered for 120 out of 126 known *Drosophila* TE sub-families, including Class I retrotransposons (the most abundant in terms of copy number and total sequence in the reference genome [38, 82]) and Class II DNA elements (Fig. 1D and Supplementary Fig. S1C). Rare insertions were represented by 34 sub-families (Supplementary Fig. S1D).

Consistent with what we reported previously and in contrast to the reference and non-reference insertions, only a few TE sub-families were represented among singleton insertions (Fig. 1E). A great majority of singletons (904 out of 1001) were retrotransposon insertions of three sub-families: ERV element *rover* (684 singletons), LTR element *copia* (140 singletons), and ERV element *springer* (80 singletons). We also recovered 22 singleton insertions of a non-LTR, LINE-like retrotransposon *I-element*, as well as 22 insertions of *Bari1*, a sub-family of DNA transposons. Notably, 965 (96%) of detected singletons were found in the libraries obtained from the intestines, while only 36 (4%) were head-specific insertions (Fig. 1E).

Taken together, the analysis of long-read DNA sequencing datasets enabled in-depth characterization of the full-length fixed TE insertions as well as detection of somatic events spe-

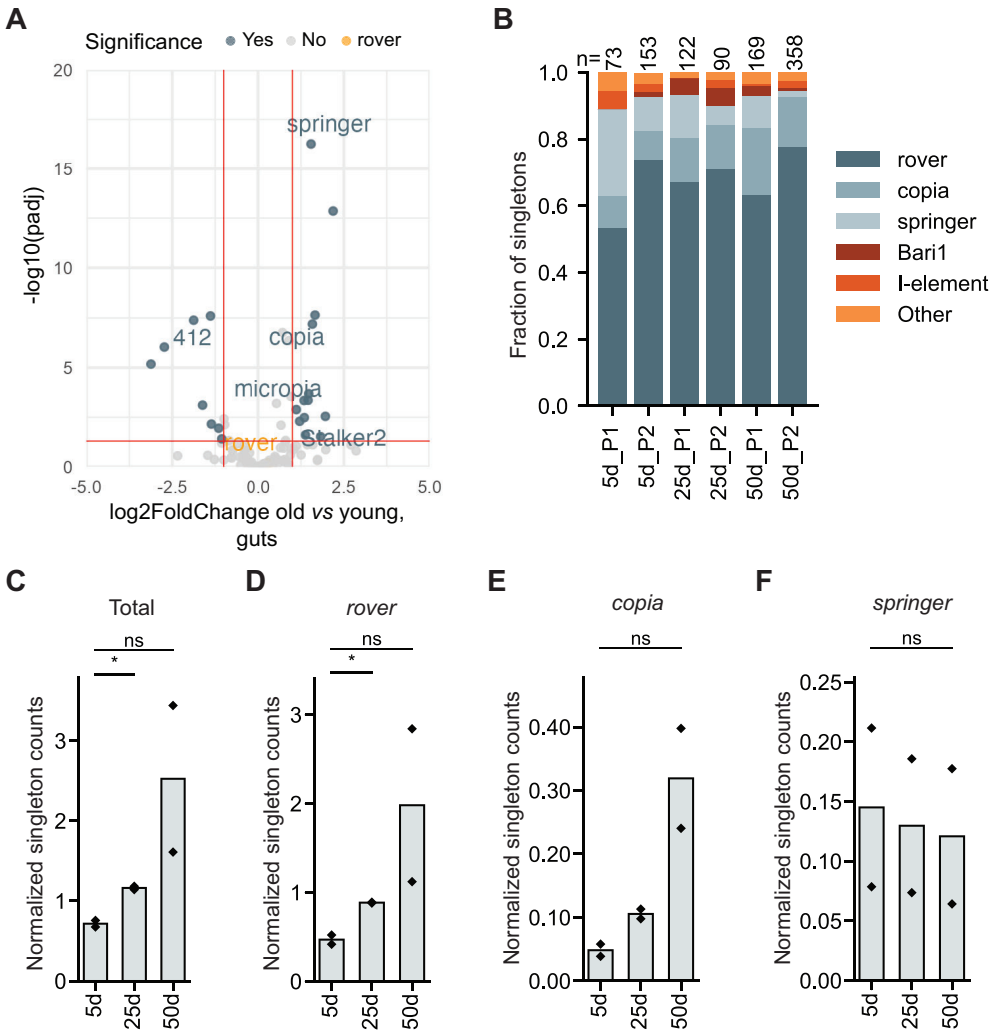


Figure 2. Somatic TE expression and mobility during aging. (**A**) Volcano plot illustrating changes in TE transcript levels upon aging in guts of the *ProsGFP* genotype. Data are from short-read RNA-seq. Significantly up- or downregulated TEs are indicated in blue and TE sub-families that were detected as mobile are labeled. (**B**) Sub-family distribution of somatic TE insertions detected in the gut DNA libraries from young (5 days), mid-aged (25 days), and old (50 days) flies. P1 and P2 are two biological replicates (pools 1 and 2). *n* = total number of "singletons" in each pool. Normalized counts of somatic insertions detected in libraries obtained from gut tissues of flies with different ages (5, 25, or 50 days). Plotted are all TE sub-families (**C**), as well as TE sub-families contributing the most *de novo* insertions: *rover* (**D**), *copia* (**E**), and *springer* (**F**). ns, not significant; * adjusted $P < 0.05$ (one-way ANOVA test).

cific for one tissue type and limited to a small number of somatically active TE sub-families. Since significant somatic mobility was detected only in the DNA sequencing libraries from the gut samples, we further focused on this tissue.

Aging is not associated with important increases in somatic *de novo* TE insertions

In different species, heterochromatin relaxation and transcriptional de-repression of DNA repeats, including TEs, are associated with aging [83–85]. Therefore, we next asked whether significant changes in somatic TE activity could be detected between young, mid-aged, or old tissues. To this end, we first carried out expression analysis using short-read RNA-seq data from young [39] as well as aged (this study) midguts to establish TE expression levels (Fig. 2A). In agreement with previously published reports from the fly intestine [86, 87], we detected increased transcript levels of selected TE sub-families in old guts, including retrotransposons *copia* and *springer*. Nevertheless, age-related upregulation in transcript levels was not widespread throughout TE sub-families (14 upregulated TE sub-families) and a comparable number of TEs were also significantly downregulated (8 TE sub-families). Out of the three LTR/ERV sub-families identified as the most mobile (Fig. 1E), *copia* and *springer* showed increased transcript levels, while *rover* transcripts were not significantly changed. Moreover, neither the expression levels nor the extent of the age-related deregulation correlated with the overall TE mobility (Supplementary Fig. S2A–C).

Since we have previously shown that TE transcript levels reflect poorly the actual TE mobility [39], we next focused on comparing somatic TE *de novo* insertions between the age groups. The contribution of TE sub-families to the total number of singleton insertions in each library was not considerably different between libraries from young, mid-aged, and old guts, with *rover*, *copia*, and *springer* LTR/ERV-type elements dominating in all samples (Fig. 2B). ONT sequencing libraries can significantly vary in terms of yield and read lengths, making direct comparison of the raw singleton counts between libraries difficult. Thus, we normalized the raw singleton counts

for sequencing depth, considering read length distributions in each library. By doing so, we observed a moderate increase with age in the total, *rover*, and *Bari1* normalized singleton counts (Fig. 2C and E, and Supplementary Fig. S2E). The significant change in the total and the *rover* counts could be detected between samples from young and mid-aged flies, but not between young and old individuals, likely stemming from the high variability between the two replicates in the latter. No significant changes between age groups were detected when singletons of *copia*, *springer*, or *I-element* sub-families were counted (Fig. 2D and F, and Supplementary Fig. S2D). Thus, only minor changes in somatic TE insertion loads could be detected between DNA samples from young and aged guts.

We then sought to determine whether similar results were obtained for the fly head samples. In the head transcriptomes, expression of 18 TE sub-families was upregulated in aged tissues, and 3 sub-families were downregulated (Supplementary Fig. S2F). Among the significantly deregulated TEs, *copia* (upregulated) and *mdg3* (downregulated) were mobile in the head tissue according to our analysis (Supplementary Fig. S2G). However, the association of the transcriptional deregulation and the mobility could not be reliably tested owing to the low numbers of the detected singletons. Thus, expression of few TE sub-families was changed in aged fly heads of the investigated genetic background, but changes in the number of the detected *de novo* TE insertions were marginal in all age groups.

In conclusion, even though transcript levels change for selected TE sub-families in the aging fly gut and head tissues, we could not observe important age-related triggering of TE mobility in terms of both insertion numbers and active TE sub-families.

Mobility and expression of the *rover*-LTR/ERV sub-family are restricted to one donor locus

To gain insights into the mechanisms of retrotransposon activation in the soma, we then sought to identify potential donor retrotransposon loci responsible for the observed somatic mobility. We focused our analysis on the *rover-ERV* sub-family that accounted for almost 70% of all somatic insertions. Taking advantage of ONT reads spanning full insertion length between genomic breakpoints, we extracted sequences of *rover* singletons, representing somatic insertions, and aligned them to the *rover* consensus sequence (Fig. 3A). The genome of the *ProsGFP* strain contains 15 nearly full-length reference and 5 non-reference *rover* copies. We created consensus sequences for the reference and non-reference *rover* insertions using ONT reads and aligned them to the published *rover* consensus sequence (https://github.com/bergmanlab/drosophila-transposons). Visual inspection of the alignments of the germline and singleton insertions showed that 1 non-reference and 11 reference insertions had large SVs that were not present in the singleton insertions (Supplementary Fig. S3), leaving 8 full-length potential candidates to be the donor locus (Fig. 3A and B). Only two out of the eight candidate loci contained the small SVs present in the singletons. Moreover, almost all singletons showed support for the single nucleotide polymorphisms (SNPs) present in only one of these two *rover* loci (96% singletons on average per SNP, Supplementary Table S3). The discrepancy from 100% was within ONT sequencing error rate and none of the singletons had the full set of SNPs that supported the second locus, suggesting that the donor locus is on second chromosome at position chr2R:14487730–14487747 (Fig. 3B). This locus is in the anti-sense orientation within the first intron of the *PRAS40* gene (Fig. 3C) and is not present in the dm6 reference genome. Thus, we concluded that the observed somatic mobility of the *rover-LTR/ERV* sub-family in the intestinal tissue is limited to one donor locus, here named *rover-2R:14M*.

