Temporal and spatial niche partitioning in a retrotransposon community of the Drosophila melanogaster genome

Marion Varoqui, Mourdas Mohamed, Bruno Mugat, Daniel Gourion, Maëlys Lemoine, Alain Pélisson, Charlotte Grimaud, Séverine Chambeyron

Temporal and spatial niche partitioning in a retrotransposon community of the *Drosophila melanogaster* genome

Marion Varoqui[1], Mourdas Mohamed[1], Bruno Mugat[1], Daniel Gourion[2], Maëlys Lemoine[1], Alain Pélisson[1], Charlotte Grimaud[1]* and Séverine Chambeyron[1]*

[1] Institute of Human Genetics, Univ. Montpellier, CNRS, Montpellier, France

[2]Avignon Université, LMA UPR 2151, 84000 Avignon, France

*co-corresponding authors

*Correspondence :charlotte.grimaud@igh.cnrs.fr; severine.chambeyron@igh.cnrs.fr

Abstract

Transposable elements (TEs) are genetic parasites that can potentially threaten the stability of the genomes they colonize. Nonetheless, TEs persist within genomes and are rarely fully eliminated, diverse TE species coexisting in various copy numbers. The TE replication strategies that enable host organisms to tolerate and accommodate the extensive diversity of TEs, while minimizing harm to the host and avoiding mutual competition among TEs, remain poorly understood. Here, by studying the spontaneous or experimental mobilization of four *Drosophila* LTR RetroTransposable Elements (LTR-RTEs), we reveal that, each of them preferentially targets open chromatin regions characterized by specific epigenetic features. Among these, gtwin and ZAM are expressed in distinct cell types within female somatic gonadal tissues and inserted into the distinct accessible chromatin landscapes of the corresponding stages of embryogenesis. These findings suggest that individual LTR-RTEs exploit unique biological niches, enabling their coexistence within the tightly regulated ecosystem of the same host genome.

Keywords

Transposable elements, retrotransposons, genome, chromatin, epigenetics, infection, germline

Introduction

Proper development of multicellular organisms relies on the temporally and spatially regulated expression of genes encoded by the genome. However, not every DNA sequence, even if it can be expressed within a genome, contributes to the fitness of the organism. Some sequences, such as transposable elements (TEs) exhibit a self-serving behavior due to their ability to take advantage of the host proteins to favor their expression and transposition in various locations within the host genome. Thus, they can be considered as genomic parasites (1). This capacity to move within the genome can generate harmful mutations such as disruptions of coding sequences, impairment of gene regulation and chromosomal rearrangements by ectopic recombination (2), which may ultimately jeopardize the integrity of the host genome. This is particularly relevant considering that maintenance of TEs in a host organism throughout generations, requires that they reach the germline, the carrier of the heritable host genetic information. In this context, the key question is to determine how each TE has been able to tune up its replication cycle in order to prevent extinction of either the host or the TE (3). On the other hand, during evolution, several defense mechanisms have been developed by the host to keep replicative transposition rates at low levels allowing proper balance between host survival and TE maintenance. Some of the mechanisms affecting the level of TE transposition, involve, in the germline and surrounding somatic cells, a specific class of small regulatory RNAs known as Piwi-interacting RNAs (piRNAs), which, when associated with PIWI proteins, a subclass of Argonaute proteins, can hybridize with nascent or cytoplasmic TE transcripts. This specific targeting by the host defense machinery leads to the silencing of TEs, either transcriptionally (TGS) or post-transcriptionally (PTGS), respectively (4–7).

We assume that present-day TE landscapes likely result from evolution of a series of host-TE interactions, including transposition repression, that, together, prevented extinction of either the host or the TE (3, 8). On the TE-host side, we hypothesize that the coexistence of numerous TE species within the same host genome was made possible by the evolution of specific features in the replication cycle of each species, which has allowed them to persist without harming the host and/or competing with each other.

As much as 20% of the *Drosophila melanogaster* genomic sequences derive from different classes of TEs (9). Within the same class, comparative analyses of conserved TE-encoding proteins have allowed their distribution into specific clades, each segregated into distinct species. For instance, comparative analyses of the conserved reverse transcriptase domain of Long Terminal Repeat-RetroTransposable Elements (LTR-RTEs), a class replicating via an RNA intermediate and representing about 10 % of the *Drosophila melanogaster* genome (9), revealed a distribution into three clades: Copia, BEL and Gypsy (10, 11). LTR-RTEs of the Copia and BEL clades encode a single open reading frame (ORF) and are represented by a few

species. The much more expanded Gypsy clade, on the other hand, displays a stronger heterogeneity in coding sequence with LTR-RTE species encoding one, two or three ORFs (10, 11). Several studies, using various *Drosophila* strains, have revealed that the Gypsy clade species are not only genetically diversified but have also adopted distinct niches of expression. While some species within this clade are expressed in germline cells, like most TEs, the majority are specifically expressed in the somatic cells surrounding the germline (12). This unique replication strategy is linked to the acquisition, by their common ancestor, of an ORF encoding a viral-like envelope protein which enables these elements to infect the germline (10, 12). By uncoupling the site of expression (tolerant, differentiated gonadal somatic cells) from the site of integration (germline), this replication cycle is likely less harmful for the host germline, providing a possible explanation for the evolutionary success of this clade. Moreover, the high diversity existing between the gonadal somatic cell types that are potential sources of viral-like particles able to infect the germline, has been fully exploited for the wide evolutionary diversification of this clade. Indeed, each type of ovarian somatic cell seems to be adapted as a specific niche for the expression of a particular LTR-RTE species (12). This expression niche partitioning probably provided diverse unique environments where species can thrive without having to compete with each other.

Similarly, the integration of TEs in eukaryotic genomes seems not random, indicating that several host-TE interactions have certainly been developed to allow such genome partitioning of TE insertions. Pioneering studies in the budding yeast *Saccharomyces cerevisiae* have revealed how co-optation of distinct endogenous proteins as TE cofactors have notably driven insertion niche partitioning for several LTR-RTEs species. Indeed, it appears that the LTR-RTE Ty5 benefits from the interaction of its integrase with the heterochromatin protein Sir4 to preferentially integrate in subtelomeric regions. On the other hand, two distinct LTR-RTEs Ty1 and Ty3 rather integrate in non-essential multicopy genes transcribed by RNA PolIII such as tRNAs genes (13). Note, although they share the same global insertion environment, each of these elements has its own insertion site preference that is mainly dependent on the interaction of its integrase with specific cellular cofactors (13).

More recent data in *Drosophila melanogaster* suggest that TEs, which belong to distinct classes and differ by their transposition mechanism, generally share insertion preference for open chromatin regions (14, 15), but display distinct insertion patterns. For example, the DNA transposon P-element favors integration of its DNA in replication origins (14, 16), while RTEs integrate their cDNAs either near the promoters and exons of active genes, for all LTR-RTEs species, or toward the telomere, for the non-LTR-RTE I-element (14). Whether integration preferences may also vary between the different species of the same LTR class and/or are influenced by specific cellular contexts are still open questions.

Studying LTR-RTE ecology regarding not only the interactions between a TE species and its host but also between members of the whole community of TE species having colonized the same host, is therefore expected to provide further insights into the ways they have successfully invaded all present-day eukaryotic genomes.

To investigate to what extent the active LTR-RTE species present in a same organism may differ in their replication cycles, we used a *Drosophila melanogaster* strain that we had previously constructed to impair the Piwi-mediated LTR-RTE repression specifically in the somatic tissue of the gonads (Supplementary Figure S1) (17). This strain contains a traffic-jam-Gal4/Gal80^{ts} inducible driver which activates, at permissive temperature, the expression of a short RNA hairpin targeting Piwi (sh-piwi) in the gonadal somatic cells. This somatic knockdown (sKD) alleviates LTR-RTE repression in these cells without causing sterility (17). When females of this strain are transferred for 5 days from the 20°C non-permissive to the 25°C permissive temperature, they display a partial depletion of the Piwi protein in their ovarian somatic cells (piwi-sKD), leading to an accumulation of LTR-RTE transcripts in these cells. *De novo* germline insertions of two LTR-RTEs from the Gypsy clade, ZAM and gtwin, were detected in short-read genomic libraries from embryos sequenced two generations (F2) after piwi-sKD (Supplementary Figure S1) (17). As a proof of concept, we also demonstrated that this strain is a powerful tool for studying the accumulation of *de novo* germline insertions of at least ZAM and gtwin LTR-RTEs. At that time, it was illustrated by the increase in their copy number, approximately estimated by genomic PCR, following the application of piwi-sKD through successive generations up to generation 72.

In the present study, by performing long-read sequencing of genomic DNA obtained from F2 male flies after 11, 31 and 73 generations of piwi-sKD, we were able to: (1) verify the increased numbers of new insertions for ZAM and gtwin across successive generations and also document new insertions for three other LTR-RTEs species, namely roo, copia and rover; (2) map enough new germinal insertions of four of these LTR-RTE species to reveal differences in their landing site preferences, particularly for distinct epigenetic marks associated with open chromatin. Furthermore, we highlighted that gtwin and ZAM replication cycles exhibit differences not only in their expression patterns but also in the timing of their integration into the different accessible chromatin landscapes of the developing embryonic germline, which could explain some of their site preferences. Our findings emphasize how, over the course of evolution, the diversity of the cell identities that different LTR-RTE species exploit for both expression and integration has facilitated their colonization of specific niches, enabling their coexistence within this ecosystem.

Materials and methods

Drosophila genotypes

As previously described (Supplementary Figure S1) (17), all flies used to determine LTR-RTE mobilization and integration, shared the genotype of the founder G0 strain : *w ; tj-Gal4 ; tubP-Gal80^{ts}, sh-piwi*. The polymorphism of this strain had been partially reduced by isolating a single pair of parents, and the strain was thereafter stably maintained at 20°C as a large population (more than 500 progenitors at each of the 100 successive generations of the G0F100 population). An independent subset of the G0 population was bred using more than 500 flies per generation, the temperature being raised at each generation from 20°C to 25°C for a 5-day period during the adult stage (Figure 1A). At the 11th (G11), 31th (G31) and 73rd (G73) generation, a large subset of GnF1 progenitors (approximately 500 flies, from the nth generation of interest) of the treated population was isolated and maintained *en masse* at 20°C, the non-permissive temperature for piwi-sKD. A strain harboring the genotype: *w ; vas::EGFP* was also used (18). The *piwi-AID-GFP* strain was a generous gift of G. Hannon (19).

Oxford Nanopore Technology (ONT) Sequencing Data Analysis

As previously described (20), genomic DNA was extracted from 100 GnF2 males (Figure 1A), and long-read sequencing data were analyzed using the TrEMOLO software (v2.2) (21) with some modifications. To detect newly integrated transposable elements, we employed the OUTSIDER TE detection module with, as a reference, the Dmel_R6.32 reference genome from FlyBase (v.104). Settings parameters for size and identity were set at 80%. The LTR-RTE database was extracted from the collection of reference TEs from Bergman's laboratory (https://github.com/bergmanlab/transposons). The quality of the reads is presented in Supplementary Table S1. According to (21), the sequencing depths of all libraries (except for that of the non polymorphic G0 strain) were estimated to be similar enough, to spare us from downsampling the largest ones (Supplementary Table S1). Frequency estimation was conducted using the TE analysis module of TrEMOLO (v2.5) and reads identified as clipped reads by TrEMOLO were excluded from the frequency calculation.

Annotation of false positive new insertions

The G0F100 library and the other libraries were respectively established with two populations that independently evolved from a shared G0 ancestor strain. Consequently, any insertion found in both the G0F100 and any other library was attributed to the G0 parental genome. This allowed us to annotate as false negative pre-existing insertions those that were likely missed in the low quality G0 parental library,

characterized by low coverage and shorter reads. All annotations were performed on the Dmel_R6.32 reference genome from FlyBase (v.104).

Annotation of newly integrated LTR-RTEs in piRNA clusters

The piRNA clusters were annotated on the Dmel_R6.32 reference genome using the published database https://www.smallrnagroup.uni-mainz.de/piRNAclusterDB/data/FASTA/Drosophila_melanogaster.piRNAclusters.gtf). Then a comparison between piRNA cluster coordinates and the LTR-RTE coordinates was used to determine the presence of new insertion in piRNA clusters.

Small RNA purification and sequencing

Small RNAs from ovaries collected at permissive (25°C) and non-permissive (20°C) temperature, for Piwi-sKD, were isolated using TraPR ion exchange spin columns (Lexogen, Catalog Nr.128.08). The libraries were performed by MGX-Biocampus Montpellier plateform using the NEBNext® Small RNA Library Prep Set for Illumina® from NEB. The sequencing was performed on flow cell SP paired-end 28-90nt on NOVASEQ 6000 apparatus by MGX. Raw reads were trimmed from their 3′ linkers and loaded on a homemade pipeline available at (https://bitbucket.org/blaiseli/pirna-pipeline) previously used (17). Briefly, trimmed reads (18–30nts in size) were mapped with Bowtie2 (22) using mismatch-tolerant settings to the *Drosophila melanogaster* genome (release 5; dm3) complemented with canonical TEs (*Drosophila* consensus TE sequences taken from https:// github.com/cbergman/transposons). Reads were annotated based on their mapping coordinates. Small RNAs mapping on piRNA clusters (23), ovary siRNA clusters (24), TEs or 3′UTR of coding genes (ftp://ftp.flybase.net/), and not to rRNAs or miRNAs were defined. Candidate piRNAs were a subset of the above defined reads with a size between 23 and 30 nucleotides. Candidate piRNAs were mapped again on canonical TE sequences. Data were normalized using the total of piRNA reads.