As the first prerequisite for retrotransposon mobility is its transcription, we next aimed to confirm the expression of the *rover-2R:14M* and check whether other, non-mobile, *rover* copies were also expressed in the tissue. Aligning the short transcriptomic reads to the *rover* consensus sequence confirmed *rover* expression in the gut tissue, regardless of the age of the flies, while minimal expression was detected in the head samples (Fig. 3C and E). Additionally, we did not detect *rover* transcripts in the gut samples of the w^{1118} genetic background (commonly used wild-type stock), which did not carry the *rover-2R:14M* copy (Fig. 3E). We then attempted to quantify per-copy expression (Fig. 3F and G). To address sequence similarity among *rover* copies, we extracted the reads that aligned to the *rover* consensus sequence and re-aligned them with stringent parameters to the consensus sequences of all reference and non-reference *rover* copies in the *ProsGFP* genome, discarding multimapping reads. This approach, which did not provide absolute per-copy expression levels but instead showed expression levels relative to other copies, revealed that the *rover-2R:14M* locus was the only full-length locus of the *rover* sub-family expressed in the gut, consistent with our evidence above from *rover* mobility. Finally, in order to further validate this result, we additionally performed long-read ONT cDNA sequencing on midguts isolated from mid-aged flies, because short-read sequencing has important limitations in the RNA-seq analysis of repeat sequences [21] (Fig. 3H and I). Consistent with the short-read datasets, we saw evidence for a predominant expression of the donor *rover-2R:14M* locus (Fig. 3I), suggesting that other *rover* loci were not or very lowly expressed in the gut tissue.

We then asked how *rover-2R:14M* expression compared to the expression of the *PRAS40* gene in which *rover-2R:14M* is embedded. Importantly, *rover-2R:14M* expression did not fully correlate with that of *PRAS40*: *PRAS40* was found to be expressed in both gut and head tissues (Fig. 3D) in contrast to *rover-2R:14M* expressed in the gut, suggesting distinct transcriptional regulation. In addition, the presence or absence of *rover-2R:14M* within *PRAS40* did not alter *PRAS40* expression patterns in either the head or the gut, as seen by comparing the *ProsGFP* background (*rover-2R:14M* positive) to the w^{1118} background (*rover-2R:14M* negative) (Fig. 3C and D). Moreover, the *rover-2R:14M* insertion did not cause mis-splicing of the *PRAS40* gene, as shown by the ONT cDNA sequencing data (Supplementary Fig. S4A). However, we observed a single transcript in the ONT cDNA sequencing data, 5′ end of which was located in the *rover-2R:14M* 3′LTR and which contained the downstream *PRAS40* intronic sequence. This indicated that *rover-2R:14M* locus could potentially enable weak aberrant transcription in the gut tissue (Supplementary Fig. S4A′). Together, this analysis hinted that the *rover-2R:14M* donor locus is expressed in the gut without affecting the transcription of the gene it is embedded in.

The control of TE transcript levels in the fly non-gonadal tissues is mostly achieved by the endogenous short interfering (siRNA) pathway, responsible for sequence-specific

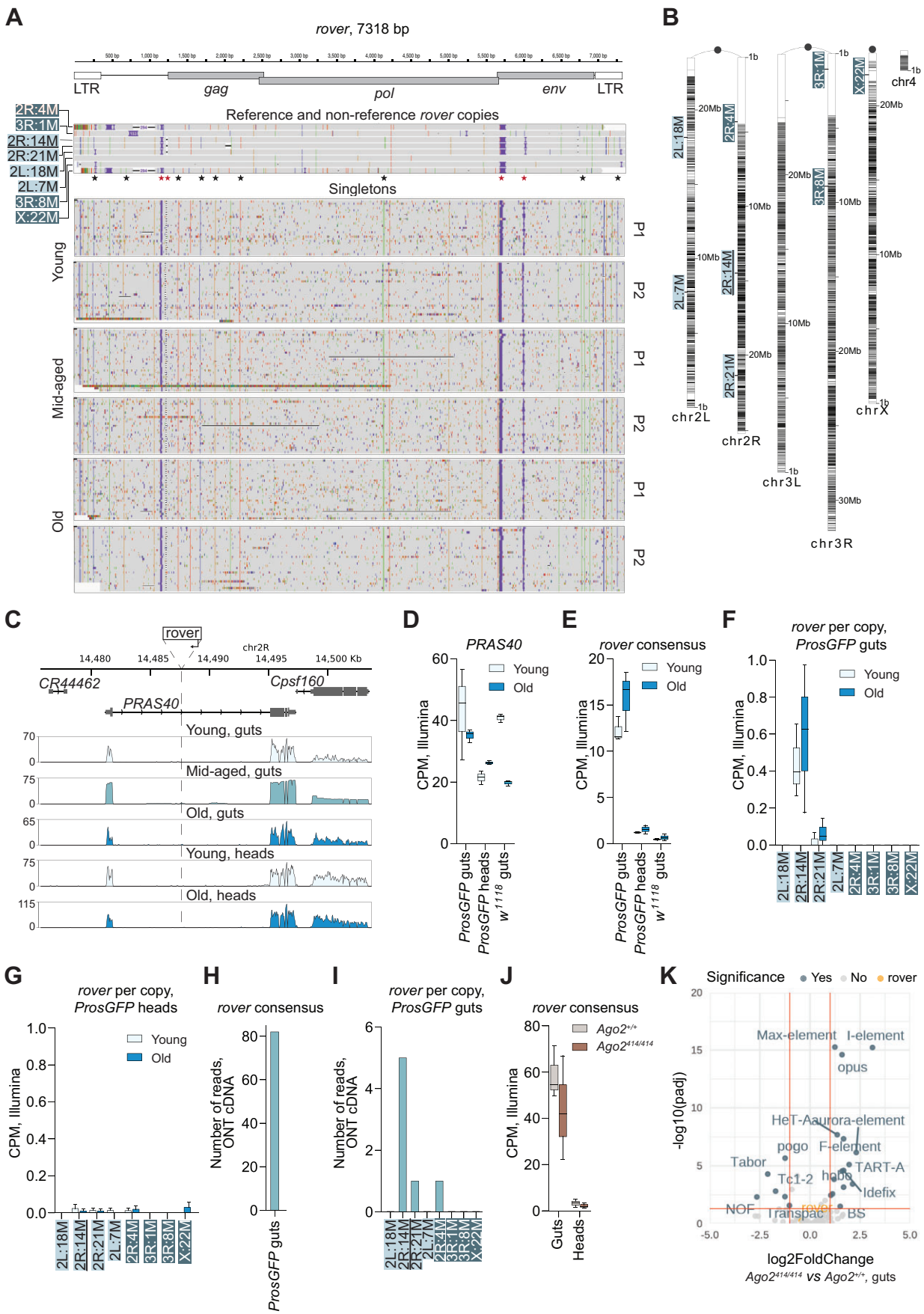


Figure 3. Somatic activity of the *rover-LTR/ERV* family is limited to one donor locus. (**A**) IGV visualization of the reference and non-reference *rover* copies present in the analyzed genome and a random selection of somatic (singleton) *rover* insertions detected in different long-read DNA-seq libraries

transcript degradation [88–90]. We have previously demonstrated that *rover* mobility in the *ProsGFP* flies occurs despite the presence of siRNAs against *rover* in the gut [39], suggesting that post-transcriptional silencing of *rover* may not be efficient in this genetic background. To further address whether the identified *rover-2R:14M* locus is regulated by the siRNA pathway, we then performed transcriptome analysis in tissues isolated from flies carrying the *rover-2R:14M* locus and mutant for *Argonaute 2* (*Ago2*), deficient in the siRNA pathway, as well as their isogenic controls (Fig. 3J and K). In both genetic backgrounds, *rover* transcripts were consistently detected only in the gut RNA transcriptomes and not in the heads. Importantly, *rover* transcript levels were not changed in the Ago2 homozygous mutant tissues (Fig. 3J), even though we detected differential expression of other TE sub-families in the gut and the head transcriptomes (Fig. 3K and Supplementary Fig. S5), as previously reported in *Ago2* mutant tissues [88–90]. Altogether, these data suggest that the identified donor *rover-2R:14M* locus expressed in the fly gut is not post-transcriptionally regulated by the siRNA pathway.

Interestingly, in a concomitant independent study using a different fly genetic background for the detection of *de novo* TE insertions, activity of the *rover* sub-family was also reported and classified as likely somatic [91]. Using their data, we have performed the same type of sequence comparison of the germline non-reference *rover* copies and the *de novo* singleton insertions from that study, and found that the *rover-2R:14M* insertion was also present in the investigated genetic background and likely acted as a source locus for the great majority of *de novo* somatic insertions (Supplementary Fig. S6).

Thus, we identified a non-reference "hot" *rover-2R:14M* LTR/ERV retroelement locus, transcribed in the fly intestinal tissue and serving as a donor locus for *de novo* somatic insertions.

The donor *rover-2R:14M LTR/ERV* locus is located in permissive chromatin

We then aimed to understand which mechanisms could underlie the somatic activity of the *rover-2R:14M* locus. Since LTR/ERV retrotransposons are believed to carry their own regulatory sequences within the LTR sequence itself and the larger 5′UTR region proceeding the LTR [40–45], we first asked whether the transcriptional activity of the *rover-2R:14M* locus could be explained by sequence variation private to the locus within the TE sequence. We analyzed the 5′ LTR sequences, sites of transcriptional initiation of LTR-type elements, of the donor locus and two non-active *rover* loci (*2R:21M* and *2L:18*), which had the highest sequence similarity to the donor locus. We first assessed the presence of the sequence motifs for the PolII transcription initiation complex, namely TATA-box, Inr, DPE, and MTE, as LTR elements are believed to be PolII transcribed [92] (Fig. 4A). We did not notice significant differences between the full-length *rover* copies that could explain gut-specific expression. We then extended the analysis to an *in silico* analysis of putative TF binding sites within the first 2 kb of the internal *rover* sequences (Fig. 4B and Supplementary Fig. S7). Again, we were unable to find motifs that were different between the active *rover-2R:14M* locus and non-expressed loci, implying that internal sequence variation could not account for the differences in the activity of the compared *rover* loci.

We thus reasoned that *rover-2R:14M* transcriptional activity (and mobility) might be conferred by its genomic position rather than its internal sequence. Given that, we inspected chromatin environment upstream (10 kb) of the active and seven inactive full-length *rover* loci present in the investigated genome, making use of the published gut cell type-specific DamID datasets profiling five chromatin binding factors associated with "active" or "silent" chromatin states [76, 93] (Fig. 4C). Only two *rover* loci (*2R:21M* and *2R:14M*) were located in "active" regions bound by PolII and Brm, with the upstream region of the donor *rover-2R:14M* locus showing the strongest binding in all intestinal cell types. The upstream regions of the remaining loci were either not bound at all or bound by the heterochromatin factors HP1, H1, or Pc. Similarly, when the same regions were inspected using published ATAC-seq data [76], the upstream region of the donor *rover-2R:14M* locus showed the highest accessibility signal (Fig. 4D), thus the most "open" chromatin environment.