Single-molecule inexpensive RNA fluorescence *in situ* hybridization (smiRNA-FISH) probe preparation

39-48 probes of 20 nucleotides targeting specifically ZAM, gtwin, roo or copia transcripts were designed using Oligostan script (25). Primary probes were produced in 96-well plates. For convenience, the oligonucleotides are delivered in Tris-EDTA pH 8.0 (TE) buffer, at final concentration of 100µM. An equimolar mixture of the different primary probes was prepared and diluted 5 times in TE buffer to obtain a final concentration of 0.833µM for each individual probe. Fluorescent labeled FLAP-X (5'-Cy3/CACT GAG TCC AGC TCG AAA CTT AGG AGG/Cy3-3' or FLAP-Y (5′-Cy3/AA TGC ATG TCG ACG AGG

TCC GAG TGT AA/Cy3-3′) were delivered lyophilized and resuspended in TE buffer at final concentration of 100µM. The reverse complement of each of these respective sequences was added at the 3'end of each specific probe (Supplementary Table S7). Annealing between specific probes and their respective FLAP was performed as previously described (25) and then diluted in hybridization buffer.

smiFISH in ovaries and embryos

Ovaries were dissected in PBS1X and fixed during 20 minutes in PBS-Triton 0,3% (PBS-Tr) containing 4% formaldehyde. After several washes in PBS-Tr, ovaries were immersed in 100% methanol by successive baths in a PBS-Tr solution containing an increasing percentage of methanol. At this stage, ovaries can be kept in methanol at -20°C for several weeks. Embryos were collected and dechorionated in 2.6% bleach. They were rinsed extensively with water and fixed in 1:1 volume of fixative solution (4%Formaldehyde, KCl 60mM, Nacl 150mM, spermidine 0,5mM, Spermine 0,15mM, EDTA 2mM, EGTA 0,5mM, PIPES 15mM) and heptane for 25min at room temperature with agitation. Upon removal of the aqueous phase, an equal volume of 100% methanol was added before a vortexing for 1 min. Devitellinized embryos were collected from the methanol phase and then washed 3 times with 100% methanol. At this stage, embryos can be kept in methanol at -20°C for several weeks. Fixed embryos or ovaries were first washed twice in 50% methanol/50% ethanol for 5 minutes, rinsed twice in 100% ethanol and then washed two times in 100% ethanol for 5 minutes. They were incubated in blocking buffer (PBS 1X, tween 0,1%, RNAsin, BSA 0,2mg/mL in nuclease-free H_2O) for 1 hour (a wash every 15 minutes) and once in wash buffer (SSC 2X, deionized formamide 10%, H2O in nuclease-free H_2O) before the O/N incubation at 37°C at 350 rpm with smiFISH probes (Supplementary Table 6) and either an anti-Rat Vasa antibody (DHSB, 1:120) or a Guinea Pig traffic jam antibody (gift from D. Godt (26), Toronto, 1: 120) diluted in the hybridization buffer (10% deionized formamide, 2X SSC, 100mg tRNA, 5% dextran sulfate, 2mM VRC (NEB), 0,2mg/mL BSA). Subsequently, embryos/ovaries were washed with a wash buffer twice for 1 hour at 37°C and once for 1 hour at room temperature. Embryos were transferred in PBS, 0,1% Tween (PBT), 10% donkey serum and either Donkey anti Rat Alexa 488 (Molecular probes) or Donkey anti Guinea Pig Alexa 594 (Molecular probes) was added at 1:500 dilution. After several washes in PBT and DAPI staining, embryos/ovaries were mounted in ProLong Gold Antifade mounting medium (Thermo Fisher Scientific). smiFISH coupled with vasa immunostaining was imaged with LSM 880 with SR Airyscan module (Zeiss) using 40X/1.4 N.A objective. Airyscan processing was performed using 2D Zen Black v3.2 (Zeiss) prior to analysis. smiFISH coupled with traffic jam immunostaining was imaged with Leica SP8 confocal microscope equipped with 40X/1.4 N.A objectives. Image acquisition was done with the following settings: 2048x2048 pixels or 1024x1024 pixels, 16-bit depth.

Immunofluorescence on *Drosophila* embryos

We performed double immunostaining on fixed embryos, from a cross between females of the G73 population and males expressing Piwi coupled to GFP (19), with a mouse anti-traffic jam antibody (M. Siomi, NIG-Fly) and a rabbit anti-GFP antibody (Abcam, ab290). Fixed embryos stored in 100% methanol were successively incubated during 15 min in 90/10, 70/30, 50/50 and 30/70 percent methanol/PBT. They were permeabilized with PBS-Tr for 30 min and blocked in a PBS-Tr solution containing 10% NDS (Normal Donkey Serum) for 1 hour. Embryos were incubated overnight at 4°C on a rotating wheel with PBSTr-10% NDS containing the two primary antibodies diluted 1:500. After several washes in PBS-Tr, embryos were incubated during 45 min at room temperature with PBS-Tr-10%NDS containing an anti-mouse-Cy5 (Jackson laboratories,715.175.150) diluted 1:500 and an anti-rabbit Alexa 488 (Molecular probes) diluted 1:800. After several washes in PBS-Tr, DNA was counterstained with DAPI, and embryos were mounted in Prolong Antifade medium (Molecular Probes). Immunostaining was imaged with LSM 880 with SR Airyscan module (Zeiss) using 40X/1.4 N.A objective. Airyscan processing was performed using 2D Zen Black v3.2 (Zeiss) prior to analysis. Image acquisition was done with the following settings: 2000x2000 pixels, 16-bit depth.

Clustering

Raw data from ModENCODE Chip-seq experiments (27) performed on 14-16h embryos (Supplementary Table S6) were analyzed to generate BigWig files using deepTools (v3.5.4.post1) bamcoverage package with default parameters, excluding regions (-bl) identified as blacklisted in dm6 (27). bigwigAverage from deepTools (v3.5.4.post1) package was then used to average duplicates. Then, the bigwigCompare package from deepTools (v3.5.4.post1) was used to obtain the log2 ratio BigWig file between the Chip-seq averaged BigWig and the Input average BigWig. From this, the enrichment signal of the histone modifications H3K4me3, H3K9ac, H3K36me3, H3K27ac, H3K4me1, H3K36me1, H3K27me3, H3K9me2 and H3K9me3 on 5kb windows were computed across the *Drosophila melanogaster* genome (BDGP6.46) using the computeMatrix Package from deepTools (v3.5.4.post1) with scale-regions mode and the the (--averageTypeBins) option with a 50 bp interval. The clustering was performed to generate 8 clusters using the plotHeatMap package from deepTools (v3.5.4.post1) using the option (--kmeans).

Distribution of LTR-RTEs insertions relative to genomic features and chromatin states

Using the chromosomal gene and exons annotations of *Drosophila melanogaster* genome (BDGP6.46) available on Ensembl Biomart (28) except for the Y chromosome, we partitioned the genome in three mutually exclusive regions corresponding to exons, introns and intergenic regions. Exons were already annotated in a bed file (28). Introns were defined as genomic regions that are present in the gene bed file

and which are not in the exon bed file. Intergenic regions are defined as genomic regions that do not overlap with the gene bed file. Using this partition and our annotations of LTR-RTEs insertion sites, we then determined the number of copia, roo, gtwin and ZAM insertions occurring in these three categories of genomic regions (Supplementary Table S5). To determine whether a specific structure (Intergenic, Intron and Exon) is enriched or depleted for insertions of each considered LTR-RTE, bilateral binomial statistical tests were performed. To do so, the size of each structure relative to the genome was computed using bedtools genomecov (29) default parameters, defining the relative size of intergenic regions (pig=0.314359), introns (pin=0.418317) and exons (pex=0.267325) (Supplementary Table S5). Null hypothesis corresponds to the probability for each LTR-RTE species to be inserted in each defined structure due to its proportion in the genome. We used the Benjamini-Hochberg step up procedure to control the false discovery rate (FDR), which is defined as the expected value of the proportion of erroneous rejection of the null hypothesis when conducting multiple comparisons.

Chromatin state annotations previously published (30) based on dm3 genome version were transformed to the latest version (Dmel_R6.32) using liftOver tool (31). As for genomic features, the genomic proportion of each chromatin state or identified cluster was computed using bedtools genomecov (29) default parameters (Supplementary Table S5). Significant enrichment or depletion of LTR-RTE insertions in the different chromatin states and clusters were calculated using bilateral binomial statistical test considering the null hypothesis as the probability of insertion in a given state for each LTR-RTE species to be equal to the relative size of this state within the genome. As in the previous part, we used the Benjamini-Hochberg step up procedure to control the false discovery rate (FDR).

Analysis of ATAC-seq and ChIP-seq available datasets

Raw data from published ATAC-seq and ChIP-seq experiments (Supplementary Table S6) were analyzed to generate BigWig files using deepTools (v3.5.4.post1) bamcoverage package with default parameters, excluding regions (-bl) identified as blacklisted in dm6 (27). BigWig files were first used with the ComputeMatrix package with reference-point mode from deepTools to filter and sort regions based on their scores in order to compute signal distributions centered on the LTR-RTE insertion sites in a region spanning 2kb upstream and downstream of the insertion. The mean number of reads across the 4kb window was calculated using the (--averageTypeBins) option from ComputeMatrix, with a 50bp interval. On the other hand, BigWig files were converted into BedGraph format using the UCSC tool bigWigToBedGraph (32). macs2 (v2.2.7.1) bdgpeakcall package (33) was used to perform a peak calling to generate bedfiles with the following parameters: (--cutoff) manually set; (--min-length) 60 and (--max-gap) 150. For the transcription factors ChIP-seq (Supplementary Table S6), a consensus bed file was created by keeping only overlapping regions from the different replicates using bedtools intersect (v2.27.1) (29). Random profile was generated

using 100 random profiles each corresponding to an average profile obtained from 100 random positions (4kb window). We used ComputeMatrix with reference-point mode from deepTools as described above. PlotProfile with the (--outFileNameData) option was used to obtain each distribution of the average read number for the 100*100 randomly selected positions generated bed files. Finally, the mean number of read matrix was computed and used with deepTools plotProfile for visualization.

Statistical analysis of available sci-ATAC-seq datasets

BigWig files normalized by Counts Per Million (CPM) from the sci-ATAC-seq atlas previously published (34) were analyzed based on two criteria: they must represent an identified cell type and cover at least 70% of genomic data. ComputeMatrix package was used to assess the average chromatin accessibility around the 210 insertions of gtwin and the 101 insertions of ZAM as previously described. To determine the number of gtwin insertions shared between the first four clusters found in the 0-2 hour window of embryonic development, bed files were created as described before and visualized with a Venn Diagram.

To analyze globally chromatin accessibility throughout the 8 time windows of embryonic development, sci-ATAC-seq signals (200 bp window) centered around ZAM or gtwin insertions were averaged for each defined cluster of each time window. The same technique was applied to 100 randomly selected regions of a 200 bp window. Ratios between the average sci-ATAC-seq ZAM (or gtwin) signals and random ones result in a single data per cluster in a defined time window. Data corresponding to the same time window were used to generate boxplots and statistical analysis.

Isolation of embryonic cells and cell sorting by flow cytometry

The embryonic cells were isolated as previously described (35). Briefly, overnight laid embryos from *vas*::EGFP line (18) were collected at 25°C and dechorionated in 2.6% bleach. Dechorionated embryos (i.e 400 mg) were transferred in a 7 mL Tenbroeck tissue grinder WHEATON™ filled with 6 mL of Schneider's insect medium for homogenization with 2 slow strokes before a 700g centrifugation for 10 minutes at 4°C. The pellet was resuspended in 4 mL of PBS 1X containing 0.1% of Trypsin-EDTA and incubated at room temperature for 20 minutes. The addition of 4 mL of ice-cold PBS 1X containing 20% fetal bovine serum is sufficient to stop Trypsin reaction before a 700g centrifugation for 10 minutes at 4°C. Pellet containing separated embryonic cells was resuspended in Schneider's insect medium (2mL) and filtered in a 40µm mesh before the addition of 1 mL of Schneider's insect medium. A final filtration in a 20µm mesh was performed before cell sorting by flow cytometry. Embryonic cellular samples were analyzed using a 4-Laser-V16-B14-R8YG10 Aurora spectral cell sorter (Cytek, Biosciences, USA) to sort GFP-positive Primordial germ cells (PGC) from GFP negative somatic cells through a measurement of complete fluorescence spectrum of individual cells. GFP signal was determined by a 488 nm excitation line and

detected in its full spectrum emission with B1 as peak channel (498nm-518nm). 2.5 x 10^5 events were recorded per sample and analyzed using the SpectroFlo software version 1.2.1 (Cytek, Biosciences USA). To define and sort the target cell populations (GFP-positive cells), three successive steps of gating were applied. First, cells were gated using the two physical parameters FSC and SSC excluding dead cells and debris. Second, doublets were excluded by comparing the width versus the area of SSC and FSC. Finally, FSC dot plot and GFP signal reported as percentage in positive or negative cells were used to gate and sort the two populations. Live cell sorting experiments were performed at 4°C with a 70µm nozzle that allows sorting at high speed (2 x 10^4 events per second). Sorted cells were collected into PBS containing 20% of Fetal Bovine serum (FBS) prior to a final centrifugation at 700g at 4°C and a -80°C freezing in DMSO supplemented with FBS.

ATAC-Seq experiments and analysis

ATAC-Seq experiments were performed using the ATAC-Seq kit from Diagenode (catalogue no. C01080002). Input material was between 100,000 to 130,000 cryopreserved PGCs (GFP-positive) cells isolated from whole embryos. Tagmentated DNA was amplified by PCR using 13 cycles and the purified DNA libraries were sequenced (paired-end sequencing 150 bp, roughly 2 Gb per sample) by Novogene (https://en.novogene.com/). ATAC-Seq were performed in duplicates, following Encode's standards (https://www.encodeproject.org/atac-seq/#standards). After initial quality checks of the sequencing data from the PGC (GFP positive cells) and the somatic cells (GFP negative cells) using FastQC (v0.12.1), the adapters (CTGTCTCTTATACACATCTNNNNNNNN) were trimmed using cutadapt (v4.2). Cleared reads were aligned to the *Drosophila* genome (Dmel_R6.32 release) using bowtie2 (v2.5.1). Duplicate alignments were removed using the fixmate and markdup packages of samtools (v1.17). The read coverage normalized by RPKM (--normalizeUsing) were computed using bamCoverage package from deepTools (v3.5.4.post1) with default parameters, excluding regions (-bl) identified as blacklisted in dm6. Chromatin accessibility averages for the duplicates were calculated using the bigwigAverage tool from deepTools package (v3.5.4.post1) to generate an averaged bedgraph. The chromatin accessibility in a region spanning 2kb upstream and downstream of ZAM insertions and 100 random genomic positions were computed with the ComputeMatrix package from deepTools as described above. The mean signal obtained from the duplicate was then computed using Matlab (R2024). To perform a peak calling, the BigWig files were converted into BedGraph files using USCS tool bigWigToBedGraph (32). The peak calling was then computed using macs2 (v2.2.7.1) bdgpeakcall package (33) with the following parameters: (--cutoff) manually set; (--min-length) 60 and (--max-gap) 150. The bedtool subtract package (29) was used to identify the genomic regions that are only open in the PGC (GFP positive) cells using the bed file generated using the sequencing of the somatic cells (GFP negative). To determine the number of gtwin insertions shared between the first four

clusters found in the 0-2 hour window of embryonic development and the PGC, bed files were created as described before and visualized with a Venn Diagram (Supplementary Table S8).

Statistical analyses and visualization

Statistics and data visualization were performed using the ggplot2 (v3.4.3) and eulerr (v7.0.2) libraries(https://ggplot2.tidyverse.org) on R (v4.3.1) (https://www.R-project.org/) and Matlab (R2024). Cytoscape (v3.10.1) (36) was used to create a graph.

Results

Four LTR-RTEs exhibit transcriptional and insertional activity across successive generations

Our previous qPCR analysis suggested a progressive genomic copy number increase of two LTR-RTEs, ZAM and gtwin, after 30, 41, and 72 successive generations of conditional piwi-sKD (i.e G30, G41, and G72)(17). For these experiments, genomic DNA was extracted from second filial generation (F2) embryos laid by flies reared at a non-permissive temperature (20°C), which allows somatic Piwi expression, while their grandparents had been exposed to a permissive (25°C) that impairs Piwi expression (Supplementary Figure S1) (17). In the present study, we employed long-read genomic sequencing to provide a comprehensive analysis of LTR-RTE genomic insertions. Genomic DNA was extracted from 100 adult males of F2 progenies from the G11, G31, or G73 populations (Figure 1A; Supplementary Table S1). Additionally, we sequenced the genome of the original G0 strain, and a G0-derived strain that had been maintained at 20°C for 100 generations (G0F100).

To annotate LTR-RTE genomic insertions, we employed the TrEMOLO method, recently developed by our group (21). Comparisons of annotated insertions for G11, G31, G73, with the parental G0 strain revealed 280, 514, and 798 novel LTR-RTE insertions, respectively, spanning 43 distinct LTR-RTE species (Figure 1B; Supplementary Table S2). Notably, the number of genomic insertions of ZAM and gtwin increased significantly and constantly from G11 to G73 (Figure 1B), which is in line with our previous findings (17). The number of copia insertions also increased but at a slower rate compared to ZAM and gtwin. Insertions of roo and rover increased between G11 and G31 but remained relatively stable thereafter (Supplementary Table S2 ; Figure 1B). Interestingly, the number of insertions for copia, roo, and rover also increased in the G0F100 strain, which was continuously reared at 20°C. This observation suggests that these three species can transpose spontaneously in our strain, independently of the piwi-sKD treatment (Figure 1B).

Some of these spontaneous insertions may have occurred during the development of F2 males raised at 20°C, possibly in their somatic cells. The case of rover supports this hypothesis, as the frequency of new rover insertions remained consistently very low (Supplementary Table S3) and no evidence of vertical transmission was observed. Indeed, unlike certain copia and roo insertions, which were shared between small samples of sequenced genomes and likely represent germline-inherited insertions, no single rover insertion was detected at the same genomic position in at least two samples (Supplementary Table S4). Based on these data, we concluded that ZAM, gtwin, roo, and copia, produced new germinal insertions that were vertically transmitted to the analyzed males.

To confirm the induced derepression of ZAM and gtwin and the spontaneous expression of copia and roo under piwi-sKD conditions, we analyzed the impact of piwi-sKD on piRNA levels in fly samples. We hypothesized that under permissive conditions for piwi-sKD, not only would Piwi levels decrease, as previously shown (17), but the levels of antisense piRNAs targeting the induced species (ZAM and gtwin) would also be reduced, while those targeting the non-induced species (copia and roo) would remain unaffected. To test this, we performed small RNA sequencing to compare antisense piRNA levels against these four species in G0F100 ovaries under permissive 25°C and non-permissive 20°C conditions. We observed that the levels of copia and roo piRNAs were slighty increased after Piwi depletion for unknown reasons. In contrast, the levels of ZAM and gtwin piRNAs were drastically reduced by 5- and 6-fold, respectively (Figure 1C).

We next conducted single molecule inexpensive fluorescent *in situ* RNA hybridization (smiRNA-FISH) to examine LTR-RTE expression levels in the ovaries at 20°C and 25°C (Figure 1D). Interestingly, our analysis revealed that roo and copia transcripts were detected in the ovaries regardless of temperature (Figure 1D). The roo transcripts accumulated in the oocyte cytoplasm at both temperatures, as it has been previously observed for transcripts of the LINE-RTE I-element (37, 38). The copia transcripts were detected in the nuclei of the follicle cells. Contrary to roo and copia, ZAM and gtwin are clearly not expressed in the ovaries at 20°C but start to express in the follicle cells of the ovaries at 25°C when Piwi is depleted (Figure 1D). ZAM expression is restricted to the posterior follicle cells as previously reported (12, 39), whereas gtwin seems to have a broader expression pattern throughout the follicle epithelium (Figure 1D). Overall, our approach enabled us to accumulate a substantial number of new germline insertions for four actively transposing LTR-RTEs (ZAM, gtwin, roo, and copia) and to observe that each appears to occupy a distinct expression niche within the ovary.

Distinct chromatin niches for genomic insertions of four LTR-RTEs

We next investigated potential biases in the genomic insertion sites of the four LTR-RTEs. For this analysis, we selected the G73 sequence dataset due to its higher number of insertions compared to other datasets.

Since we worked with large populations of flies harboring polymorphic LTR-RTEs insertions, selection could have favored the survival of individuals with beneficial insertions, potentially at the expense of those with neutral or deleterious ones. To test for evidence of positive selection, we first estimated the frequency of each new insertion in the G73 population using our lond read-sequencing data. Our analysis revealed that most insertions were segregating at low frequencies in the population (Supplementary Figure S2A). The few insertions detected at high frequencies warrant further study. We then focused on another possible signature of positive selection: the integration of LTR-RTE into piRNA clusters. Indeed, among the new insertions, those occurring in piRNA clusters are of particular interest, as these regions are known to act as sources of piRNAs (4, 40). Such insertions are expected to influence the piRNA population, favoring the production of piRNAs that silence the expression of the corresponding TE. However, as shown in Supplementary Figure S2B, none of the 17 new insertions identified in annotated piRNA clusters appeared significantly more frequent than the others (Supplementary Figure S2C).

Moreover, purifying selection is also known to be stronger against insertions into genes, and even more into coding sequences, than into intergenic regions. To determine the distribution of newly integrated LTR-RTEs, we partitioned the *D. melanogaster* genome into intergenic, intronic, and exonic regions, and quantified the number of LTR-RTE insertions within each category (Figure 2A). As a reference, "expected" values correspond to the proportional sizes of these three genomic bins (Supplementary Table S5). The number of LTR-RTE insertions in the G73 dataset was then distributed according to these proportions. For example, the "observed" number of intronic insertions for gtwin slightly but significantly exceeded the "expected" values, whereas the "observed" number of exonic insertions for gtwin were correspondingly lower than the "expected" values. Similar trends were observed for the other three LTR-RTEs. Interestingly, the levels of this exonic depletion were always much lower than those previously noticed for older germline insertions that had been subjected to long periods of purifying selection in natural populations (41). Instead, the observed patterns in our G73 dataset were more reminiscent of those of recent insertions that would not have yet been eliminated by purifying selection (14).

Given the observation that natural selection does not appear to have significantly influenced the distribution of LTR-RTEs in our G73 population, we decided to investigate whether specific chromatin landscapes could define the localization of each of the four LTR-RTEs within the genome. Using chromatin immunoprecipitation sequencing (ChIP-seq) dataset from *Drosophila* embryos at 14–16 hours post-egg-laying (AEL) provided by ModEncode (Supplementary Table S6), we classified the genome into clusters. This classification was conducted across 5 kb genomic windows, following previously established methods in S2 cells (30). As a result, eight distinct clusters were identified, each defined by unique epigenetic profiles (Figure 2B). The distribution of the four LTR-RTE insertions in the G73 genome was analyzed in relation to the proportion of the genome occupied by each chromatin cluster (Supplementary Table S5). Genomic

insertions of all four LTR-RTEs were detected across all clusters; however, significant enrichment was observed in clusters 1–4 (Figure 2C). These clusters are characterized by histone marks commonly associated with open chromatin, including H3K4me1, H3K9ac, and H3K27ac. Notably, each LTR-RTE exhibited a distinct distribution pattern. Gtwin showed pronounced enrichment in clusters 1 and 3, while copia displayed marked enrichment in cluster 3. In contrast, roo and ZAM displayed a weak preference for cluster 4. This chromatin niches partitioning was confirmed by analyzing the chromatin states defined in S2 cells (30) (Supplmentary Figure S3).

Gtwin and ZAM exhibit distinct timing of genomic insertion during embryogenesis

To investigate the timing of the LTR-RTE genomic insertions during development, we focused on the sequences of events from oogenesis to embryogenesis. Oocytes undergo meiosis, and arrest at metaphase I during the final stages of oogenesis. Upon ovulation, the oocytes are released into the oviduct, where they become activated (42) (Figure 3A). This activation involves the resumption of meiosis, fertilization, and the initiation of embryogenesis. Previous studies have shown that gypsy/mdg4, an infectious, enveloped LTR-RTE belonging to the Gypsy clade, is expressed in follicular cells during oogenesis. However, its insertion into the germline genome occurs later, during embryogenesis (43). This was further supported by the rescue of the *ovoD1* mutant phenotype, which was linked to the insertion of gypsy/mdg4 into the paternal allele of *ovoD1* (44). Given these prior findings, and the facts that both gtwin and ZAM are enveloped similar to gypsy/mdg4, we hypothesized that integration of gtwin and ZAM may also occur during embryogenesis. To test this hypothesis, we sought to determine the specific stage of embryonic development during which gtwin and ZAM insertions occur. To this end, we reanalyzed single-cell ATAC-seq (sci-ATAC-seq) data from embryos spanning 0 to 16 hours AEL (34). We calculated the average chromatin accessibility within a 200 bp window centered on the gtwin or ZAM insertion sites, normalizing these values to signals from 100 randomly selected genomic regions. Our analysis revealed significant enrichment of gtwin insertion sites in open chromatin across all studied time windows, including the 14-16 h AEL window (Figure 3B), which is consistent with our previous analysis performed on 14-16h whole embryos (Figures 2B and 2C). This analysis further demonstrates that gtwin insertions are also enriched in open chromatin as early as the 0-2 h AEL window (Figure 3B). During this early stage, the embryo consists of undifferentiated syncytial nuclei, with only primordial germ cells (PGCs) beginning to cellularize and specify (Figure 3A). For ZAM insertion sites, chromatin accessibility was consistently lower than that of gtwin across all time windows. However, accessibility gradually increased throughout embryogenesis, with a marked enrichment observed during the 10-16 h AEL window (Figure 3B). At this stage, PGCs are identified (Figure 3A), suggesting that ZAM integration may occur in PGCs. To investigate this possibility, we performed ATAC-seq experiments on purified PGCs. Using a *Drosophila* strain expressing GFP-Vasa

(39), we separated GFP-positive PGCs and GFP-negative somatic cells from overnight egg collection by fluorescence-activated cell sorting (FACS) (40). ATAC-seq was performed in duplicate for both cell types. We then analyzed the averaged ATAC-seq signals within a 4 kb window centered on 101 ZAM insertion sites and compared these profiles to 100 randomly selected 4 kb genomic regions in the same cell types. As expected, the averaged ATAC-seq profiles showed a distinct peak centered on the ZAM insertion site, with the signal being more pronounced in PGCs than in somatic cells (Figure 3C). This observation indicates that ZAM insertion sites are associated with open chromatin in PGCs compared to somatic cells. Taken together, these findings suggest that gtwin and ZAM genomic insertions occur at distinct, but not mutually exclusive, stages of embryogenesis.

Additionally, we investigated whether the number of ZAM and gtwin insertions into open chromatin regions in early embryos (AEL 0-2 hr) and PGCs deviated significantly from the expected values. The expected values were calculated based on the relative size of open chromatin regions identified in the two experiments described above, compared to the whole genome (Supplementary Table S8). Insertions were then classified into three categories: embryo-specific open chromatins, PGC-specific open chromatins, and regions shared by both (Figure 3D). Gtwin insertions were found to be significantly enriched in embryo-specific and shared open chromatin regions (Figure 3D, left), while the ZAM insertions were predominantly located in PGC-specific open chromatin regions (Figure 3D, right). To refine this analysis, we isolated the PGC-specific chromatin accessibility profile (ATAC-seq PGC unique) by subtracting the ATAC-seq signal of GFP-negative somatic cells from the PGC signal of late embryos (Supplementary Table S8). This revealed that a subset of ZAM insertions were specific to late embryonic PGCs, whereas gtwin insertions were more enriched in the earlier developmental window (Figure 3D, right panel). These findings indicate a distinct distribution of ZAM and gtwin insertions within open chromatin during embryogenesis.

Gtwin preferentially inserts into the open chromatin before cellularization

To better understand the timing of gtwin insertion during germline specification, we analyzed the sci-ATAC-seq data of 0-2 hours embryos (34). In this early time window, the data revealed four distinct groups of nuclei, each characterized by unique chromatin accessibility landscapes. We calculated the average chromatin accessibility within a 4 kb window surrounding the gtwin insertion sites belonging to each cluster. Across all clusters, gtwin insertion sites were predominantly localized within open chromatin regions. In contrast, as anticipated, ZAM insertion sites were not enriched in open chromatin at this stage (Figure 4A). Peak-calling identified 76 gtwin insertions in at least one of these clusters, with 25 of them significantly enriched in all four clusters (Figure 4B). These findings suggest that most gtwin insertions occur early in embryogenesis, likely before the differentiation of nuclear clusters and likely before germ

cell specification (45). The early open chromatin state in which gtwin insertions are observed is characterized by low nucleosome occupancy and minimum higher-order chromatin structure, features typically associated with pioneer transcription factor binding. Pioneer factors can overcome nucleosome barriers and establish accessibility of cis-regulatory elements to transcription factors at a crucial step for cellular differentiation (46, 47). To identify whether specific pioneer factors are associated with gtwin insertions, we analyzed ChIP-seq data from stage 5 embryos for four known pioneer factors; GAGA Factor (GAF), Opa, Chromatin-linked Adaptor for MSL proteins (CLAMP), and Zelda (48–50). Among the 84 gtwin insertions located in open chromatin regions of stage 5 embryos, 69 were bound by at least one of these pioneer factors. However, no single factor predominated, with GAF, Opa, CLAMP, and Zelda binding sites corresponding to 32, 24, 6, and 7, respectively. These results suggest that chromatin accessibility, rather than the specific binding of pioneer factors, is the key determinant of gtwin insertion site selection during early embryogenesis.

Late ZAM insertion sites correlate with late ZAM embryonic expression

Two hypotheses could explain the late germline genomic insertion of ZAM. The first hypothesis is that the germ cells infected by ZAM particles during oogenesis remain dormant during early embryogenesis. The second hypothesis postulates the expression of a "second-wave" of ZAM, potentially facilitated by the permissive 25°C temperature at which piwi-sKD is possible. To test this possibility, we combined ZAM smiFISH and germline-specific vasa immunostaining on late embryos. High expression of ZAM was observed in the gonads of embryos laid at 25°C but not at 20°C (Figure 5A). At these late stages of embryogenesis, the PGCs have migrated away from the midgut toward the adjacent mesoderm and have become associated with somatic gonadal precursors (SGPs) (45, 51) that express traffic jam (tj) (52). As expected, ZAM expression was specifically detected in the tj-positive cells at 25°C (Figure 5B). Additionally, zygotically expressed Piwi levels were lower at 25°C compared to 20°C, as anticipated (Figure 5C). Overall, these analyses show that ZAM is expressed in tj-positive SGPs cells in embryogenesis and is sensitive to Piwi depletion. Collectively, these results suggest that ZAM insertion sites correlate with a specific chromatin accessibility landscape in late PGCs. We propose that somatic ZAM expression leads to PGC infection, with ZAM invading open genomic regions in late-stage PGCs.