Altogether, this analysis indicated that the epigenomic environment, rather than copy-specific sequence variants, is likely responsible for the *rover-2R:14M* activity in the gut tissue.

rover-2R:14M expression is driven by its upstream genomic sequence

To experimentally test the transcriptional regulation of the active *rover* copy, we engineered expression reporters where the *rover-2R:14M* 5′UTR and upstream genomic sequences were placed in front of a *lacZ.NLS* (NLS, nuclear localization signal) reporter gene (Fig. 5A). We inserted the reporters into

aligned to the published *rover* consensus sequence. The reference and non-reference copies (covered by many reads) are named following their position in the genome and each line represents a consensus sequence of the copy. For somatic insertions, each line represents a single read covering an insertion. Regions with SVs and SNPs that distinguish the donor *rover-2R:14M* copy from other *rover* copies are highlighted with red and black stars, respectively. (**B**) Schematic representation of the *Drosophila* chromosomes with the positions of reference and non-reference *rover* copies in the *ProsGFP* genome. (**C**) Position of the donor *rover-2R:14M* locus on the chromosome 2R within the intron of the *PRAS40* gene (upper panel). RNA-seq coverage of the *PRAS40* locus showing expression of the gene in libraries obtained from gut or head tissues from flies of different ages (lower panels). Young and old gut, and young and old head data are from short-read RNA-seq. Mid-aged gut data are from ONT long-read cDNA sequencing. Normalized expression levels (CPM, counts per million) in gut and head transcriptomes of the *ProsGFP* genotype (with the *rover-2R:14M* insertion) and in gut transcriptomes of the w^{1118} control genotype (without the *rover-2R:14M* insertion). Quantified are transcript levels of *PRAS40* (**D**) and of *rover* upon read mapping to the consensus sequence (**E**). Data are from short-read RNA-seq. Normalized expression levels (CPM, counts per million) of *rover* upon copy-specific read mapping in transcriptomes from guts (**F**) and heads (**G**) of the *ProsGFP* strain. Data are from short-read RNA-seq. Number of long reads mapping to the *rover* consensus (**H**) and to the different *rover* copies present in the genome (**I**) in long-read transcriptomes from the *ProsGFP* background. Data are from ONT cDNA sequencing. (**J**) Normalized *rover* expression levels (CPM, counts per million) in gut and head transcriptomes of the *Ago2* mutant (*rover-2R:14M*$^{+/+}$;*Ago2*$^{414/414}$) and the *rover-2R:14M*$^{+/+}$;*Ago2*$^{+/+}$ control flies. Reads were mapped to the *rover* consensus sequence. (**K**) Volcano plot illustrating changes in TE transcript levels in *Ago2* mutant (*rover-2R:14M+/+;Ago2*$^{414/414}$) guts as compared to *rover-2R:14M*$^{+/+}$;*Ago2*$^{+/+}$ control tissues. Both genotypes carry the *rover-2R:14M* active copy. Significantly up- or downregulated TE sub-families are indicated in blue and labeled. In panels A, B, F and G, reference copies are marked with dark blue and white lettering, non-reference copies with light blue and the identified donor *rover-2R:14M* copy is underlined.

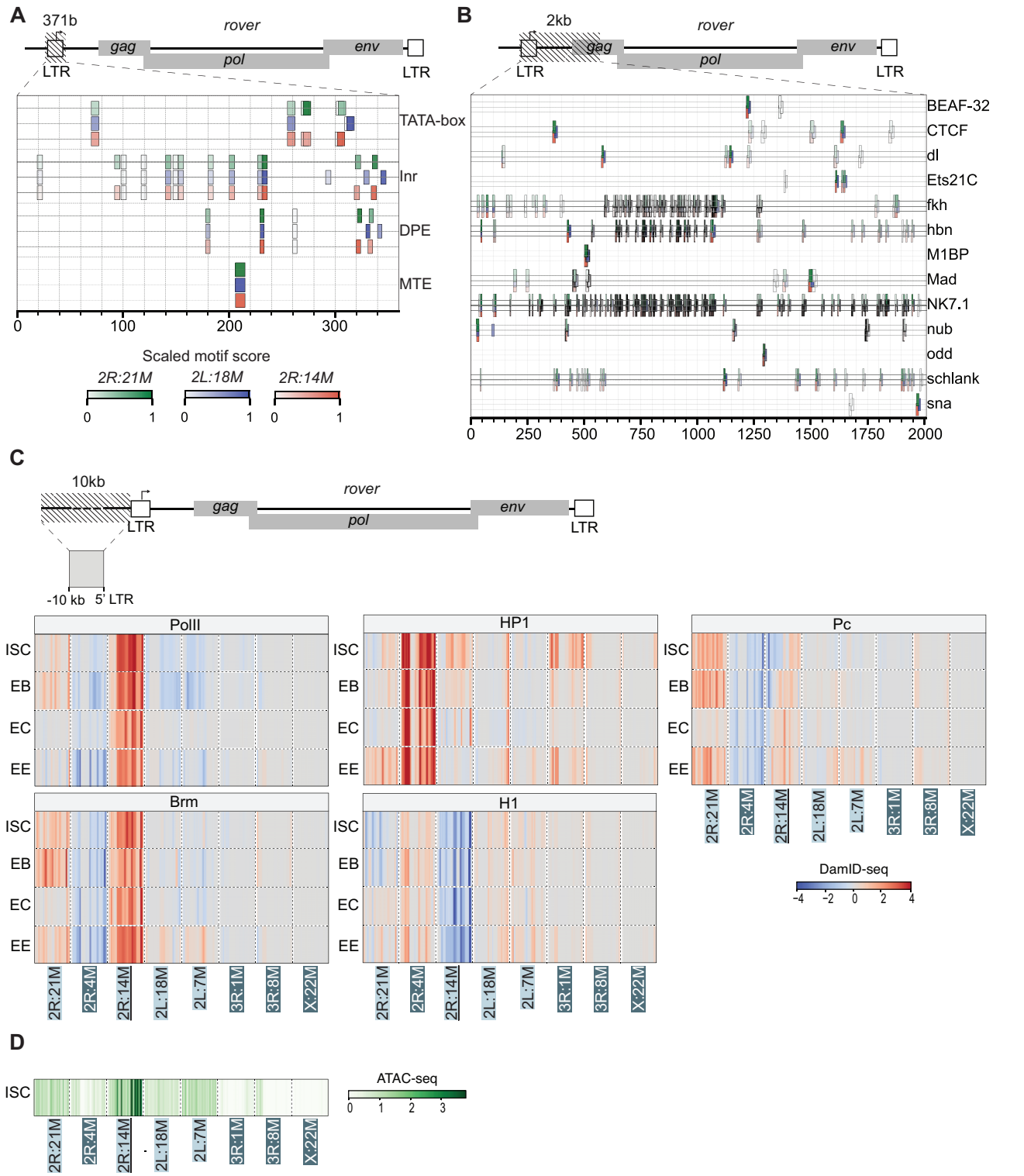


Figure 4. The donor *rover-2R:14M* locus is located in permissive chromatin. (**A**) *In silico* detection of DNA motifs for PolII transcription initiation complex (TATA-box, Inr, DPE, and MTE) in the 5′LTR region of the donor *rover-2R:14M* copy (red) and two non-active copies with the highest sequence similarity (*2R:21M*, green, and *2L:18*, blue). (**B**) *In silico* search for TF binding sites within the first 2 kb of the internal sequences of the donor *rover-2R:14M* copy (red) and two non-active copies with the highest sequence similarity (*2R:21M*, green, and *2L:18*, blue). Only TFs highly expressed (RPKM > 10) in the gut lineage are plotted (for all TFs, see Supplementary Fig. S7). (**C**) Gut cell type-specific heatmaps of DamID (DNA adenine methyltransferase identification) chromatin profiles upstream (10 kb) of the active *rover-2R:14M* copy and seven inactive full-length *rover* loci present in the investigated genome. Five chromatin binding factors are shown: PolII, Brahma (Brm), Polycomb (Pc), heterochromatin protein 1a (HP1), and histone H1 (H1). Intestinal stem cells (ISCs) and enteroblasts (EBs) are gut progenitor cells. Enterocytes (ECs) and enteroendocrine cells (EEs) are differentiated cell types. *Rover* reference copies are marked with dark blue and white lettering, and non-reference copies with light blue. The identified donor *rover-2R:14M* copy is underlined. (**D**) Heatmaps of ISC ATAC-seq profiles upstream (10 kb) of the active *rover-2R:14M* copy and seven inactive full-length *rover* loci present in the investigated genome. Reference copies are marked with dark blue and white lettering, and non-reference copies with light blue. The identified donor *rover-2R:14M* copy is underlined.