Discussion

Transposable elements may alternate rapid bursts of activity and prolonged phases of repression during which their replication within the host genome is limited (53). The interactions observed today between TEs and their host genomes, as well as among TEs themselves, likely reflect the outcomes of extensive co-

549 evolution. This process has enabled the coexistence of several TE species within the same genome while
550 minimizing detrimental impacts on the host. In an attempt to describe the diversity of these interactions, we
551 took advantage of a particular *Drosophila melanogaster* laboratory strain (17) to simultaneously impair the
552 repression of several LTR-RTEs. We observed the efficient germline transposition of four active elements
553 (roo, copia, gtwin and ZAM). This approach revealed the existence of two distinct categories of LTR-RTEs:
554 ZAM and gtwin, on one hand, whose transposition is induced by the depletion of Piwi in gonadal somatic
555 tissues (Piwi-sKD), and roo and copia, on the other hand, whose activity is independent of this treatment
556 and which transpose spontaneously, even in the presence of Piwi. We found that they all displayed distinct
557 characteristics at various stages of their replication cycles, as follows.

558 Regarding expression, we observed cell-type-specific patterns for all four elements. For example, roo
559 transcripts were exclusively expressed in the germinal nurse cells, whereas ZAM, gtwin, and copia were
560 transcribed in various somatic follicular epithelial cells. Specifically, ZAM was predominantly expressed
561 at the posterior pole, gtwin was ubiquitously transcribed, and copia transcripts were expressed throughout
562 the follicular epithelium but sequestered into nuclei. Additionally, ZAM displayed a second somatic
563 expression window in the SGPs of late embryonic gonads, emphasizing its distinct temporal and spatial
564 regulation. Concerning the integration step, our approach was based on characterization of the overall
565 epigenetic specificity of their genomic insertion sites. While all four elements predominantly inserted into
566 regions of open chromatin (clusters 1 to 4), each displayed distinct preferences for specific chromatin
567 clusters associated with different histone modifications. Gtwin showed significant enrichment in clusters 1
568 and 3, which are associated with the histone modifications H3K4me3 in cluster 1 and H3K27ac and
569 H3K4me1 in cluster 3. Copia was enriched in cluster 3, whereas roo and ZAM exhibited a preference for
570 cluster 4, which is characterized by open chromatin but lacks a distinct enrichment for specific histone
571 modifications (Figure 2C).

572 Moreover, our data indicated that the preference for specific genomic insertion sites may follow the
573 differentiation of the chromatin landscape of the cells that are invaded by gtwin and ZAM at different stages
574 of the embryonic development. Indeed, for maternally deposited gtwin, a significant proportion of the
575 insertions seemed to have occurred as soon as their landing sites had begun accessible, at the very beginning
576 of embryogenesis. Conversely, consistent with the late embryonic wave of ZAM somatic expression,
577 several ZAM insertions were located within different open chromatin regions that were accessible only in
578 late embryonic germ cells. However, although the insertion of a LTR-RTE into closed chromatin is
579 generally considered unlikely, it cannot be entirely ruled out. Therefore, we cannot exclude the possibility
580 that some maternally deposited ZAM virus-like particles would have driven integration at early stages into
581 close chromatin landing sites that would open later in the gonadic PGCs. Altogether, our findings disclosed

a novel level of LTR-RTE niche partitioning, linking temporal and spatial features of the integration step of the replication cycle.

Further evidence for diversity of expression and integration niches

The diversity of ovarian expression patterns reported here (Figure 1D) has also been observed recently for 16 species of evolutionarily related LTR-RTEs (12). These different patterns in the onset of the replication cycles of this class, indicate that each LTR-RTE species has evolved specific host-TE interactions, hijacking tissue-specific transcription factors to adapt their proper expression niche to a specific cell type of the ovary. In our study, we identified a novel cell type in which ZAM is also expressed, the SGPs of late embryonic gonads. Future experiments will be necessary to determine whether the transcription factor called Pointed, which drives ZAM transcription in the posterior part of the ovarian follicular cells (54, 55), is also responsible for its expression in the SGPs.

In our study, we also revealed a novel level of LTR-RTE niche partitioning, at the integration step of the replication cycle. It is well-documented that different TE species, belonging to various classes, exhibit diverse target site preferences due to distinct transposition mechanisms. For example, DNA transposons, like the P-element, manage to create new copies by integrating near the replication origins of the *Drosophila* genome (14, 16), whereas retrotransposons Ty1 and Ty3 specifically insert into Pol III promoters of *S. cerevisiae* (13, 56). Recently, it has been suggested that LTR-RTEs are rather attracted by open chromatin of active genes, whereas LINE elements, such as the I-element, target AT-rich sites and tend to integrate near telomeres (14). These TE-specific host affinities have been described to depend on the enzymes driving their integration such as transposases and integrases. We found here that even LTR-RTEs of the same class, despite using the same integration mechanism, preferentially integrate into open chromatin domains harboring distinct chromatin features (Figure 2C). This finding suggests that each LTR-RTE species has evolved specific interactions between its integrase and host co-factors (DNA- and/or chromatin-binding proteins) providing different affinities for specific genetic and epigenetic marks.

Note that, although specific for each LTR-RTE, their preferred epigenetic landscapes share a common feature, open chromatin, a permissive location for subsequent efficient transcription. These similarities might be considered as cases of concerted evolution by sharing general molecular mechanisms of targeting. A famous mechanism of decompacted chromatin targeting operates *via* the histone H4 tail that can no longer be targeted by HIV when embedded in closed chromatin (57, 58).

Our data also suggest that the specific integration of LTR-RTEs into distinct epigenetically defined domains might, at least partly, result from different integration timings during development. Strikingly, gtwin and

ZAM landing site landscapes correlated with chromatin accessibility data sets extracted from early and late stages of embryogenesis, respectively. The hypothesis, assuming replication cycles with different timings of integration, is supported by the second wave of ZAM expression observed later in embryonic gonads. An obvious difference between these two putative cellular integration niches concerns their ability to proliferate. It is indeed worth noting that, unlike early embryonic nuclei that are rapidly cycling, gonadic germ cells are no longer dividing. As it has been suggested for HIV, further experiments will be necessary to know whether gtwin and ZAM integrases have distinct abilities to be imported into non dividing nuclei.

A complex interplay of positive and negative selection forces, applied to TEs and their hosts, likely lead to niche partitioning.

By studying the simultaneous replication of four LTR-RTEs in the *Drosophila* germline, we observed distinct patterns suggesting that these LTR-RTEs occupy different ecological niches within the TE community. Here, we briefly speculate about the putative selective forces that might have led to host-TE and TE-TE coexistence *via* niche partitioning.

First, we can notice that the four LTR-RTEs species are expressed in gonadal tissues, the only host compartment supporting vertical transmission of the new TE copies. On the contrary, replication in non-gonadal tissues is not only useless for the TE replication but could have been counter-selected by the host as a possible cause of diseases like cancer and aging-related decline (59, 60). A second type of selective pressure might have prevented toxic TE expression (60) in the germline stem cells, the immortal cell lineage of the gonad. That is probably why ZAM and gtwin are expressed in differentiated somatic gonadal cells, while roo, despite being a germline-specific TE, is expressed in nurse cells, which are differentiated germ cells destined to disappear at the end of oogenesis. Third, on one hand, the new TE copies need to be inserted into the germinal genome, but, on the other hand, the resulting DNA damage may be even more deleterious for the germline survival than the toxicity of the expression step. As a possible trade-off, integration is delayed until the DNA damage-tolerant embryonic stage of development (61), followed by larval stages where germ cell division may compensate for previous cell death (62). Fourth, further research is needed to characterize the putatively detrimental phenotypic effects of the TE insertions we studied and determine if their preferred integration sites correspond to safe havens within the host genome. Similarly, regarding TE-TE interactions, it is unknown whether the TE-specificity of these integration niches results from detrimental fitness effects of competition between different TE species for common insertion sites. Finally, our non-overlapping TE expression patterns are in agreement with previous observations (12) suggesting that such a competition between somatic TEs might have led to expression niche partitioning.

In conclusion, TE niche partitioning highlights the complex interplay of positive and negative selection forces applied to TEs and their hosts leading to their stable coexistence.

Data availability

Long reads sequencing data previously published and presented in this study have been deposited at ENA (https://www.ebi.ac.uk/ena) under the accession numbers ERP122844 and PRJEB75331 respectively. The source code of TrEMOLO as well as all the accessory codes are available at https://github.com/DrosophilaGenomeEvolution/TrEMOLO. The ATAC-seq, and small-RNA-seq raw data are available on GEO under accession number: GSE274394.

Acknowledgments

We thank Mikiko C. Siomi for participating in the discussions and helping to edit the manuscript, Bernd Schuettengruber for comments on the manuscript, Callum Burnard and Gonzalo Sabaris for scientific discussions. We acknowledge the ISO 9001 certified IRD itrop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported in this paper (URLs: https://bioinfo.ird.fr/ and http://www.southgreen.fr); the Genotoul platform (https://genotoul.fr/) and (https://www.france-bioinformatique.fr/) for providing calculation time on their servers; BioCampus MRI platform for microscopy, Drosophila core facilities and MGX platform. We thank Akira Nakamura for the drosophila line *w* ; *vas::EGFP*, Makoto Hayashi for his help on the PGC isolation protocol, Felicia Leccia from the MRI-Cyto IRMB Cytometry platform for the cell sorting and Bernd Schuettengruber for his help on the ATAC-seq experiments. We thank BioRender.com for the drawings used in the creation of the illustrations included in this article.

Author contributions

Conception: S.C; A.P.; Computational analysis of NGS and genomics data, M.M.; M.V.; Statistics, D.G.; M.V.; Experiments C.G, M.L., B.M., M.V.; Methodology and analyses: B.M.; C.G.; A.P.; S.C.; Supervision, C.G.; S.C.; Visualization: B.M.; D.G.; M.V.; Writing: C.G.; A.P.; S.C.; Funding & infrastructure: S.C.

Declaration of interests

The authors declare no competing interests.

Funding

This research was funded by the Fondation pour la Recherche Médicale, grant number "EQU202303016294" and French National Research Agency "ANR-20-CE12-0015-01" to S.C., the CNRS and the University of Montpellier. M.V. was funded by CNRS – University of Tokyo "Excellence Science" Joint Research Program and supported by the Fondation ARC pour la recherche sur le cancer.

References

1. Feschotte,C. (2023) Transposable elements: McClintock's legacy revisited. *Nat. Rev. Genet.*, **24**, 797–800.

2. Payer,L.M. and Burns,K.H. (2019) Transposable elements in human genetic disease. *Nat. Rev. Genet.*, **20**, 760–772.

3. Lawlor,M.A. and Ellison,C.E. (2023) Evolutionary dynamics between transposable elements and their host genomes: mechanisms of suppression and escape. *Curr. Opin. Genet. Dev.*, **82**, 102092.

4. Czech,B., Munafò,M., Ciabrelli,F., Eastwood,E.L., Fabry,M.H., Kneuss,E. and Hannon,G.J. (2018) piRNA-Guided Genome Defense: From Biogenesis to Silencing. *Annu. Rev. Genet.*, **52**, 131–157.

5. Sato,K. and Siomi,M.C. (2018) Two distinct transcriptional controls triggered by nuclear Piwi-piRISCs in the Drosophila piRNA pathway. *Curr. Opin. Struct. Biol.*, **53**, 69–76.

6. Osumi,K., Sato,K., Murano,K., Siomi,H. and Siomi,M.C. (2019) Essential roles of Windei and nuclear monoubiquitination of Eggless/SETDB1 in transposon silencing. *EMBO Rep.*, **20**, e48296.

7. Sienski,G., Donertas,D. and Brennecke,J. (2012) Transcriptional silencing of transposons by piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*, **151**, 964–80.

8. Cosby,R.L., Chang,N.-C. and Feschotte,C. (2019) Host–transposon interactions: conflict, cooperation, and cooption. *Genes Dev.*, **33**, 1098–1116.

9. Mérel,V., Boulesteix,M., Fablet,M. and Vieira,C. (2020) Transposable elements in Drosophila. *Mob. DNA*, **11**, 23.

10. Malik,H.S., Henikoff,S. and Eickbush,T.H. (2000) Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses. *Genome Res.*, **10**, 1307–1318.

11. Nefedova,L. and Kim,A. (2017) Mechanisms of LTR-Retroelement Transposition: Lessons from Drosophila melanogaster. *Viruses*, **9**, 81.

12. Senti,K.-A., Handler,D., Rafanel,B., Kosiol,C., Schloetterer,C. and Brennecke,J. (2023) Functional Adaptations of Endogenous Retroviruses to the Drosophila Host Underlie their Evolutionary Diversification. 10.1101/2023.08.03.551782.

13. Sultana,T., Zamborlini,A., Cristofari,G. and Lesage,P. (2017) Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.*, **18**, 292–308.

14. Cao,J., Yu,T., Xu,B., Hu,Z., Zhang,X., Theurkauf,W.E. and Weng,Z. (2023) Epigenetic and chromosomal features drive transposon insertion in Drosophila melanogaster. *Nucleic Acids Res.*, 10.1093/nar/gkad054.

15. Wang,L., Dou,K., Moon,S., Tan,F.J. and Zhang,Z.Z. (2018) Hijacking Oogenesis Enables Massive Propagation of LINE and Retroviral Transposons. *Cell*, 10.1016/j.cell.2018.06.040.

16. Spradling,A.C., Bellen,H.J. and Hoskins,R.A. (2011) Drosophila P elements preferentially transpose to replication origins. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 15948–15953.

17. Barckmann,B., El-Barouk,M., Pélisson,A., Mugat,B., Li,B., Franckhauser,C., Fiston Lavier,A.-S., Mirouze,M., Fablet,M. and Chambeyron,S. (2018) The somatic piRNA pathway controls germline transposition over generations. *Nucleic Acids Res.*, **46**, 9524–9536.

18. Kina,H., Yoshitani,T., Hanyu-Nakamura,K. and Nakamura,A. (2019) Rapid and efficient generation of GFP-knocked-in Drosophila by the CRISPR-Cas9-mediated genome editing. *Dev. Growth Differ.*, **61**, 265–275.

19. Fabry,M.H., Falconio,F.A., Joud,F., Lythgoe,E.K., Czech,B. and Hannon,G.J. (2021) Maternally inherited piRNAs direct transient heterochromatin formation at active transposons during early Drosophila embryogenesis. *eLife*, **10**, e68573.

20. Mohamed,M., Dang,N.T.-M., Ogyama,Y., Burlet,N., Mugat,B., Boulesteix,M., Mérel,V., Veber,P., Salces-Ortiz,J., Severac,D., *et al.* (2020) A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore. *Cells*, **9**, 1776.