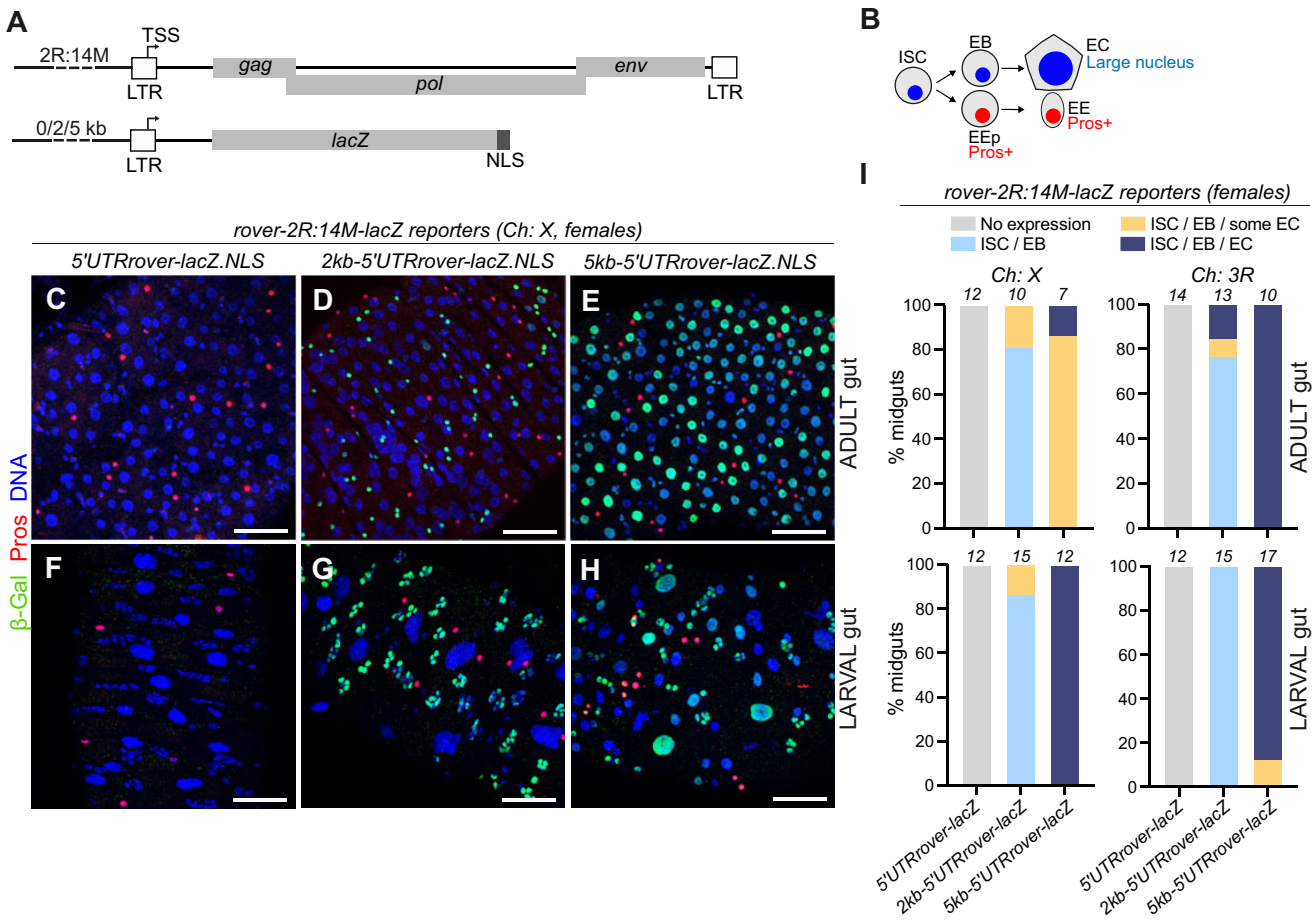


Figure 5. Upstream genomic sequence drives *rover-2R:14M* expression in the gut. (**A**) Schematic representation of the *in vivo rover-2R:14M* reporter constructs. lacZ reporter was placed downstream of the *rover-2R:14M* 5′UTR region replacing all *rover* native ORFs. 0, 2, or 5 kb of the upstream genomic sequence was also added. (**B**) Schematic representation of the *Drosophila* intestinal lineage. ISCs give rise to committed progenitors: EBs and enteroendocrine precursors (EEp). The progenitors then differentiate into ECs and enteroendocrine cells (EEs). The EEp and EE express Prospero (Pros). The ECs are distinguished by a large, polyploid nucleus. In the larval gut, the stem cells, named adult midgut precursors (AMPs), form distinct clusters. (**C**–**H**) Representative images of female midguts carrying *rover-2R:14M-lacZ* reporter constructs inserted on the X chromosome stained for β-galactosidase (β-Gal, green), Prospero (Pros, red), and DNA (blue). Adult (C–E) and larval (F–H) tissues are shown. Scale bar: 50 μm (**I**) Quantifications of the expression patterns of *rover-2R:14M-lacZ* reporter constructs in female adult and larval midguts. Numbers of midguts scored are indicated above each bar.

the fly genome using two different landing sites on the X and the third chromosomes. The sites were chosen based on their similarity in the DamID chromatin landscapes to the original *rover-2R:14M* genomic location ("active" chromatin in all gut cell types; Fig. 4C and Supplementary Fig. S8).

To begin to define the patterns or the reporter activity *in vivo*, we first tested *lacZ* expression in the female gonads, where similar reporters for other retrotransposon families were already used [94–96]. As previously shown, in ovaries with functional small RNA-driven TE silencing, we detected no or very weak expression of *lacZ* under the control of the 5′UTR of *rover-2R:14M* (Supplementary Fig. S9A, B′, E, and F′). However, *lacZ* expression was readily detected upon knockdown of the piRNA pathway with germline- (Supplementary Fig. S9C and D′) or somatic follicle cell-specific drivers (Supplementary Fig. S9G and H′). This confirmed that, as expected, in female reproductive tissues, the *rover-2R:14M* reporter can be efficiently transcribed, but is silenced by the piRNA pathway.

To gain insight into *rover-2R:14M* expression in the intestine, we then examined the lacZ reporter expression in this tissue in the adult stage, as well as during development, in the third instar larval stage. The adult and the larval gut consist of progenitors and differentiated cell types, which can be distinguished by the expression of cell type-specific markers and by the nuclear size (Fig. 5B). *LacZ* was not expressed in the gut when driven only by the 5′UTR region of *rover-2R:14M* (Fig. 5C, F, and I). However, by extending the regulatory region with 2 or 5 kb of the upstream genomic sequence, we could achieve reporter expression in adult (Fig. 5D, E, and I) and larval intestinal tissues (Fig. 5G, H, and I), even without interfering with the TE silencing pathways. Patterns of reporter expression were different depending on the length of the upstream regulatory sequence. Two kilobases of the upstream region drove *rover-2R:14M* reporter expression mostly in gut progenitor cells: stem cells (ISCs) and EBs (Fig. 5D, G, and I, unmarked small diploid nuclei). Increasing the length of the sequence to 5 kb allowed for broader reporter expression patterns in progenitors as well as differentiated absorptive ECs (large polyploid nuclei) (Fig. 5E, H, and I). We did not detect reporter activity in the secretory EE cells, marked by the expression of the Prospero (Pros) TF. These expression patterns were largely consistent between females and males (Supplementary Fig. S10). Notably, we obtained similar re-

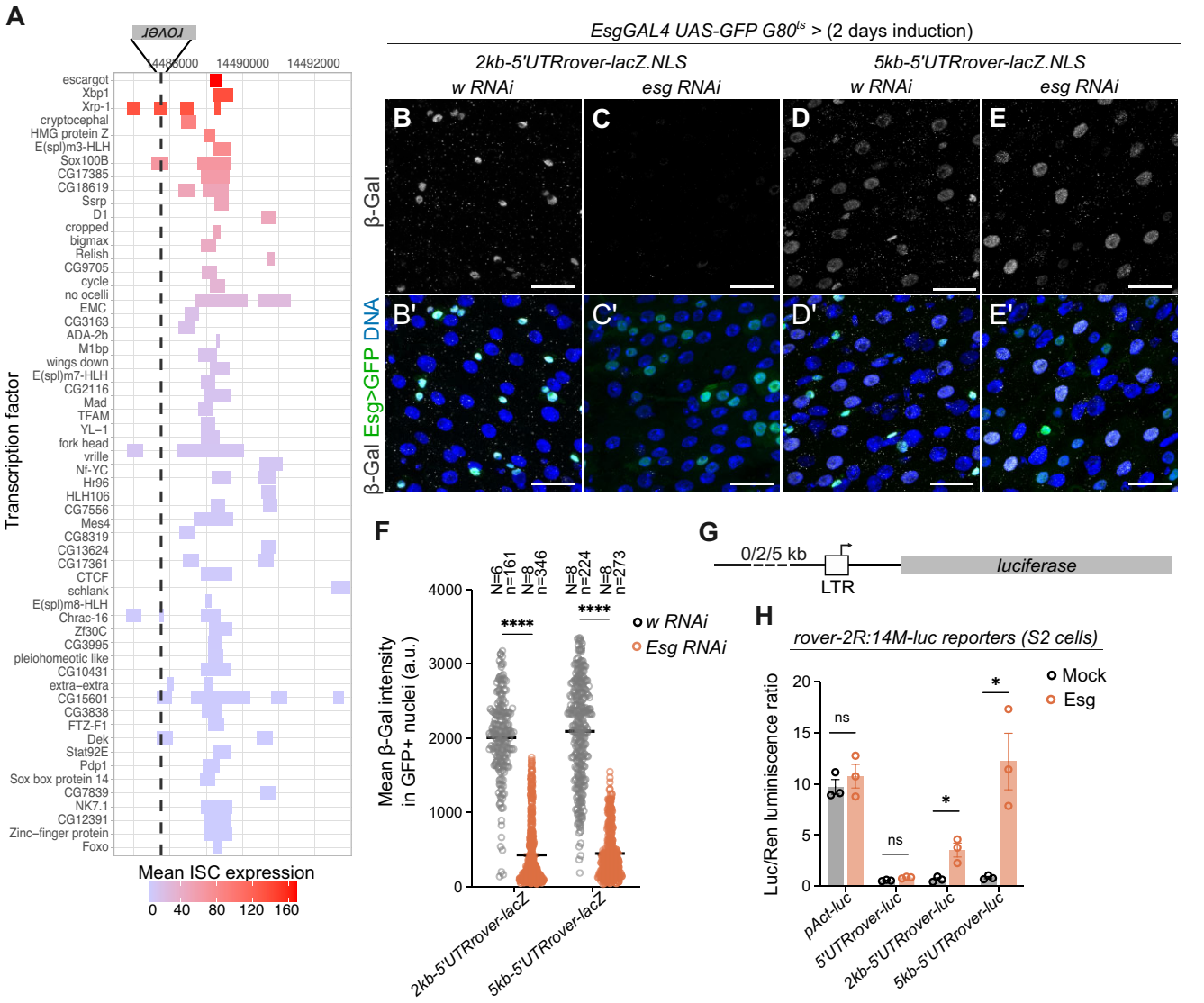


Figure 6. Esg TF drives *rover-2R:14M* expression via the upstream genomic sequence. (**A**) TF binding sites identified bound to the genomic region upstream of the *rover-2R:14M* locus based on the published ChiP-seq datasets (modENCODE [67]). TFs were ordered based on mean expression levels in ISCs obtained from [68]. (**B**–**E**′) Representative confocal images of female midguts with *rover-2R:14M-lacZ* reporter constructs inserted on the X chromosome stained for β-galactosidase (β-Gal, gray), GFP (green), and DNA (blue). White RNAi (w, control, B, B′, D, and D′) or Esg RNAi (C, C′, E, and E′) was induced only in progenitor cells marked with GFP (*EsgGAL4 UAS-GFP G80^{ts}*) for 2 days. Scale bar: 25 μm. (**F**) Quantification of β-galactosidase signal intensities in GFP-positive progenitor cells upon *white* or *esg* knockdown in guts carrying the *rover-2R:14M-lacZ* reporter constructs inserted on the X chromosome. *N*: number of guts scored; *n*: number of GFP+ nuclei scored. ****$P < .0001$ (two-tailed Mann–Whitney test). (**G**) Schematic representation of the *rover-2R:14M* cell culture reporter constructs. Luciferase reporter was placed downstream of the *rover-2R:14M* 5′UTR region replacing all *rover* native ORFs. 0, 2, or 5 kb of the upstream genomic sequence was also added. (**H**) Luciferase assays for transcriptional activation in S2 cells. Luciferase to renilla luminescence ratios are plotted. *pAct-luc* represents a luciferase reporter under control of a constitutive promoter. All constructs were tested in the presence and absence of the Esg TF. ns, non-significant; * $P < .05$ (two-tailed Mann–Whitney tests). The graph is representative of three independent experiments, with three experimental replicates each.

sults with both insertion sites tested (Fig. 5I), suggesting that, at least for the two insertion sites tested by us, reporter expression was not influenced by the genomic position.