21. Mohamed,M., Sabot,F., Varoqui,M., Mugat,B., Audouin,K., Pélisson,A., Fiston-Lavier,A.-S. and Chambeyron,S. (2023) TrEMOLO: accurate transposable element allele frequency estimation using long-read sequencing data combining assembly and mapping-based approaches. *Genome Biol.*, **24**, 1–20.

22. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

23. Brennecke,J., Aravin,A.A., Stark,A., Dus,M., Kellis,M., Sachidanandam,R. and Hannon,G.J. (2007) Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell*, **128**, 1089–1103.

24. Czech,B., Malone,C.D., Zhou,R., Stark,A., Schlingeheyde,C., Dus,M., Perrimon,N., Kellis,M., Wohlschlegel,J.A., Sachidanandam,R., *et al.* (2008) An endogenous small interfering RNA pathway in Drosophila. *Nature*, **453**, 798–802.

25. Tsanov,N., Samacoits,A., Chouaib,R., Traboulsi,A.-M., Gostan,T., Weber,C., Zimmer,C., Zibara,K., Walter,T., Peter,M., *et al.* (2016) smiFISH and FISH-quant – a flexible single RNA detection approach with super-resolution capability. *Nucleic Acids Res.*, **44**, e165.

745 26. Gunawan,F., Arandjelovic,M. and Godt,D. (2013) The Maf factor Traffic jam both enables
746 and inhibits collective cell migration in Drosophila oogenesis. *Development*, **140**, 2808–
747 2817.

748 27. Amemiya,H.M., Kundaje,A. and Boyle,A.P. (2019) The ENCODE Blacklist: Identification of
749 Problematic Regions of the Genome. *Sci. Rep.*, **9**, 9354.

750 28. Martin,F.J., Amode,M.R., Aneja,A., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Becker,A.,
751 Bennett,R., Berry,A., Bhai,J., *et al.* (2023) Ensembl 2023. *Nucleic Acids Res.*, **51**, D933–
752 D941.

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

755 30. Kharchenko,P.V., Alekseyenko,A.A., Schwartz,Y.B., Minoda,A., Riddle,N.C., Ernst,J.,
756 Sabo,P.J., Larschan,E., Gorchakov,A.A., Gu,T., *et al.* (2011) Comprehensive analysis of
757 the chromatin landscape in Drosophila. *Nature*, **471**, 480–485.

758 31. Hinrichs,A.S. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids*
759 *Res.*, **34**, D590–D598.

760 32. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and
761 BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.

762 33. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C.,
763 Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS).
764 *Genome Biol.*, **9**, R137.

765 34. Calderon,D., Blecher-Gonen,R., Huang,X., Secchia,S., Kentro,J., Daza,R.M., Martin,B.,
766 Dulja,A., Schaub,C., Trapnell,C., *et al.* (2022) The continuum of Drosophila embryonic
767 development at single-cell resolution. *Science*, **377**, eabn5800.

768 35. Shigenobu,S., Arita,K., Kitadate,Y., Noda,C. and Kobayashi,S. (2006) Isolation of germline
769 cells from Drosophila embryos by flow cytometry. *Dev. Growth Differ.*, **48**, 49–57.

770 36. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N.,
771 Schwikowski,B. and Ideker,T. (2003) Cytoscape: A Software Environment for Integrated
772 Models of Biomolecular Interaction Networks. *Genome Res.*, **13**, 2498–2504.

773 37. Brennecke,J., Malone,C.D., Aravin,A.A., Sachidanandam,R., Stark,A. and Hannon,G.J.
774 (2008) An epigenetic role for maternally inherited piRNAs in transposon silencing.
775 *Science*, **322**, 1387–1392.

776 38. Chambeyron,S., Popkova,A., Payen-Groschene,G., Brun,C., Laouini,D., Pelisson,A. and
777 Bucheton,A. (2008) piRNA-mediated nuclear accumulation of retrotransposon transcripts
778 in the Drosophila female germline. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 14964–14969.

779 39. Yoth,M., Maupetit-Méhouas,S., Akkouche,A., Gueguen,N., Bertin,B., Jensen,S. and
780 Brasset,E. (2023) Reactivation of a somatic errantivirus and germline invasion in
781 Drosophila ovaries. *Nat. Commun.*, **14**, 6096.

40. Kofler,R. (2019) Dynamics of Transposable Element Invasions with piRNA Clusters. *Mol. Biol. Evol.*, **36**, 1457–1472.

41. Petrov,D.A., Fiston-Lavier,A.-S., Lipatov,M., Lenkov,K. and González,J. (2011) Population genomics of transposable elements in Drosophila melanogaster. *Mol. Biol. Evol.*, **28**, 1633–1644.

42. Berg,C., Sieber,M. and Sun,J. (2023) Finishing the egg. *Genetics*, 10.1093/genetics/iyad183.

43. Prud'homme,N., Gans,M., Masson,M., Terzian,C. and Bucheton,A. (1995) Flamenco, a gene controlling the gypsy retrovirus of Drosophila melanogaster. *Genetics*, **139**, 697–711.

44. Mével-Ninio,M., Mariol,M.C. and Gans,M. (1989) Mobilization of the gypsy and copia retrotransposons in Drosophila melanogaster induces reversion of the ovoD dominant female-sterile mutations: molecular analysis of revertant alleles. *EMBO J*, **8**, 1549–1558.

45. Starz-Gaiano,M. and Lehmann,R. (2001) Moving towards the next generation. *Mech. Dev.*, **105**, 5–18.

46. Barral,A. and Zaret,K.S. (2024) Pioneer factors: roles and their regulation in development. *Trends Genet. TIG*, **40**, 134–148.

47. Zaret,K.S. and Mango,S.E. (2016) Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr. Opin. Genet. Dev.*, **37**, 76–81.

48. Harrison,M.M., Li,X.-Y., Kaplan,T., Botchan,M.R. and Eisen,M.B. (2011) Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition. *PLOS Genet.*, **7**, e1002266.

49. Gaskill,M.M., Gibson,T.J., Larson,E.D. and Harrison,M.M. (2021) GAF is essential for zygotic genome activation and chromatin accessibility in the early Drosophila embryo. *eLife*, **10**, e66668.

50. Urban,J., Kuzu,G., Bowman,S., Scruggs,B., Henriques,T., Kingston,R., Adelman,K., Tolstorukov,M. and Larschan,E. (2017) Enhanced chromatin accessibility of the dosage compensated Drosophila male X-chromosome requires the CLAMP zinc finger protein. *PLoS ONE*, **12**, e0186855.

51. Santos,A.C. and Lehmann,R. (2004) Germ Cell Specification and Migration in Drosophila and beyond. *Curr. Biol.*, **14**, R578–R589.

52. Li,M.A., Alls,J.D., Avancini,R.M., Koo,K. and Godt,D. (2003) The large Maf factor Traffic Jam controls gonad morphogenesis in Drosophila. *Nat Cell Biol*, **5**, 994–1000.

53. Kofler,R., Nolte,V. and Schlötterer,C. (2015) Tempo and Mode of Transposable Element Activity in Drosophila. *PLOS Genet.*, **11**, e1005406.

54. Leblanc,P., Desset,S., Giorgi,F., Taddei,A.R., Fausto,A.M., Mazzini,M., Dastugue,B. and Vaury,C. (2000) Life Cycle of an Endogenous Retrovirus,ZAM, in Drosophila melanogaster. *J. Virol.*, **74**, 10658–10669.

55. Meignin,C., Dastugue,B. and Vaury,C. (2004) Intercellular communication between germ line and somatic line is utilized to control the transcription of ZAM, an endogenous retrovirus from Drosophila melanogaster. *Nucleic Acids Res.*, **32**, 3799–3806.

56. Bridier-Nahmias,A., Tchalikian-Cosson,A., Baller,J.A., Menouni,R., Fayol,H., Flores,A., Saïb,A., Werner,M., Voytas,D.F. and Lesage,P. (2015) An RNA polymerase III subunit determines sites of retrotransposon integration. *Science*, **348**, 585–588.

57. Benleulmi,M.S., Matysiak,J., Robert,X., Miskey,C., Mauro,E., Lapaillerie,D., Lesbats,P., Chaignepain,S., Henriquez,D.R., Calmels,C., *et al.* (2017) Modulation of the functional association between the HIV-1 intasome and the nucleosome by histone amino-terminal tails. *Retrovirology*, **14**, 54.

58. Lagadec,F., Parissi,V. and Lesbats,P. (2022) Targeting the Nucleosome Acidic Patch by Viral Proteins: Two Birds with One Stone? *mBio*, **13**, e01733-21.

59. Burns,K.H. (2017) Transposable elements in cancer. *Nat. Rev. Cancer*, **17**, 415–424.

60. Dubnau,J. (2018) The Retrotransposon storm and the dangers of a Collyer's genome. *Curr. Opin. Genet. Dev.*, **49**, 95–105.

61. Sullivan,W., Daily,D.R., Fogarty,P., Yook,K.J. and Pimpinelli,S. (1993) Delays in anaphase initiation occur in individual nuclei of the syncytial Drosophila embryo. *Mol. Biol. Cell*, **4**, 885–896.

62. Gilboa,L. and Lehmann,R. (2006) Soma-germline interactions coordinate homeostasis and growth in the Drosophila gonad. *Nature*, **443**, 97–100.

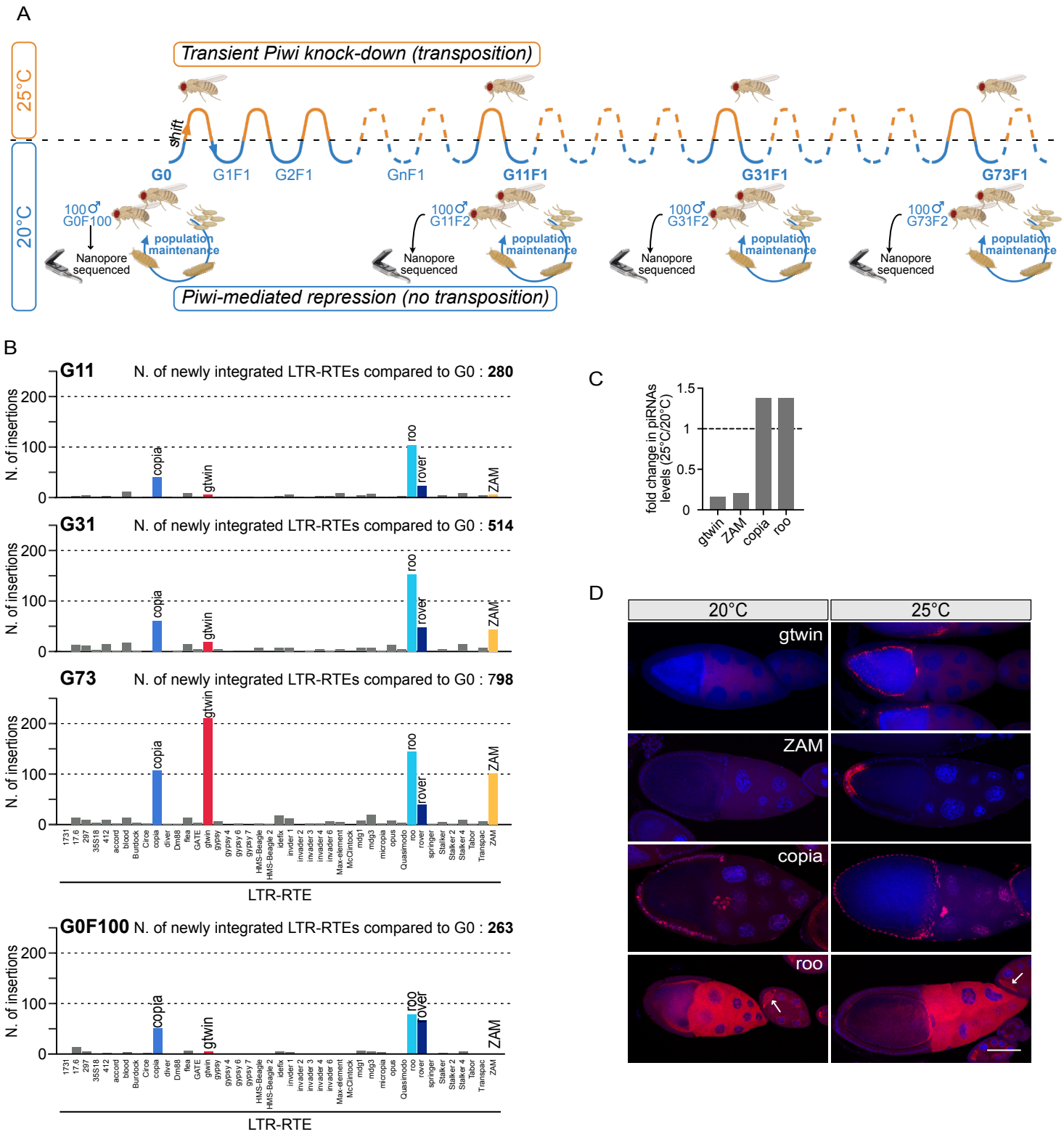


Figure 1: Four LTR-RTEs integrate into the Drosophila germline after successive generations of somatic Piwi knockdown (A) Schematic representation of the temperature change from 20°C to 25°C applied to adult flies for 5 days at each generation to induce a transient somatic knockdown of Piwi (Piwi-sKD) followed by constant maintenance at 20°C until the next generation. G0F100 corresponds to a sub-population of the initial G0 parental strain that was constantly kept at 20°C for 100 generations. During the successive Piwi-sKDs, three large populations corresponding to the offsprings of flies at generations 11, 31 and 73 (G11F1, G31F1, G73F1) were isolated and kept at 20°C. The offsprings of these isolated populations (G11F2, G31F2 and G73F2) were raised at 20°C and were sequenced using Nanopore technology. (B) Quantification of the new LTR-RTE insertions annotated in G11F2, G31F2, G73F2 and in the control population G0F100, as compared to the initial parental G0 strain. Total number of newly integrated LTR-RTEs is indicated for each generation in bold at the top of each chart. (C) Bar plots showing the fold change in antisense piRNA reads, normalized to the total piRNA reads, in Piwi-sKD ovaries (25°C) compared to control ovaries (20°C) for the four LTR-RTEs (gtwin, ZAM, copia, roo). (D) Representative images of stage 10 ovarian expression patterns obtained for gtwin, ZAM, copia and roo LTR-RTEs by smiFISH (in red) at the non-permissive temperature, 20°C, or after 5 days at the permissive temperature, 25°C (Piwi-sKD). DNA was stained with 4,6-diamidino-2-phenylindole (DAPI; blue). Bar represents 50 µm.

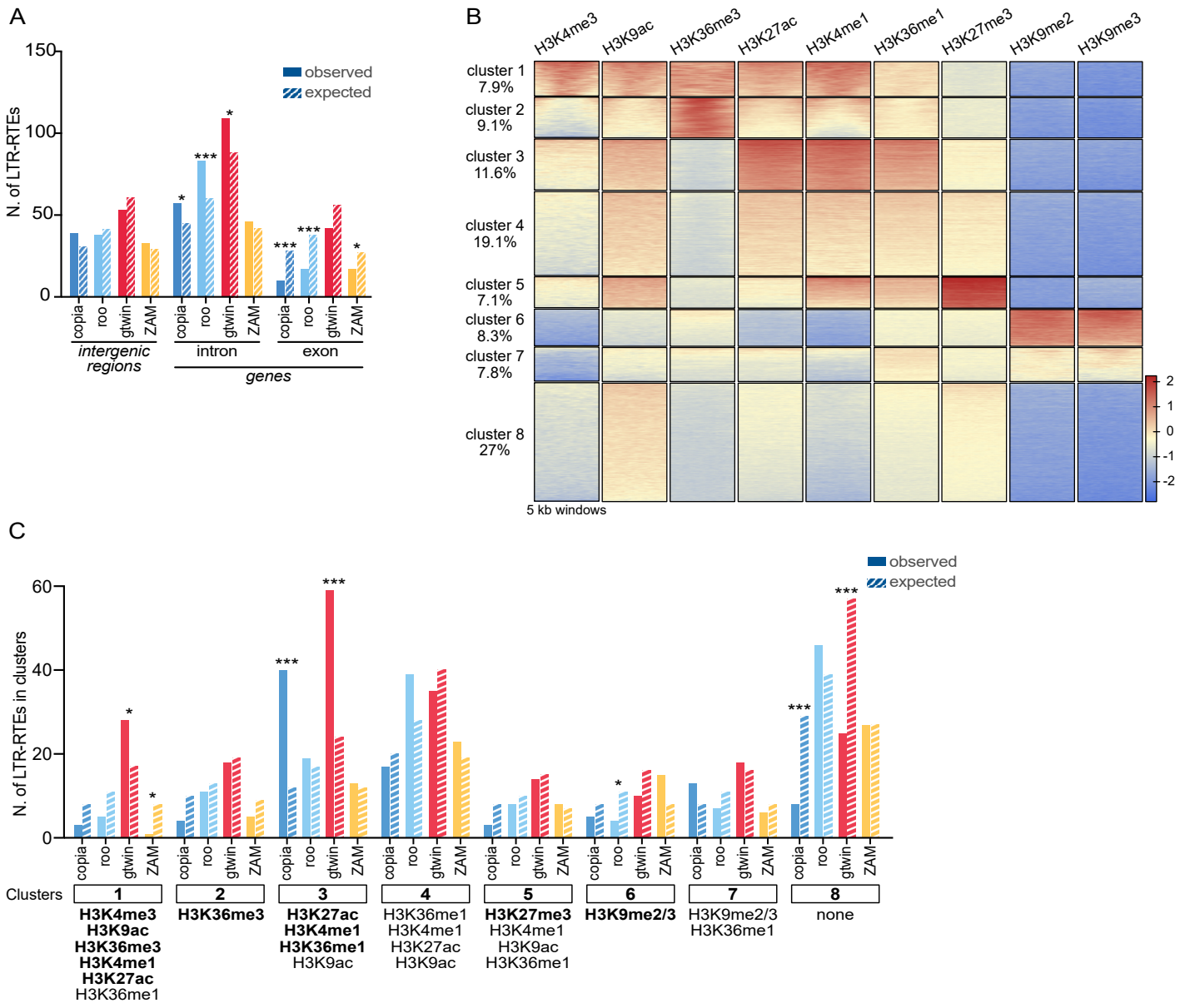


Figure 2: Each LTR-RTE species has its own specific chromatin domain preferences for genomic integration
(A) Bar plots showing the observed (filled bars) and expected (dashed bars) numbers of new LTR-RTE insertions in the intergenic, intronic and exonic regions of the genome. Expected values were calculated based on the proportional size of each genomic region (Supplementary Table S5) and the total number of new insertions identified for each LTR-RTE species. Statistical significance was assessed using binomial tests corrected with the Benjamini–Hochberg procedure to control the false discovery rate (FDR); p-values: * < 0.05, ** < 0.005, *** < 0.001. (B) Heatmap illustrating genome-wide clustering of nine post-translational histone modifications based on ChIP-seq data from 14–16 h Drosophila embryos, segmented into non overlapping 5 kb genomic bins. This analysis identified 8 distinct clusters, each representing a defined proportion of the genome (indicated on the left). The intensity of ChIP-seq signal for each histone modification is displayed using a color gradient (shown at the bottom right) with red indicating an enrichment and blue indicating a depletion. (C) Bar plots displaying the observed (filled bars) and expected (dashed bars) numbers of LTR-RTE insertions for each species in each cluster, as defined in panel B. Expected values were calculated similarly as in panel A considering the genomic proportion of each chromatin cluster (Supplementary Table S5). Statistical significances (p-values) were calculated using binomial tests corrected by the Benjamini–Hochberg step up procedure to control the false discovery rate; p-value: * < 0.05, ** < 0.005, *** < 0.001.

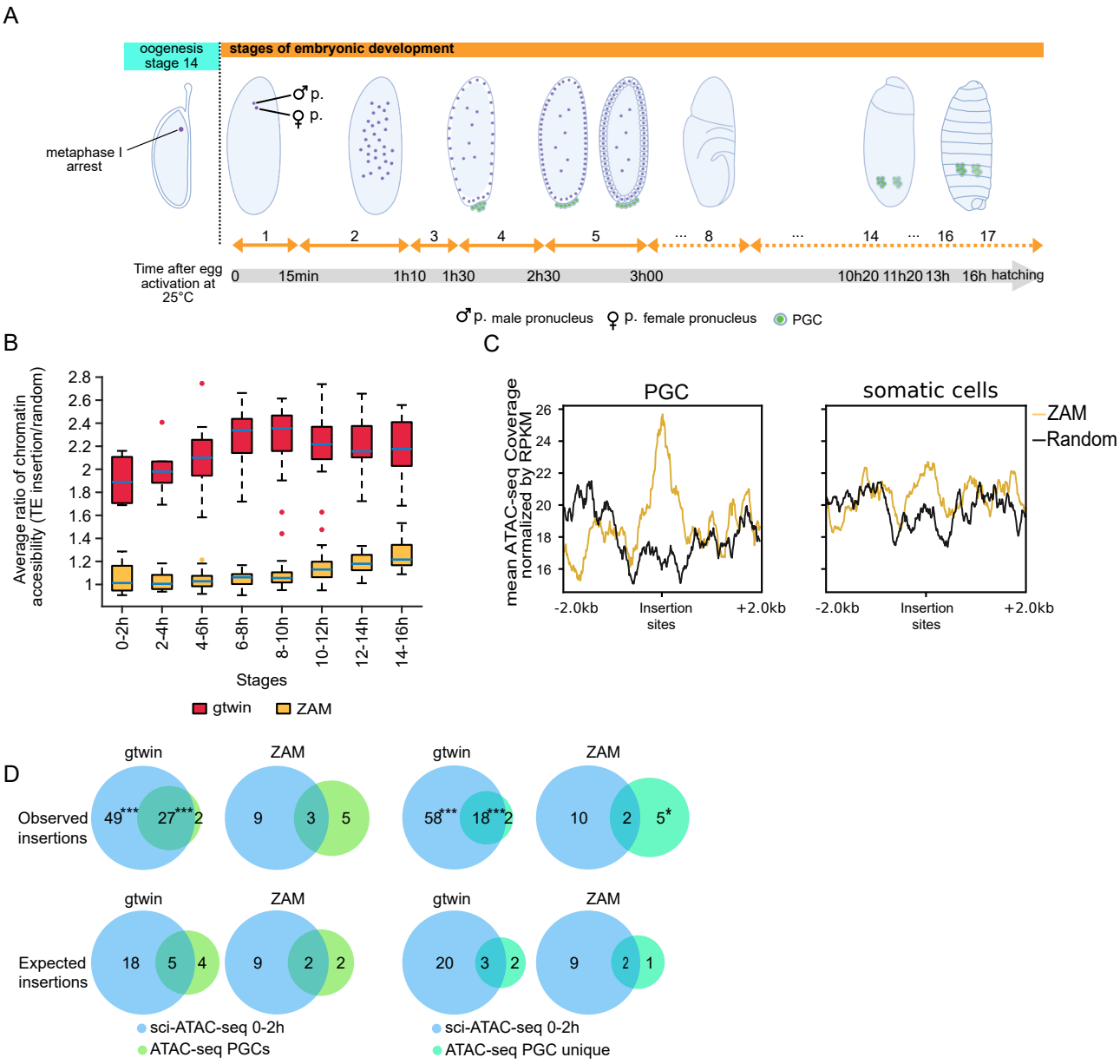


Figure 3: Differential timing of integration during embryogenesis
(A) Schematic representation of the last stage of oogenesis containing the arrested oocyte in metaphase I and the different stages of Drosophila embryonic development before hatching. Time after egg laying (AEL) is indicated at the bottom. Primordial Germ Cells (PGC) are in green. (B) Boxplots representing the temporal kinetics of the average chromatin accessibility around gtwin (red) and ZAM (orange) insertions, relative to random profiles. For each time window, the ratio of accessibility is defined as the average sci-ATAC-seq signal of the 200 bp windows centered on LTR-RTE insertions for each defined cluster (at each time window) divided by the average signal obtained for 100 randomly selected 200 bp windows in the same cluster. (C) Metaplots showing mean ATAC-seq signals within 4 kb windows centered on 101 ZAM insertion sites (yellow) and 100 random insertion sites (black). The signals are averaged from two replicates and normalized by coverage (RPKM: Reads Per Kilobase per Million mapped reads). The left panel shows data from PGCs, while the right panel displays data from somatic cells, both sorted out of overnight embryos. (D) Venn diagrams illustrating the distribution of observed (top) and expected (bottom) gtwin or ZAM insertions in the chromatin accessibility domains detected in 0–2h embryos (sci-ATAC-seq pooled data, in blue) and in primordial germ cells (PGCs) of late embryos (ATAC-seq, in green). The numbers are those of the insertions that are within stage-specific ATAC-seq peaks as well as those that are located in chromatin domains that are accessible at both stages of embryonic development (overlaps). The right panels (PGC unique) are missing those ATAC-seq PGC peaks that are also present in the somatic cells of the corresponding late embryos. The expected values are based on random distribution, adjusted to the size of each defined regions in the whole genome (Supplementary Table S8).
Significant enrichments in the observed insertions are indicated with asterisks. p-value: * < 0.05, ** < 0.005, *** < 0.001.

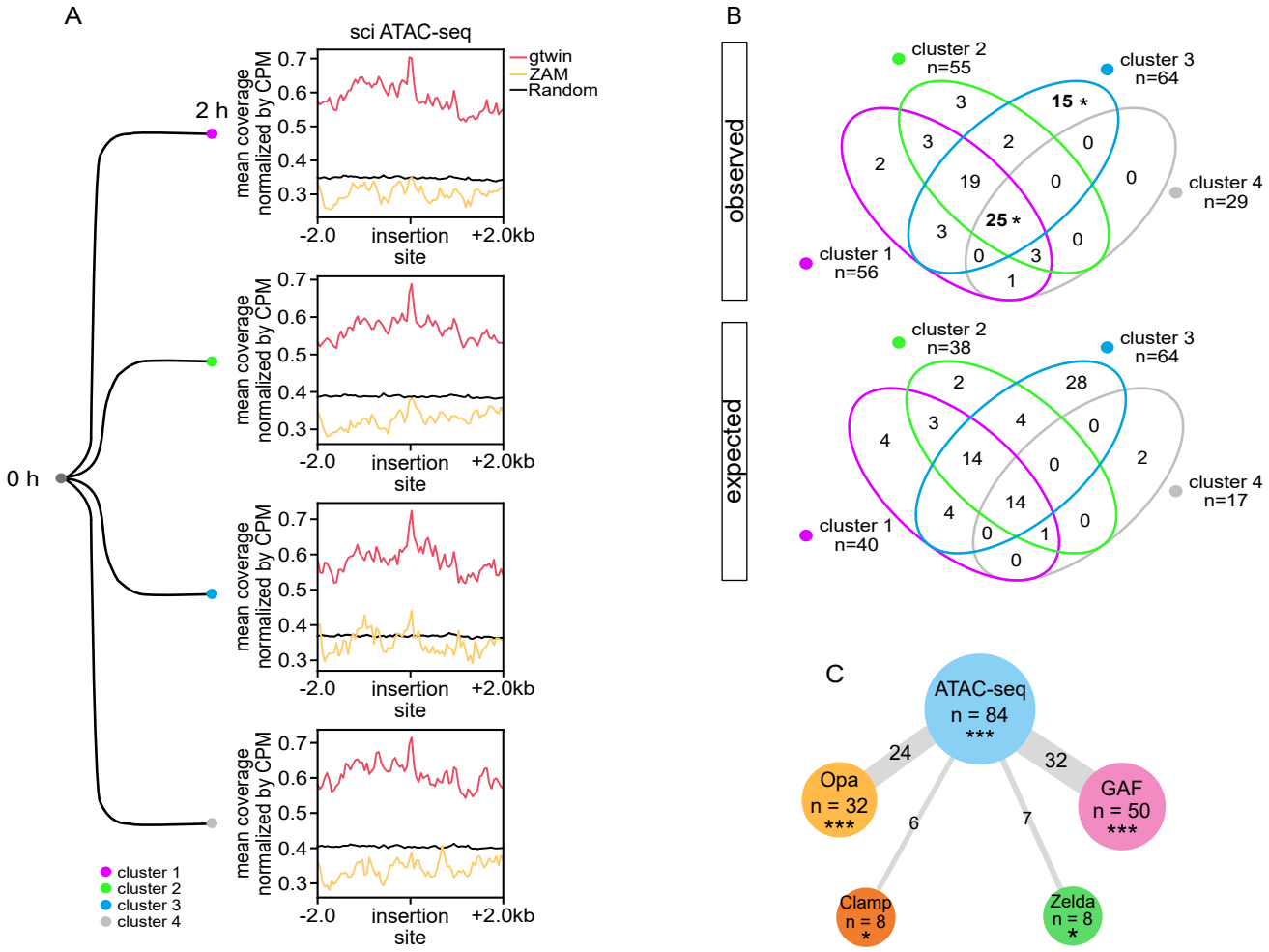


Figure 4: gtwin preferentially inserts into the open chromatin before cellularization
(A) Metaplots depicting mean coverage of single-cell ATAC-seq (sci-ATAC-seq) in 4 kb windows centered on 210 gtwin (red), 101 ZAM (yellow) or 100 random insertion sites (black) in the 4 distinct groups of nulei previously identified in 0-2 h embryos (34). Values were normalized by Counts Per Million (CPM) unique mapped reads. (B) Venn diagram highlighting the distribution of the 76 gtwin insertions sites enriched in sci-ATAC-seq signal among the 4 clusters. The numbers represent the observed (top) and expected (bottom) count distribution of gtwin insertions in each cluster. The expected values (bottom) are based on random distribution, adjusted for the size of each subset. Significant enrichment or depletion of the insertions observed in two subsets is marked with asterisk. p-value: * < 0.05. (C) Overlap of pioneer-factor-rich and sci-ATAC-seq-rich gtwin insertion sites in stage 5 embryos. The size of each circle is scaled by the number of gtwin insertions that are enriched for accessible chromatin (blue), GAF (pink), Zelda (green), Clamp (orange) and Opa (yellow). The number indicated in the edges corresponds to the number of gtwin insertions possessing both features. Expected values were calculated according to the proportion of ChIP-seq signal of the protein of interest and the total number of new gtwin insertions. P-values were calculated using binomial tests corrected by Benjamini–Hochberg step up procedure to control the false discovery rate; p-value: * < 0.05, ** < 0.005, *** < 0.001.