Taken together, these results support the idea that *rover-2R:14M* expression in the fly gut may be driven in *cis* by its upstream genomic sequence.

Escargot transcription factor regulates *rover-2R:14M* expression

To better understand how the upstream genomic region could allow for *rover-2R:14M* expression in the gut, we next analyzed published chromatin immunoprecipitation (ChiP-seq) data (ModEncode Project [75]) to identify TF binding sites present in the region (Fig. 6A). Indeed, many TFs were found to bind to the region, particularly within the first 2 kb upstream of the *rover-2R:14M* insertion. We sorted the TFs based on their expression in the adult gut ISCs (Fig. 6A), EBs, or differentiated ECs (Supplementary Fig. S11) using previously published transcriptomic data [74]. Based on this analysis, we identified a set of candidate factors that could regulate *rover-2R:14M* expression through the upstream genomic sequence. These included TFs with established gut function, such as Escargot (Esg), Sox100B, or Fork head.

For further analysis, we decided to focus on Esg, since the observed patterns of *rover-2R:14M-lacZ* reporter activity were reminiscent of *esg* expression, enriched in progenitor cells (ISCs and EBs), and expressed at lower levels in differentiated ECs [74]. In flies carrying the *rover-2R:14M-lacZ* reporters, we performed a transient, progenitor-specific knockdown of Esg using a temperature-sensitive GAL4-UAS system (*Esg>GAL4 UAS-GFP GAL80^{ts}*) [48] (Fig. 6B–E′). Control GFP-marked progenitor cells (expressing *white* RNAi) were positive for β-Gal staining (Fig. 6B, B′, D, and D′); however, we observed marked decrease in β-Gal staining in GFP-positive cells upon *esg* knockdown. This effect was observed for both the *2kb-5′UTRrover*-lacZ (Fig. 6C, C′, and F) and the *5kb-5′UTRrover*-lacZ (Fig. 6E, E′, and F) reporters, suggesting that Esg could positively regulate *rover* reporter expression *in vivo*.

Importantly, in the fly gut, Esg maintains the progenitor cell fate and its depletion deregulates many genes, switching on the EC differentiation markers and, in longer term, leading to stem cell loss [97, 98]. Indeed, visibly enlarged nuclei of GFP-positive cells upon *esg* knockdown indicated ongoing loss of the progenitor fate in these cells. Thus, we cannot exclude that the decrease in the *lacZ* reporter expression was a secondary consequence of deregulation of other genes than *esg* itself. Hence, in order to directly test the involvement of Esg in the *rover-2R:14M* transcriptional regulation, we made use of *in vitro* luciferase reporter assays in *Drosophila* S2 cells (Fig. 6G). Transient transfection of the luciferase reporters under control of the *rover-2R:14M* 5′UTR alone or with addition of the upstream sequences (2 or 5 kb) did not result in detectable luciferase activity, as compared to the positive control with a constitutive actin promoter (*pAct-luc*) (Fig. 6H). However, strikingly, co-expression with Esg led to a significant increase in luciferase activity of *rover-luc* reporters. The transcriptional upregulation was observed only in the presence of the genomic sequence (2 or 5 kb) upstream of the *rover-2R:14M*, suggesting that the sequence is required for reporter activation and that Esg acts directly or indirectly on the sequence, regulating reporter expression.

Together, these results, obtained with the *in vivo* and cell culture *rover* expression reporters, support the idea that the active *rover-2R:14M* LTR/ERV retrotransposon locus may be regulated by Esg, a TF critical for cell fate regulation in the fly intestine. This regulation requires the genomic sequence present upstream of the *rover-2R:14M* locus.

Discussion

Although expression and mobility of retroelements are increasingly reported in diverse organisms and somatic contexts, the regulation of such activity is not well understood. Here, we provide new insight into somatic activity of endogenous LTR/ERV retrotransposons.

First, through whole genome long-read DNA sequencing of *Drosophila* tissues, we deliver a detailed landscape of germline and somatic TE insertions in the analyzed genome. As previously reported by us [39], as well as by others [38], our analysis highlights the important and often unappreciated variability in TE composition between fly strains. Indeed, about half of all full-length TE insertions mapped by us are not found in the reference genome. Notably, in contrast to reference TEs, many non-reference TEs are found in euchromatic regions, which, as we show here, may contribute to their transcriptional activity or increase the chance that such insertions could influence expression of neighboring genes. This emphasizes the frequently overlooked importance of carefully characterizing or standardizing genetic backgrounds, particularly when somatic TE expression or mobility is studied and compared in different contexts.

Second, using our datasets, we address a commonly presumed idea that TE activity is unleashed upon aging [83, 85]. While we report changes in TE transcript levels in the fly gut and head as seen previously [86, 87, 99, 100], we detect no striking increases in *de novo* somatic TE insertions in these tissues. This result, while contradictory to previous, reporter-based data [86, 99, 101, 102], is consistent with recent sequencing-based studies [100, 103–105]. We cannot exclude that different results could be obtained when systemically analyzing tissues isolated from flies with other genetic backgrounds. Nonetheless, to date, evidence for age-related unleashed mobility of endogenous TEs in the *Drosophila* soma remains scarce. In contrast, we detected many *de novo* retrotransposon insertions in guts isolated from young flies. We cannot eliminate the possibility that these insertions arose during the first few days after eclosion, but the presence of somatic insertions in young adults suggests that retrotransposon mobility could have occurred in pre-adult stages and continued to arise in the adult tissues. This is also consistent with the fact that *rover-2R:14M-lacZ* reporter expression was readily detected in larval as well as adult intestines. Furthermore, the potential developmental TE activity observed by us goes along with recently reported *mdg4* retrotransposon activation in the *Drosophila* hindgut during metamorphosis, proposed to contribute to anti-viral immunity [9]. It is also interesting to note that, although somatic transposition has mostly been studied in the context of the mammalian neuronal lineages [106], relatively high levels of somatic L1 retrotransposition have also been reported in the human intestinal lineage [34, 107, 108], similarly to our results. This raises a question of whether the intestinal tissue is in general predisposed to retrotransposition, and if so, what features enable it. Since the intestine is self-renewing in both humans and *Drosophila* [109], a plausible hypothesis could be that the mechanism of retrotransposition of LTR/ERV elements is coordinated with the cell cycle, similar to that of the L1 elements [110, 111]. However, other tissue- or cell type-specific levels of regulation may also play a role.

Further, we deliver new insight into transcriptional regulation of retrotransposons in the fly gut. Through our analysis of *de novo* somatic insertions with long-read DNA sequencing, we identify a "hot" LTR/ERV locus, *rover-2R:14M*. This example of a somatically mobile LTR/ERV retrotransposon adds to the previously characterized cases of mobile non-LTR L1 elements in mammalian contexts [26, 31, 32, 35, 112]. Hence, we were able to investigate copy-specific LTR-element regulation, in line with a growing body of evidence demonstrating that examining TE activity necessitates locus-specific approaches [24, 25, 58, 113], the same way as genes are studied.

To date, studies on the regulation of LTR/ERV elements have been focusing on the sequences carried by the retrotransposons themselves [16–19, 40, 41, 43–45]. In contrast, we revealed that *rover-2R:14M* is located in permissive chromatin environment, within an intron of an expressed gene, and its transcription is regulated in *cis* by the genomic sequence upstream to the *rover-2R:14M* sequence. Interestingly, in a recent

preprint concomitant with our work, Glaser *et al*. characterize a polymorphic mouse LTR retrotransposon, MusD, which achieves its expression during limb development by adopting the expression pattern of neighboring genes contained within the same regulatory domain [114]. Thus, this study, along with ours, provides complementary example of how LTR retroelement activity may be regulated tissue-specifically by the genomic environment of the locus. It suggests that such level of regulation is conserved and, as a consequence, there are likely many endogenous retrotransposons regulated in the manner we describe.

In addition to highlighting the role of the genomic environment in the regulation of LTR retroelements, we also identify a TF contributing to this regulation. Numerous TFs were shown to bind to retroelement sequences and regulate retrotransposon expression in different biological contexts [115, 116]. Examples include YY1 [117], RUNX3 [118], p53 [119], SRY [120], MeCP2 [121], SOX2 [122], PAX5 [123], or SOX6 [36]. In contrast to these findings, our work implies that regulation of the donor *rover-2R:14M* locus by a gut lineage-specific TF relies not on the *rover-2R:14M* itself, but on the genomic sequence upstream to the locus. As previously shown for other LTR retroelements [42–45] and in agreement with our *in silico* analysis, the 5′UTR *rover* sequence itself also carries multiple TF binding motifs and might interact with other regulatory proteins. Thus, we cannot exclude that TF binding within the *rover-2R:14M* locus could also contribute to the locus expression. However, based on the results obtained with our reporter assays, the *rover-2R:14M* 5′UTR sequence itself is not sufficient to allow for its expression in the gut. Considering this additional level of regulation by the upstream genomic region, which has not been investigated so far, the complete spectrum of tissue-specific TFs that regulate somatic retrotransposon expression in different biological contexts is certainly yet to be discovered.

Here, we identify Esg as a regulator of *rover-2R:14M* expression in the fly intestinal tissue. Esg, a snail-family TF, is well known to control stem cell fate in the fly intestinal lineage [97, 98, 124, 125]. In this context, Esg, enriched in progenitor cells, is known to act as a repressor of differentiation genes. Thus, its function as a transcriptional activator may appear surprising. However, Esg was shown to bind to a relatively large set of genes, including those with downregulated expression upon Esg knockdown [97, 98]. Furthermore, Snail, another closely related TF promoting mesoderm development in the fly embryo through repression of ectodermal genes [126], also functions as a transcriptional activator [127]. Thus, it is probable that Esg, perhaps via other co-factors and/or TFs, may promote transcription of some targets in the fly gut, including the *rover-2R:14M* .

Altogether, our work highlights a new level of locus-specific regulation of LTR/ERV elements by the genomic environment in which these retrotransposons are inserted. This notion is particularly relevant in light of high TE polymorphisms already mentioned above and documented not only in *Drosophila* [79, 38, 128], but also in humans [129–132] or other animal and plant species [133–135]. Indeed, the *rover-2R:14M* locus identified by us is a non-reference polymorphic insertion, as are many of the described somatically active human L1 retrotransposons [35]. The full extent to which TE polymorphism contributes to their somatic expression and mobility in different species remains to be addressed.

The biological significance of somatic retrotransposon expression and mobility, although not yet well understood, is now well appreciated and intensively investigated [136]. LTR/ERV-type retroelements are not mobile in humans, but they continue to mobilize in other mammalian species. Furthermore, retrotransposon activity may affect host tissues in multiple ways, including transposition-independent, through their transcripts or protein products. Thus, studies such as ours, helping to better understand retroelement regulation in diverse lineages, are relevant.

Acknowledgements

We thank Marius van den Beek, whose initial ONT DNA-seq data analysis led to the identification of the *rover-2R:*14M locus. We would also like to acknowledge S. Chambeyron for sharing and discussing unpublished data, A. Boivin for his comments on the manuscript, and J. Crocker, L. Teysset, C. Carré, E. Brasset, C. Saleh, Bloomington, and the Drosophila Genomics Resource Center (NIH Grant 2P40OD010949) for providing fly stocks or reagents. The high-throughput sequencing involved in this study was performed by the ICGex NGS platform of the Institute Curie [supported by the grants ANR-10-EQPX-03 (Equipex) and ANR-10-INBS-09–08 (France Génomique Consortium) from the Agence Nationale de la Recherche ("Investissements d'Avenir" program), by the Canceropole Ile-de-France, and by the SiRIC-Curie program—SiRIC Grant "INCa-DGOS4654", as well as the I2BC High-throughput Sequencing Facility, supported by France Génomique (funded by the French National Program "Investissement d'Avenir" ANR-10-INBS-09). The present work has also benefited from Imagerie-Gif core facility supported by l'Agence Nationale de la Recherche (ANR-10-INBS- 04/FranceBioImaging; ANR-11-IDEX-0003-02/Saclay Plant Sciences). We thank Valerie Nicolas for her assistance in quantification of confocal images. All reagents generated in this study, including plasmids and *Drosophila* stocks, are available from the lead contact upon request.

Author contributions: Natalia Rubanova (Conceptualization [equal], Formal analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), Darshika Singh (Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), Louis Barolle (Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), Fabienne Chalvet (Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), Sophie Netter (Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), Mickaël Poidevin (Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Resources [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), Nicolas Servant (Formal analysis [supporting], Funding acquisition [supporting], Methodology [supporting], Software [supporting], Supervision [equal], Writing—original draft [supporting], Writing—review & editing [supporting]), Allison J. Bardin (Conceptualization [equal], Formal analysis [supporting], Funding acquisition [equal], Investigation [supporting], Methodology [supporting], Resources [equal], Supervi-

sion [equal], Writing—original draft [supporting], Writing—review & editing [supporting]), and Katarzyna Siudeja (Conceptualization [lead], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [lead], Resources [lead], Supervision [equal], Visualization [equal], Writing—original draft [lead], Writing—review & editing [lead]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Fondation pour la Recherche Médicale (A.J.B., EQ202003010251) and the ERC StG "gutTEimpact" (K.S., 101078070). Salary support of K.S. is from Inserm; A.J.B. and M.P. from CNRS, F.C. from University Paris-Saclay, S.N. from University of Versailles St-Quentin, and L.B. from Ministère de l'Enseignement Supérieur (doctoral grant). Funding to pay the Open Access publication charges for this article was provided by ERC Starting grant "gutTEimpact" (101078070).

Data availability

Sequencing data generated for this study have been deposited at GEO (BioProject ID PRJNA1202082; ONT DNA-seq: PRJNA1202082; RNA-seq and ONT cDNA-seq: GSE285324) and are publicly available. Accession numbers of datasets published previously are provided in Supplementary Table S1. All original code is available from https://github.com/nrubanova/rover_ERV and https://doi.org/10.6084/m9.figshare.28903466.v1. The table "sequences.tab" that contains read sequences for singleton insertions and consensus sequences for all other TE insertions, along with their coordinates and sub-family information, is available in the Figshare repository: https://doi.org/10.6084/m9.figshare.28736633.v1.

References

1. Wells JN, Feschotte C. A field guide to eukaryotic transposable elements. *Annu Rev Genet* 2020;**54**:539–61. https://doi.org/10.1146/annurev-genet-040620-022145
2. Reus K, Mayer J, Sauter M *et al.* HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). *J Virol* 2001;**75**:8917–26. https://doi.org/10.1128/JVI.75.19.8917-8926.2001
3. Cosby RL, Chang N-C, Feschotte C. Host–transposon interactions: conflict, cooperation, and cooption. *Genes Dev* 2019;**33**:1098–116. https://doi.org/10.1101/gad.327312.119
4. Dopkins N, O'Mara MM, Lawrence E *et al.* A field guide to endogenous retrovirus regulatory networks. *Mol Cell* 2022;**82**:3763–8. https://doi.org/10.1016/j.molcel.2022.09.011
5. Jachowicz JW, Bing X, Pontabry J *et al.* LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet* 2017;**49**:1502–10. https://doi.org/10.1038/ng.3945
6. Chang N-C, Wells JN, Wang AY *et al.* Gag proteins encoded by endogenous retroviruses are required for zebrafish development. *Genetics* 2025;**122**:e2411446122. https://www.pnas.org/doi/10.1073/pnas.2411446122.
7. Simon M, Meter MV, Ablaeva J *et al.* LINE1 derepression in aged wild-type and SIRT6-deficient mice drives inflammation. *Cell Metab* 2019;**29**:871–85. https://doi.org/10.1016/j.cmet.2019.02.014
8. Zhao Y, Oreskovic E, Zhang Q *et al.* Transposon-triggered innate immune response confers cancer resistance to the blind mole rat. *Nat Immunol* 2021;**22**:1219–30. https://doi.org/10.1038/s41590-021-01027-8
9. Wang L, Tracy L, Su W *et al.* Retrotransposon activation during *Drosophila* metamorphosis conditions adult antiviral responses. *Nat Genet* 2022;**54**:1933–45. https://doi.org/10.1038/s41588-022-01214-9
10. Mietz JA, Fewell JW, Kuff EL. Selective activation of a discrete family of endogenous proviral elements in normal BALB/c lymphocytes. *Mol Cell Biol* 1992;**12**:220–8.
11. Deininger P, Morales ME, White TB *et al.* A comprehensive approach to expression of L1 loci. *Nucleic Acids Res* 2017;**45**:e31. https://doi.org/10.1093/nar/gkw1067
12. Pehrsson EC, Choudhary MNK, Sundaram V *et al.* The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat Commun* 2019;**10**:5640. https://doi.org/10.1038/s41467-019-13555-x
13. Chung N, Jonaid GM, Quinton S *et al.* Transcriptome analyses of tumor-adjacent somatic tissues reveal genes co-expressed with transposable elements. *Mobile DNA* 2019;**10**:39. https://doi.org/10.1186/s13100-019-0180-5
14. Ansaloni F, Scarpato M, Di Schiavi E *et al.* Exploratory analysis of transposable elements expression in the *C. elegans* early embryo. *BMC Bioinformatics* 2019;**20**:484.https://doi.org/10.1186/s12859-019-3088-7
15. Treiber CD, Waddell S. Transposon expression in the *Drosophila* brain is driven by neighboring genes and diversifies the neural transcriptome. *Genome Res* 2020;**30**:1559–69. https://doi.org/10.1101/gr.259200.119
16. Burn A, Roy F, Freeman M *et al.* Widespread expression of the ancient HERV-K (HML-2) provirus group in normal human tissues. *PLoS Biol* 2022;**20**:1559–69. https://doi.org/10.1371/journal.pbio.3001826
17. Carter TA, Singh M, Dumbović G *et al.* Mosaic *cis*-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. *eLife* 2022;**11**:e76257. https://doi.org/10.7554/eLife.76257
18. She J, Du M, Xu Z *et al.* The landscape of hervRNAs transcribed from human endogenous retroviruses across human body sites. *Genome Biol* 2022;**23**:231. https://doi.org/10.1186/s13059-022-02804-w
19. Chang N-C, Rovira Q, Wells J *et al.* Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Res* 2022;**32**:1408–23. https://doi.org/10.1101/gr.275655.121
20. Faulkner GJ, Garcia-Perez JL. L1 mosaicism in mammals: extent, effects, and evolution. *Trends Genet* 2017;**33**:802–16. https://doi.org/10.1016/j.tig.2017.07.004
21. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. *Nat Rev Genet* 2020;**21**:721–36. https://doi.org/10.1038/s41576-020-0251-y
22. Athanikar JN, Badge RM, Moran JV. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res* 2004;**32**:3846–55. https://doi.org/10.1093/nar/gkh698
23. Lavie L, Maldener E, Brouha B *et al.* The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res* 2004;**14**:2253–60. https://doi.org/10.1101/gr.2745804
24. Philippe C, Vargas-Landin DB, Doucet AJ *et al.* Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife* 2016;**5**:e13926. https://doi.org/10.7554/eLife.13926