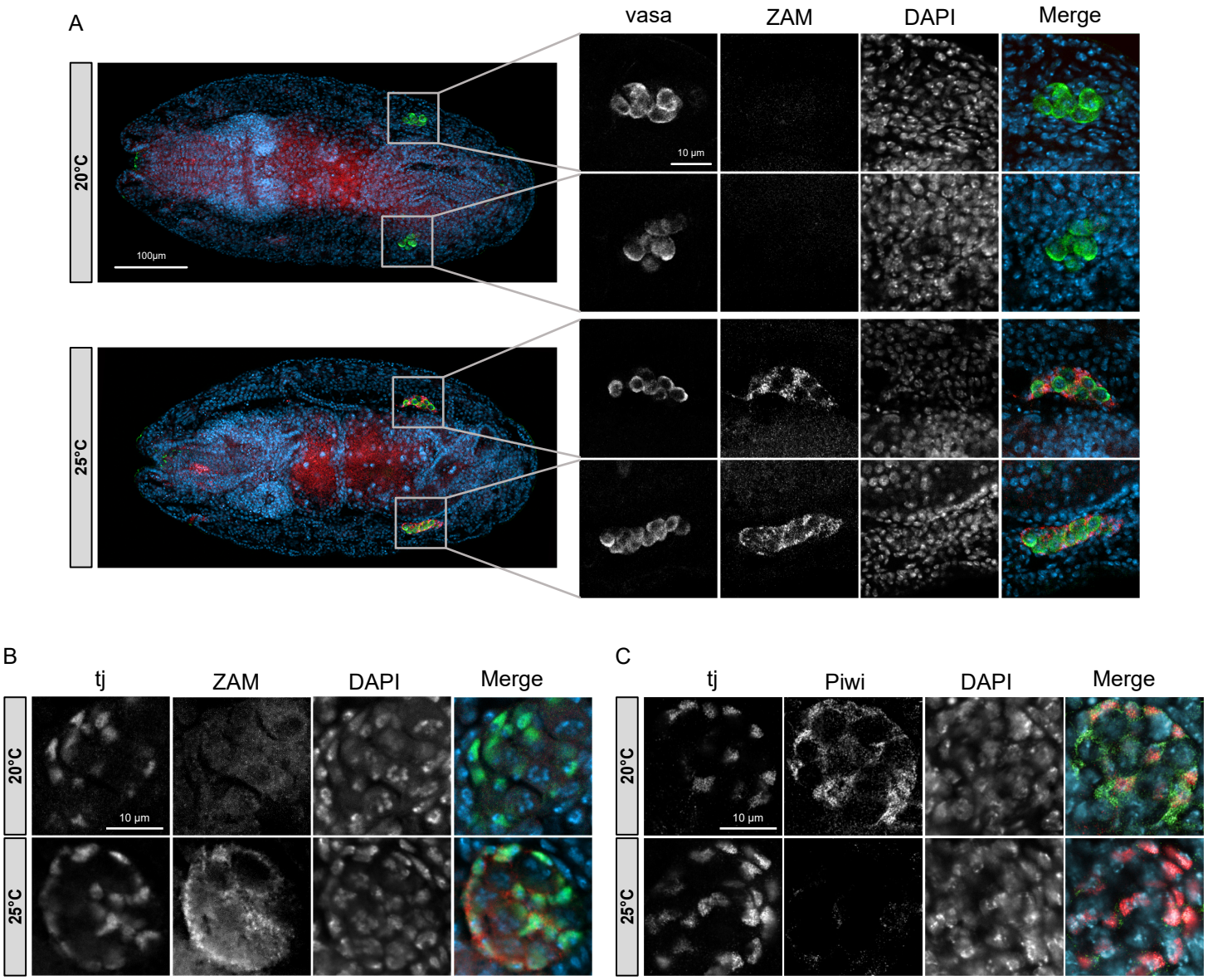
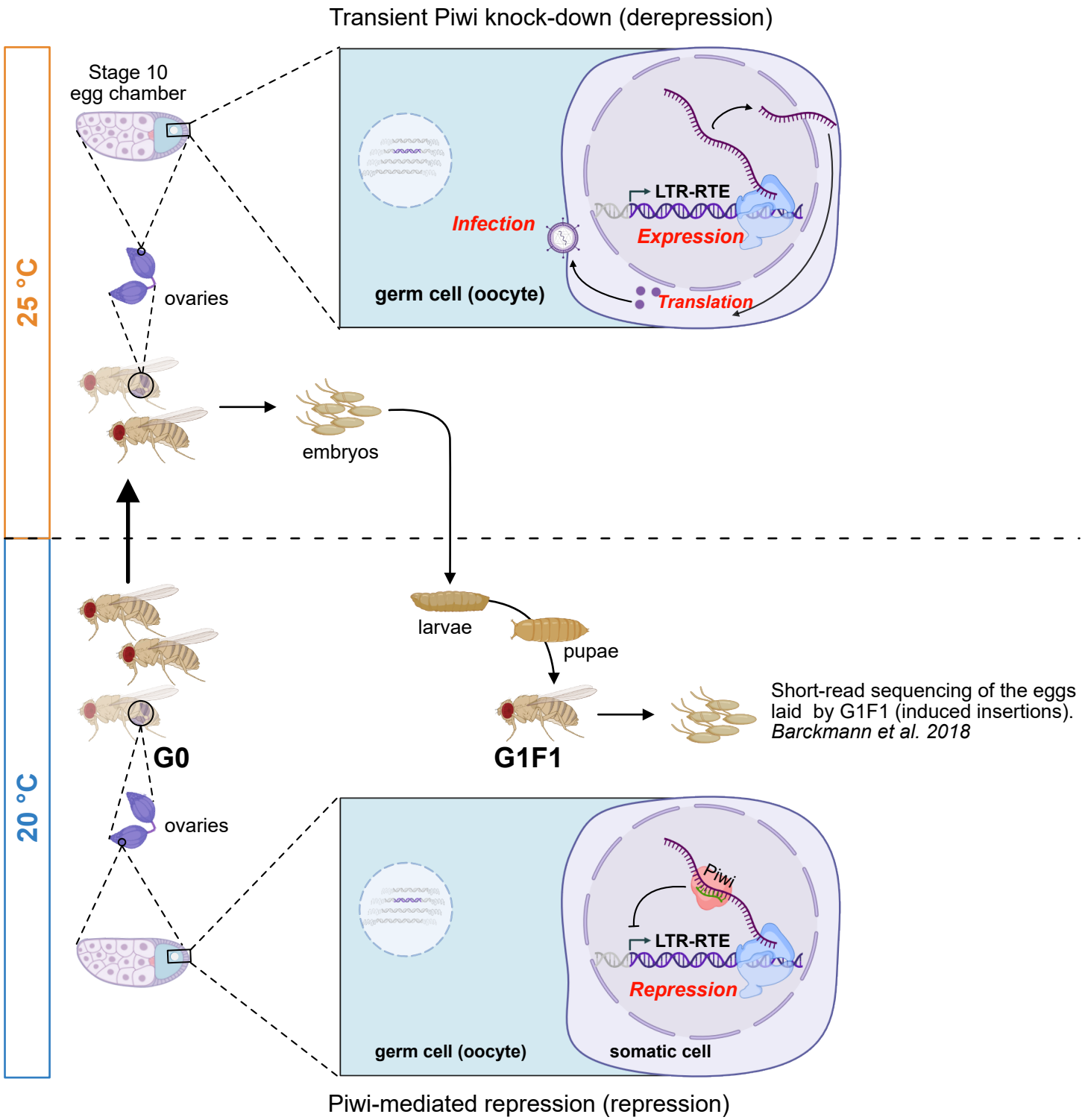
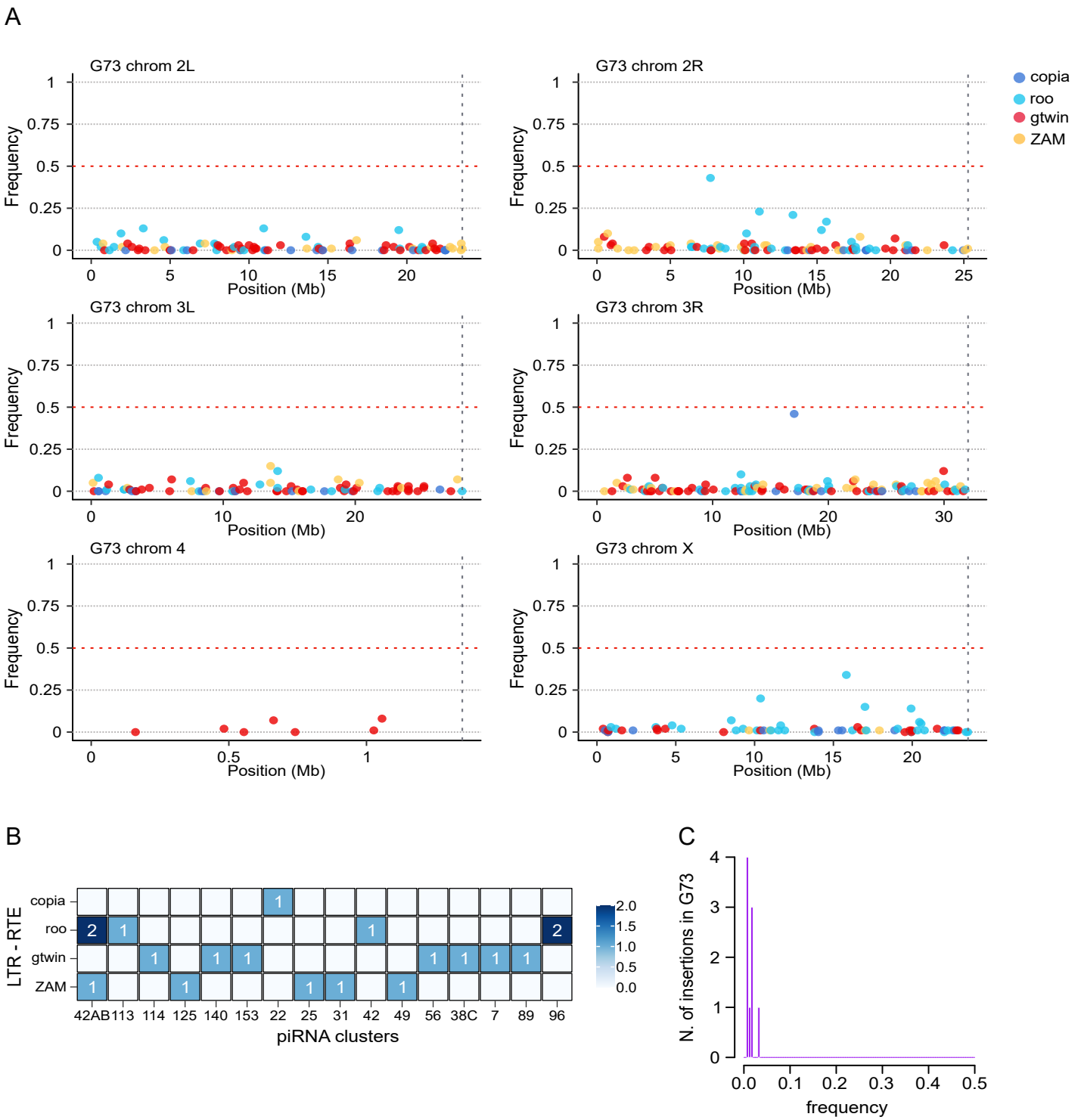


Figure 5: Chromatin accessibility of ZAM insertion sites correlates with ZAM late embryonic expression
(A) Vasa immunostaining combined with ZAM smiFISH on 12- to 16-h whole-mount G73 embryos at 20°C (upper panel) and 25°C (lower panel). The right panel shows higher magnification of embryonic gonads. Anti-vasa antibody (green) labelled the primordial germ cells (PGCs) of the gonads. ZAM transcripts labelled in red are detected in cells surrounding vasa-positive cells at 25°C (lower panels). DNA is labelled with 4,6-diamidino-2-phenylindole (DAPI; blue).
(B) Traffic jam immunostaining combined with ZAM smiFISH of gonadal cells of 12- to 16-h G73 embryos at 20°C (upper panel), 25°C (lower panel). The SPGs labelled in green are tj-positive cells.
ZAM transcripts labelled in red are detected in tj-positive cells at 25°C (lower panel). DNA is labelled with 4,6-diamidino-2-phenylindole (DAPI; blue). (C) Immunostaining of gonadal cells from 12-16 h embryos laid by G73 females crossed with males expressing a GFP-tagged version of Piwi, at 20°C (upper panel) or 25°C (lower panel). Traffic jam (red) and GFP (Piwi, green) antibodies revealed the zygotic expression of Piwi in tj-positive cells at 20°C. A decrease in Piwi zygotic expression was observed at 25°C, consistent with the activation of the thermo-inducible RNAi system targeting Piwi.