25. Berrens RV, Yang A, Laumer CE *et al.* Locus-specific expression of transposable elements in single cells with CELLO-seq. *Nat Biotechnol* 2022;**40**:546–54. https://doi.org/10.1038/s41587-021-01093-1
26. Gerdes P, Chan D, Lundberg M *et al.* Locus-resolution analysis of L1 regulation and retrotransposition potential in mouse embryonic development. *Genome Res* 2023;**33**:1465–81. https://doi.org/10.1101/gr.278003.123
27. Baillie JK, Barnett MW, Upton KR *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 2011;**479**:534–7. https://doi.org/10.1038/nature10531
28. Evrony GD, Cai X, Lee E *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* 2012;**151**:483–96. https://doi.org/10.1016/j.cell.2012.09.035
29. Evrony GD, Lee E, Mehta BK *et al.* Cell lineage analysis in human brain using endogenous retroelements. *Neuron* 2015;**85**:49–59. https://doi.org/10.1016/j.neuron.2014.12.028
30. Upton KR, Gerhardt DJ, Jesuadian JS *et al.* Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* 2015;**161**:228–39. https://doi.org/10.1016/j.cell.2015.03.026
31. Tubio JMC, Li Y, Ju YS *et al.* Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* 2014;**345**:1251343. https://doi.org/10.1126/science.1251343
32. Scott EC, Gardner EJ, Masood A *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res* 2016;**26**:745–55. https://doi.org/10.1101/gr.201814.115
33. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet* 2020;**52**:306–19. https://doi.org/10.1038/s41588-019-0562-0
34. Nam CH, Youk J, Kim JY *et al.* Widespread somatic L1 retrotransposition in normal colorectal epithelium. *Nature* 2023;**617**:540–7. https://doi.org/10.1038/s41586-023-06046-z
35. Sanchez-Luque FJ, Kempen M-JHC, Gerdes P *et al.* LINE-1 evasion of epigenetic repression in humans. *Mol Cell* 2019;**75**:590–604. https://doi.org/10.1016/j.molcel.2019.05.024
36. Bodea GO, Botto JM, Ferreiro ME *et al.* LINE-1 retrotransposons contribute to mouse PV interneuron development. *Nat Neurosci* 2024;**27**:1274–84. https://doi.org/10.1038/s41593-024-01650-2
37. Gagnier L, Belancio VP, Mager DL. Mouse germ line mutations due to retrotransposon insertions. *Mobile DNA* 2019;**10**:15. https://doi.org/10.1186/s13100-019-0157-4
38. Mérel V, Boulesteix M, Fablet M *et al.* Transposable elements in *Drosophila*. *Mobile DNA* 2020;**11**:23. https://doi.org/10.1186/s13100-020-00213-z
39. Siudeja K, van den Beek M, Riddiford N *et al.* Unraveling the features of somatic transposition in the *Drosophila* intestine. *EMBO J* 2021;**40**:e106388. https://doi.org/10.15252/embj.2020106388
40. Araujo P, Casacuberta J, Costa A *et al.* Retrolyc1 subfamilies defined by different U3 LTR regulatory regions in the *Lycopersicon* genus. *Mol Gen Genomics* 2001;**266**:35–41. https://doi.org/10.1007/s004380100514
41. Mugnier N, Biémont C, Vieira C. New regulatory regions of *Drosophila* 412 retrotransposable element generated by recombination. *Mol Biol Evol* 2005;**22**:747–57. https://doi.org/10.1093/molbev/msi060
42. Minervini CF, Marsano RM, Casieri P *et al.* Heterochromatin protein 1 interacts with 5′UTR of transposable element ZAM in a sequence-specific fashion. *Gene* 2007;**393**:1–10. https://doi.org/10.1016/j.gene.2006.12.028
43. Göke J, Lu X, Chan Y-S *et al.* Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* 2015;**16**:135–41. https://doi.org/10.1016/j.stem.2015.01.005
44. Ito J, Sugimoto R, Nakaoka H *et al.* Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet* 2017;**13**:e1006883. https://doi.org/10.1371/journal.pgen.1006883
45. Geng LN, Yao Z, Snider L *et al.* DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev Cell* 2012;**22**:38–51. https://doi.org/10.1016/j.devcel.2011.11.013
46. Malik HS, Henikoff S, Eickbush TH. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* 2000;**10**:1307–18. https://doi.org/10.1101/gr.145000
47. Terzian C, Pélisson A, Bucheton A. Evolution and phylogeny of insect endogenous retroviruses. *BMC Evol Biol* 2001;**1**:3. https://doi.org/10.1186/1471-2148-1-3
48. Jiang H, Patel PH, Kohlmaier A *et al.* Cytokine/Jak/Stat signaling mediates regeneration and homeostasis in the *Drosophila* midgut. *Cell* 2009;**137**:1343–55. https://doi.org/10.1016/j.cell.2009.05.014
49. Bischof J, Maeda RK, Hediger M *et al.* An optimized transgenesis system for *Drosophila* using germ-line-specific φC31 integrases. *Proc Natl Acad Sci USA* 2007;**104**:3312–7. https://doi.org/10.1073/pnas.0611511104
50. Schindelin J, Arganda-Carreras I, Frise E *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* 2012;**9**:676–82. https://doi.org/10.1038/nmeth.2019
51. Buchon N, Osman D, David FPA *et al.* Morphological and molecular characterization of adult midgut compartmentalization in *Drosophila*. *Cell Rep* 2013;**3**:1725–38. https://doi.org/10.1016/j.celrep.2013.04.001
52. De Coster W, D'Hert S, Schultz DT *et al.* NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;**34**:2666–9. https://doi.org/10.1093/bioinformatics/bty149
53. Öztürk-Çolak A, Marygold SJ, Antonazzo G *et al.* FlyBase: updates to the *Drosophila* genes and genomes database. *Genetics* 2024;**227**:iyad211. https://doi.org/10.1093/genetics/iyad211
54. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191
55. Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools *Gigascience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008
56. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* 2023;**39**:btad311. https://doi.org/10.1093/bioinformatics/btad311
57. Leger A, Leonardi T. pycoQC, interactive quality control for Oxford Nanopore Sequencing. *JOSS* 2019;**4**:1236. https://doi.org/10.21105/joss.01236
58. Ewing AD, Smits N, Sanchez-Luque FJ *et al.* Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol Cell* 2020;**80**:915–28. https://doi.org/10.1016/j.molcel.2020.10.024
59. Storer J, Hubley R, Rosen J *et al.* The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* 2021;**12**:2. https://doi.org/10.1186/s13100-020-00230-y
60. Bailly-Bechet M, Haudry A, Lerat E . "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA* 2014;**5**:13. https://doi.org/10.1186/1759-8753-5-13
61. Han S, Basting PJ, Dias GB *et al.* Transposable element profiles reveal cell line identity and loss of heterozygosity in *Drosophila* cell culture. *Genetics* 2021;**219**:iyab113. https://doi.org/10.1093/genetics/iyab113
62. Rozewicki J, Li S, Amada KM *et al.* MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res* 2019;**47**:W5–10. https://doi.org/10.1093/nar/gkz342

63. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000;**16**:276–7. https://doi.org/10.1016/S0168-9525(00)02024-2
64. Robinson JT, Thorvaldsdóttir H, Winckler W *et al.* . Integrative Genomics Viewer. *Nat Biotechnol* 2011;**29**:24–26. https://doi.org/10.1038/nbt.1754
65. Darling ACE, Mau B, Blattner FR *et al.* Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004;**14**:1394–403. https://doi.org/10.1101/gr.2289704
66. Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635
67. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40. https://doi.org/10.1093/bioinformatics/btp616
68. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8
69. Wickham H. Build a plot layer by layer. In: Wickham H (ed.), *ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer International Publishing, 2016, 89–107.
70. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30. https://doi.org/10.1093/bioinformatics/btt656
71. Gershenzon NI, Trifonov EN, Ioshikhes IP. The features of *Drosophila* core promoters revealed by statistical analysis. *BMC Genomics* 2006;**7**:161. https://doi.org/10.1186/1471-2164-7-161
72. Tan G, Lenhard B. TFBSTools: an R/Bioconductor package for transcription factor binding site analysis. *Bioinformatics* 2016;**32**:1555–6. https://doi.org/10.1093/bioinformatics/btw024
73. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2024;**52**:D174–82. https://doi.org/10.1093/nar/gkad1059
74. Dutta D, Dobson AJ, Houtz PL *et al.* Regional cell-specific transcriptome mapping reveals regulatory complexity in the adult *Drosophila* midgut. *Cell Rep* 2015;**12**:346–58. https://doi.org/10.1016/j.celrep.2015.06.009
75. The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 2010;**330**:1787–97.
76. Josserand M, Rubanova N, Stefanutti M *et al.* Chromatin state transitions in the *Drosophila* intestinal lineage identify principles of cell-type specification. *Dev Cell* 2023;**58**:3048–63. https://doi.org/10.1016/j.devcel.2023.11.005
77. Ramírez F, Ryan DP, Grüning B *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;**44**:W160–5. https://doi.org/10.1093/nar/gkw257
78. Riddiford N, Siudeja K, Beek Mv *et al.* Evolution and genomic signatures of spontaneous somatic mutation in *Drosophila* intestinal stem cells. *Genome Res* 2021;**31**:1419–32. https://doi.org/10.1101/gr.268441.120
79. Rahman R, Chirn G, Kanodia A *et al.* Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. *Nucleic Acids Res* 2015;**43**:10655–72. https://doi.org/10.1093/nar/gkv1193
80. Mohamed M, Dang NT-M, Ogyama Y *et al.* A transposon story: from TE content to TE dynamic invasion of *Drosophila* genomes using the single-molecule sequencing technology from Oxford Nanopore. *Cells* 2020;**9**:1776. https://doi.org/10.3390/cells9081776
81. Bergman CM, Quesneville H, Anxolabéhère D *et al.* Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol* 2006;**7**:R112. https://doi.org/10.1186/gb-2006-7-11-r112
82. Kaminker JS, Bergman CM, Kronmiller B *et al.* The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol* 2002;**3**:RESEARCH0084. https://doi.org/10.1186/gb-2002-3-12-research0084
83. Dubnau J. The retrotransposon storm and the dangers of a Collyer's genome. *Curr Opin Genet Dev* 2018;**49**:95–105.
84. Cardelli M. The epigenetic alterations of endogenous retroelements in aging. *Mech Ageing Dev* 2018;**174**:30–46. https://doi.org/10.1016/j.mad.2018.02.002
85. Gorbunova V, Seluanov A, Mita P *et al.* The role of retrotransposable elements in ageing and age-associated diseases. *Nature* 2021;**596**:43–53. https://doi.org/10.1038/s41586-021-03542-y
86. Sousa-Victor P, Ayyaz A, Hayashi R *et al.* Piwi is required to limit exhaustion of aging somatic stem cells. *Cell Rep* 2017;**20**:2527–37. https://doi.org/10.1016/j.celrep.2017.08.059
87. Tang X, Liu N, Qi H *et al.* Piwi maintains homeostasis in the *Drosophila* adult intestine. *Stem Cell Rep* 2023;**18**:503–18.
88. Chung W-J, Okamura K, Martin R *et al.* Endogenous RNA interference provides a somatic defense against *Drosophila* transposons. *Curr Biol* 2008;**18**:795–802. https://doi.org/10.1016/j.cub.2008.05.006
89. Czech B, Malone CD, Zhou R *et al.* An endogenous small interfering RNA pathway in *Drosophila*. *Nature* 2008;**453**:798–802. https://doi.org/10.1038/nature07007
90. Ghildiyal M, Seitz H, Horwich MD *et al.* Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* 2008;**320**:1077–81. https://doi.org/10.1126/science.1157396
91. Varoqui M, Mohamed M, Mugat B *et al.* Temporal and spatial niche partitioning in a retrotransposon community of the *Drosophila melanogaster* genome. *Nucleic Acids Res* 2025;**53**:gkaf516. https://doi.org/10.1093/nar/gkaf516
92. Johnson WE. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol* 2019;**17**:355–70. https://doi.org/10.1038/s41579-019-0189-2
93. Gervais L, van den Beek M, Josserand M *et al.* Stem cell proliferation is kept in check by the chromatin regulators Kismet/CHD7/CHD8 and trr/MLL3/4. *Dev Cell* 2019;**49**:556–73. https://doi.org/10.1016/j.devcel.2019.04.033
94. Desset S, Meignin C, Dastugue B *et al.* COM, a heterochromatic locus governing the control of independent endogenous retroviruses from *Drosophila melanogaster*. *Genetics* 2003;**164**:501–9. https://doi.org/10.1093/genetics/164.2.501
95. Sarot E, Payen-Groschêne G, Bucheton A *et al.* Evidence for a *piwi*-dependent RNA silencing of the *gypsy* endogenous retrovirus by the *Drosophila melanogaster flamenco* gene. *Genetics* 2004;**166**:1313–21. https://doi.org/10.1534/genetics.166.3.1313
96. Senti K-A, Rafanel B, Handler D *et al.* Co-evolving infectivity and expression patterns drive the diversification of endogenous retroviruses. *EMBO J* 2025;1–20. https://doi.org/10.1038/s44318-025-00471-8
97. Korzelius J, Naumann SK, Loza-Coll MA *et al. Escargot* maintains stemness and suppresses differentiation in *Drosophila* intestinal stem cells. *EMBO J* 2014;**33**:2967–82. https://doi.org/10.15252/embj.201489072
98. Antonello ZA, Reiff T, Ballesta-Illan E *et al.* Robust intestinal homeostasis relies on cellular plasticity in enteroblasts mediated by miR-8-Escargot switch. *EMBO J* 2015;**34**:2025–41. https://doi.org/10.15252/embj.201591517
99. Li W, Prazak L, Chatterjee N *et al.* Activation of transposable elements during aging and neuronal decline in *Drosophila*. *Nat Neurosci* 2013;**16**:529–31. https://doi.org/10.1038/nn.3368
100. Treiber CD, Waddell S. Resolving the prevalence of somatic transposition in *Drosophila*. *eLife* 2017;**6**:e28297. https://doi.org/10.7554/eLife.28297

101. Jones BC, Wood JG, Chang C *et al.* A somatic piRNA pathway in the *Drosophila* fat body ensures metabolic homeostasis and normal lifespan. *Nat Commun* 2016;**7**:13856. https://doi.org/10.1038/ncomms13856
102. Chang Y-H, Keegan RM, Prazak L *et al.* Cellular labeling of endogenous retrovirus replication (CLEVR) reveals *de novo* insertions of the gypsy retrotransposable element in cell culture and in both neurons and glial cells of aging fruit flies. *PLoS Biol* 2019;**17**:e3000278. https://doi.org/10.1371/journal.pbio.3000278
103. Yang N, Srivastav SP, Rahman R *et al.* Transposable element landscapes in aging *Drosophila*. *PLoS Genet* 2022;**18**:e1010024. https://doi.org/10.1371/journal.pgen.1010024
104. Rigal J, Martin Anduaga A, Bitman E *et al.* Artificially stimulating retrotransposon activity increases mortality and accelerates a subset of aging phenotypes in *Drosophila*. *eLife* 2022;**11**:e80169. https://doi.org/10.7554/eLife.80169
105. Schneider BK, Sun S, Lee M *et al.* Expression of retrotransposons contributes to aging in *Drosophila*. *Genetics* 2023;**224**:iyad073. https://doi.org/10.1093/genetics/iyad073
106. Bodea GO, McKelvey EGZ, Faulkner GJ. Retrotransposon-induced mosaicism in the neural genome. *Open Biol* 2018;**8**:180074. https://doi.org/10.1098/rsob.180074
107. Solyom S, Ewing AD, Rahrmann EP *et al.* Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res* 2012;**22**:2328–38. https://doi.org/10.1101/gr.145235.112
108. Ewing AD, Gacita A, Wood LD *et al.* Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome Res* 2015;**25**:1536–45. https://doi.org/10.1101/gr.196238.115
109. Biteau B, Hochmuth CE, Jasper H. Maintaining tissue homeostasis: dynamic control of somatic stem cell activity. *Cell Stem Cell* 2011;**9**:402–11. https://doi.org/10.1016/j.stem.2011.10.004
110. Mita P, Wudzinska A, Sun X *et al.* LINE-1 protein localization and functional dynamics during the cell cycle. *eLife* 2018;**7**:e30058. https://doi.org/10.7554/eLife.30058
111. Ardeljan D, Steranka JP, Liu C *et al.* Cell fitness screens reveal a conflict between LINE-1 retrotransposition and DNA replication. *Nat Struct Mol Biol* 2020;**27**:168–78. https://doi.org/10.1038/s41594-020-0372-1
112. Schauer SN, Carreira PE, Shukla R *et al.* L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. *Genome Res* 2018;**28**:639–53. https://doi.org/10.1101/gr.226993.117
113. Lanciano S, Philippe C, Sarkar A *et al.* Locus-level L1 DNA methylation profiling reveals the epigenetic and transcriptional interplay between L1s and their integration sites. *Cell Genom* 2024;**4**:100498. https://doi.org/10.1016/j.xgen.2024.100498
114. Glaser J, Cova G, Fauler B *et al.* Enhancer adoption by an LTR retrotransposon generates viral-like particles causing developmental limb phenotypes. bioRxiv, https://doi.org/10.1101/2024.09.13.612906, 15 September 2024, preprint: not peer reviewed.
115. Sun X, Wang X, Tang Z *et al.* Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proc Natl Acad Sci USA* 2018;**115**:E5526–35. https://doi.org/10.1073/pnas.1722565115
116. Hermant C, Torres-Padilla M-E. TFs for TEs: the transcription factor repertoire of mammalian transposable elements. *Genes Dev* 2021;**35**:22–39. https://doi.org/10.1101/gad.344473.120
117. Becker KG, Swergold G, Ozato K *et al.* Binding of the ubiquitous nuclear transcription factor YY1 to a *cis* regulatory sequence in the human LINE-1 transposable element. *Hum Mol Genet* 1993;**2**:1697–702. https://doi.org/10.1093/hmg/2.10.1697
118. Yang N. An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res* 2003;**31**:4929–40. https://doi.org/10.1093/nar/gkg663
119. Wylie A, Jones AE, D'Brot A *et al.* p53 genes function to restrain mobile elements. *Genes Dev* 2016;**30**:64–77. https://doi.org/10.1101/gad.266098.115
120. Tchenio T. Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res* 2000;**28**:411–5. https://doi.org/10.1093/nar/28.2.411
121. Muotri AR, Marchetto MCN, Coufal NG *et al.* L1 retrotransposition in neurons is modulated by MeCP2. *Nature* 2010;**468**:443–6. https://doi.org/10.1038/nature09544
122. Kuwabara T, Hsieh J, Muotri A *et al.* Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. *Nat Neurosci* 2009;**12**:1097–105. https://doi.org/10.1038/nn.2360
123. Tang H, Yang J, Xu J *et al.* The transcription factor PAX5 activates human LINE1 retrotransposons to induce cellular senescence. *EMBO Rep* 2024;**25**:3263–75. https://doi.org/10.1038/s44319-024-00176-9
124. Micchelli CA, Perrimon N. Evidence that stem cells reside in the adult *Drosophila* midgut epithelium. *Nature* 2006;**439**:475–9. https://doi.org/10.1038/nature04371
125. Ohlstein B, Spradling A. The adult *Drosophila* posterior midgut is maintained by pluripotent stem cells. *Nature* 2006;**439**:470–4. https://doi.org/10.1038/nature04333
126. Kosman D, Ip YT, Levine M *et al.* Establishment of the mesoderm-neuroectoderm boundary in the *Drosophila* embryo. *Science* 1991;**254**:118–22. https://doi.org/10.1126/science.1925551
127. Rembold M, Ciglar L, Yáñez-Cuna JO *et al.* A conserved role for Snail as a potentiator of active transcription *Genes Dev* 2014;**28**:167–81. https://doi.org/10.1101/gad.230953.113
128. Rech GE, Radío S, Guirao-Rico S *et al.* Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. *Nat Commun* 2022;**13**:1948. https://doi.org/10.1038/s41467-022-29518-8
129. Stewart C, Kural D, Strömberg MP *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* 2011;**7**:e1002236. https://doi.org/10.1371/journal.pgen.1002236
130. Rishishwar L, Tellez Villa CE, Jordan IK. Transposable element polymorphisms recapitulate human evolution. *Mobile DNA* 2015;**6**:21. https://doi.org/10.1186/s13100-015-0052-6
131. Sudmant PH, Rausch T, Gardner EJ *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;**526**:75–81. https://doi.org/10.1038/nature15394
132. Chu C, Borges-Monroy R, Viswanadham VV *et al.* Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun* 2021;**12**:3836. https://doi.org/10.1038/s41467-021-24041-8
133. Zhou X, Sam TW, Lee AY *et al.* Mouse strain-specific polymorphic provirus functions as *cis*-regulatory element leading to epigenomic and transcriptomic variations. *Nat Commun* 2021;**12**:6462. https://doi.org/10.1038/s41467-021-26630-z
134. Alonge M, Wang X, Benoit M *et al.* Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 2020;**182**:145–61. https://doi.org/10.1016/j.cell.2020.05.021
135. Domínguez M, Dugas E, Benchouaia M *et al.* The impact of transposable elements on tomato diversity. *Nat Commun* 2020;**11**:4058. https://doi.org/10.1038/s41467-020-17874-2
136. Burns KH. Our conflict with transposable elements and its implications for human disease. *Annu Rev Pathol* 2020;**15**:51–70. https://doi.org/10.1146/annurev-pathmechdis-012419-032633

Received: January 17, 2025. **Revised:** April 24, 2025. **Editorial Decision:** April 27, 2025. **Accepted:** May 21, 2025