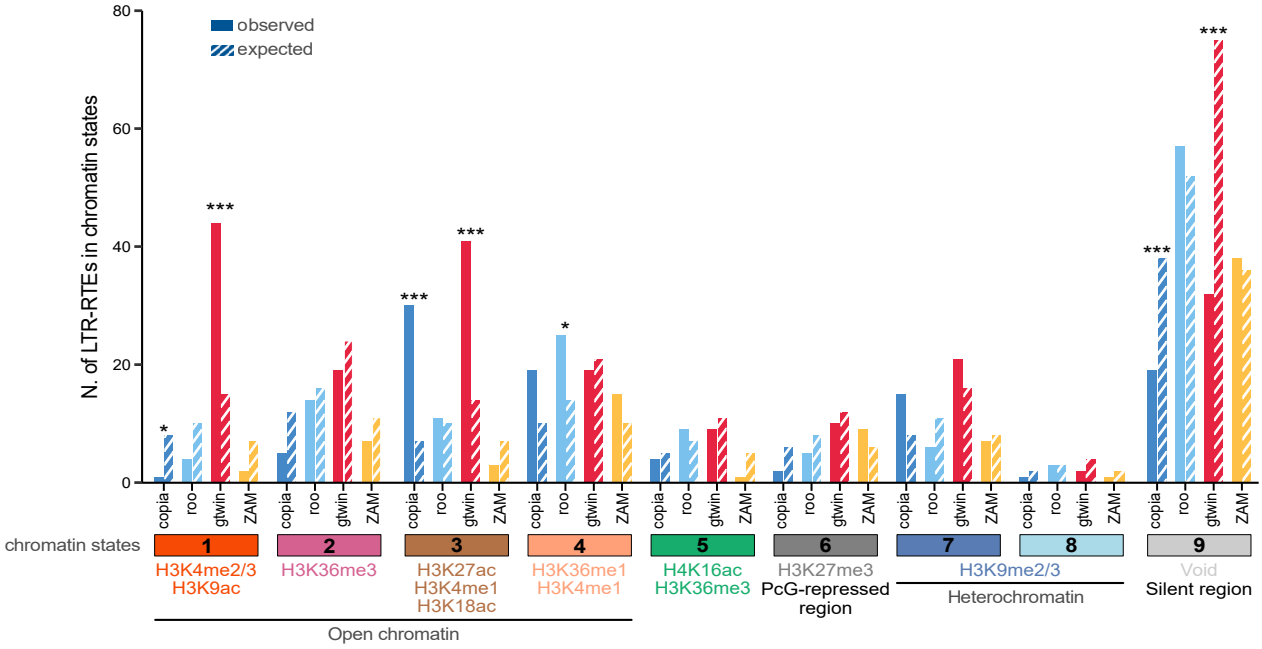


:

Supplementary Figure S1: Genetic model. Schematic representation of our working model and the previous data published in Barckmann et al, 2018. Lower panel: Piwi-dependent repression of LTR-RTEs at 20°C. Upper panel: Transient Piwi knock-down, by temperature change from 20°C to 25°C applied to adult flies for 5 days, allowing LTR-RTE transcriptional derepression in somatic ovarian cells (purple), completion of the LTR-RTE replication cycle, including the production of viral parti-cles that infect the germline (oocyte, blue). The embryos are then maintained at 20°C for the rest of development. The eggs laid by the G1F1 were sequenced using short-reads to detect newly integrated LTR-RTE in the germline.



Supplementary Figure S2: New LTR-RTE insertions are not positively selected in G73 generation (A) Metaplots depicting the frequency of each new LTR-RTE insertion according to its positioning along each chromosome. Chromosomes 2 and 3 are separated in two metaplots corresponding the left and right arms of these chromosomes. Each family of LTR-RTEs is represented with a color code indicated on the top right-hand side. (B) Table indicating the numbers of LTR-RTE insertions in the different piRNA clusters. (C) Frequency of each LTR-RTE insertion found in piRNA clusters.



Supplementary Figure S3: Each LTR-RTE species has its own specific chromatin states preferences for genomic integration

Barplots depicting the observed (filled bars) and expected (dashed bars) numbers of LTR-RTEs insertions across the nine described chromatin states in S2 cells (14). Chromatin states are clustered with typical post-translational histone marks associated with each chromatin state. Expected values were calculated according to the size of each chromatin state in the genome and the total number of new insertions obtained for the indicated LTR-RTEs. Statistical significances (p-values) were calculated using binomial tests corrected by Benjamini–Hochberg step up procedure to control the false discovery rate; p-value: * < 0.05, ** < 0.005, *** < 0.001

Name	Numbers of Reads	N50	Mean QuaI, Phred	Total bases	Depth, in X
G0	2,252,087	3.768	9.1	5,418,397,754	30.1
G0F100	3,358,451	11.5	12.5	16,964,518,776	94.2
G11	1,557,842	16.339	12,34	10,749,554,082	59.7
G31	5,681,195	12.0	12.20	23,420,484,737	130.1
G73	3,567,671	19.65	12.4	24,980,845,478	138.8

All lengths are expressed in bases. Quality is expressed in standard Phred scale.

Supplementary Table S1: Sequencing data statistics

Families	G11	G31	G73
1731	1	1	0
17.6	3	13	13
297	4	11	9
3S18	2	3	3
412	3	15	10
accord	0	2	2
blood	11	18	13
Burdock	0	2	3
Circe	1	0	2
copia	**40**	**61**	**107**
diver	1	1	1
Dm88	0	2	1
flea	8	14	14
GATE	1	5	3
gtwin	**6**	**19**	**210**
gypsy	1	4	6
gypsy4	1	1	0
gypsy6	1	0	2
gypsy7	1	0	0
HMS-Beagle	0	7	2
HMS-Beagle2	1	0	1
Idefix	3	7	18
invader1	6	8	12
invader2	1	1	0
invader3	0	2	2
invader4	3	4	2
invader6	3	2	7
Max-element	8	4	5
McClintock	1	0	0
mdg1	5	7	8
mdg3	7	8	19
micropia	1	2	2
opus	1	13	8
Quasimodo	3	5	4
roo	**104**	**153**	**144**
rover	**23**	**48**	**39**
springer	0	0	1
Stalker	4	4	5
Stalker2	1	0	1
Stalker4	9	14	10
Tabor	1	1	1
Transpac	4	8	7
ZAM	**6**	**44**	**101**

Supplementary Table S2: Annotation of newly integrated LTR-RTEs in the different generations compared to G0 parental genome.

	rover			roo			copia		
Frequency Range (%)	G11	G31	G73	G11	G31	G73	G11	G31	G73
[0-1]	**2**	**36**	**34**	**2**	**45**	**10**	**1**	**35**	**70**
[1-2]	**20**	**9**	**5**	**41**	**42**	**57**	**17**	**7**	**13**
[2-3]	**1**	**1**	0	**15**	**17**	**26**	**16**	**3**	**10**
[3-4]	0	0	0	**10**	**12**	**11**	**2**	**4**	**6**
[4-5]	0	0	0	**7**	**6**	**9**	**1**	**2**	**3**
[5-6]	0	0	0	**5**	**3**	**5**	0	**2**	**1**
[6-7]	0	0	0	**4**	**4**	**6**	0	**2**	**1**
[7-8]	0	0	0	**2**	**2**	**1**	**2**	**2**	**1**
[8-9]	0	0	0	**6**	**2**	**2**	0	0	0
[9-10]	0	0	0	**2**	**2**	0	**1**	**2**	**1**
[10-20]	0	**1**	0	**9**	**13**	**12**	0	**1**	0
[20-30]	0	0	0	**1**	**3**	**3**	0	0	0
[30-40]	0	0	0	0	**2**	**2**	0	**1**	**1**
[40-50]	0	0	0	0	0	0	0	0	0
[50-60]	0	**1**	0	0	0	0	0	0	0
[60-70]	0	0	0	0	0	0	0	0	0
[70-80]	0	0	0	0	0	0	0	0	0
[80-90]	0	0	0	0	0	0	0	0	0
[90-100]	0	0	0	0	0	0	0	0	0

Supplementary Table S3: Distribution of the rover, roo and copia new insertions according to their sequencing frequencies in the G11, G31 and G73 populations

	copia			roo			rover		
Generation	G11	G31	G73	G11	G31	G73	G11	G31	G73
Total insertion number	40	61	107	104	153	144	23	48	39
Shared Insertion number *	1	1	2	11	7	13	0	0	0

* Shared insertions (that were detected in at least two different generations) were very likely vertically transmitted.

Supplementary Table S4 : Evidence for the existence of copia and roo germinal insertions.

Regions	Proportion of the genome	copia (n = 107)			roo (n = 144)			gtwin (n = 210)			ZAM (n = 101)		
		Obs	Exp	P.adj	Obs	Exp	P.adj	Obs	Exp	P.adj	Obs	Exp	P.adj
Intergenic	0.314359	40	34	0.281411	44	45	0.857967	59	66	0.400694	38	32	0.281411
Intron	0.418317	57	45	0.044382	83	60	7.18924E-4	109	88	0.012066	46	42	0.524310
Exon	0.267325	10	29	1.09180E-4	17	38	1.09180E-4	42	56	0.049317	17	27	0.048480
Clustering													
Cluster 1	0.078988	3	8	0.139088	5	11	0.139088	28	17	0.030257	1	8	0.030257
Cluster 2	0.116372	40	12	1.81247E-10	19	17	0.719466	59	24	1.32809E-09	13	12	0.816913
Cluster 3	0.070921	3	8	0.204977	8	10	0.816913	14	15	1	8	7	0.826181
Cluster 4	0.091350	4	10	0.152861	11	13	0.816913	18	19	0.964852	5	9	0.357746
Cluster 5	0.083450	5	9	0.389563	4	12	0.046022	10	17	0.152861	15	8	0.100669
Cluster 6	0.191103	17	20	0.670884	39	28	0.076920	35	40	0.670884	23	19	0.631236
Cluster 7	0.077645	13	8	0.389563	7	11	0.670884	18	16	0.862047	6	8	0.862047
Cluster 8	0.270241	8	29	4.44044E-06	46	39	0.378466	25	57	1.45777E-06	27	27	1

Supplementary Table S5: Distribution of the copia, roo, gtwin and ZAM new insertions according to the genomic or chromatin features of their landing sites. “Obs” and “Exp” refer to the observed and expected numbers of LTR-RTE insertions for each genomic and chromatin features. Adjusted p-values "P.adj" were calculated using binomial tests corrected by Benjamini–Hochberg step up procedure to control the false discovery rate.

Antigen	Type of Data	Tissue	GSE	Reference
H3K4me3	Chip-seq of Histone Modification	Embryo 14-16h	GSE47285	DOI: 10.1038/nature11247
H3K9ac	Chip-seq of Histone Modification	Embryo 14-16h	GSE55557	DOI: 10.1038/nature11247
H3K36me3	Chip-seq of Histone Modification	Embryo 14-16h	GSE47256	DOI: 10.1038/nature11247
H3K27ac	Chip-seq of Histone Modification	Embryo 14-16h	GSE47237	DOI: 10.1038/nature11247
H3K4me1	Chip-seq of Histone Modification	Embryo 14-16h	GSE47281	DOI: 10.1038/nature11247
H3K36me1	Chip-seq of Histone Modification	Embryo 14-16h	GSE47241	DOI: 10.1038/nature11247
H3K27me3	Chip-seq of Histone Modification	Embryo 14-16h	GSE47241	DOI: 10.1038/nature11247
H3K9me2	Chip-seq of Histone Modification	Embryo 14-16h	GSE47247	DOI: 10.1038/nature11247
H3K9me3	Chip-seq of Histone Modification	Embryo 14-16h	GSE47246	DOI: 10.1038/nature11247
GAF	**Chip-seq of Transcription factor**	**Embryo 2-4h**	**GSM6045770 GSM6045771 GSM6045772**	**DOI: 10.1038/s41467-023-41408-1**
OPA	**Chip-seq of Transcription factor**	**Embryo 3h**	**GSM4182505 GSM4182506**	**DOI: 10.7554/eLife.59610**
Clamp	Chip-seq of Transcription factor	Embryo 2-4h	GSM4618642 GSM4618660 GSM4618678	DOI: 10.7554/eLife.69937
Zelda	Chip-seq of Transcription factor	Embryo 2-4h	GSM4618643 GSM4618661 GSM4618679	DOI: 10.7554/eLife.69937

In bold : transcription factor profiles in bigWig format on the dm6 directly available.

REFERENCE TABLE web link

Structure	Original genome built	Reference	Link
Genes	dm6	https://doi.org/10.1093/nar/gkac958	https://www.ensembl.org/biomart/:/836e762aea9a306041c6e5ed15cd59e7

Exon	dm6	https://doi.org/10.1093/nar/gkac958	https://www.ensembl.org/biomart/martview/836e762aea9a306041c6e5ed15cd59e7
Chromatin States	dm3	https://doi.org/10.1093/genetics/iyad211	ftp://ftp.flybase.org/flybase/associated_files/r6_liftedover_DataS5_Chromatin9state.S2.bed.gz
Chromatin accessibility single cell ATAC-seq	dm6	DOI: 10.1126/science.abn5800	https://shendure-web.gs.washington.edu/content/members/DEAP_website/public/ATAC/revision/bigwigs/
Blacklist of the dm6	dm6	DOI: 10.1038/s41598-019-45839-z	https://github.com/Boyle-Lab/Blacklist/raw/master/lists/dm6-blacklist.v2.bed.gz

Supplementary Table S6 : Key Resources Table

smiFISH-primers	sequences
copia-1-FLAPX	TCATTTAGGTCAGTCATCCTAAACTTTTCCATCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-2-FLAPX	CTAGCCGCTTGCTTGAGTCCGTAAATTGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-3-FLAPX	TGATCCCTTACTTTCATTCAGGTGATCATCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-4-FLAPX	GTAGTCGGACAATACCACGCTTAGTGGCACCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-5-FLAPX	AATTTTATCCATCTCTTCTATTTTTGCACCAGCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-6-FLAPX	CAAACGGCTTAATATTACGTTTAGCCTTGTCCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-7-FLAPX	AGTAACTATCAAATGTGGGTGGTGTGCATTCTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-8-FLAPX	GAAGTGTTAACTGATCCAGCATTTGCTGCGCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-9-FLAPX	AATTTCATTCAGCATTCGATTGGTCGTCTTGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-10-FLAPX	AAGACAAATCACATTATTCTGAACTTGCTCTCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-11-FLAPX	TTAGCTCGTTTATGACATGAGGGGTTGTTTGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-12-FLAPX	CAAATAGGGCCATATACTCAGCTTCAGTTGACCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-13-FLAPX	AGGCTGCTACTGAGTTCTGTCTCTTTGTACCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-14-FLAPX	TTGTACTTTTTCTATCAATTTCACTACCAGCCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-15-FLAPX	ATTCGGAGTTATTTTTGCTACTATATCTGCTCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-16-FLAPX	TGGCTTAAATAGATTTTATCTTCCTGCATCTCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-17-FLAPX	AAGTTATTCATTCTTGTCATATCTCCTGTAGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-18-FLAPX	GCGATCAACTGAAGAGTTTACAAACTCACACTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-19-FLAPX	TCCTCTTTTAACGTGCCATTTAAGAAAGCTGTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-20-FLAPX	TTGGTATTTTTGAGTGAATCCTCGTGCAACCACCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-21-FLAPX	TATGAGCATTTAACTCTGTATTGATGGCTTCTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-22-FLAPX	TATTATCCTCTTCATTATAGGATATCTGAGGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-23-FLAPX	CTCTCACTTCTTCTATTAATAATTTCTATGCCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-24-FLAPX	ATCAATTCCAATTTCTTTTAAGTGCTCTGCTGCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-25-FLAPX	TCACTTTCCCTACTCTCATTCGGGTTGCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-26-FLAPX	ATCTTTCAGGAATTGTATGTTGTCGCATTCCTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-27-FLAPX	ATCTTTCAGGAATTGTATGTTGTCGCATTCCTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-28-FLAPX	TTTCACTTTCCTTACTATCTTTCAGGAACACTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-29-FLAPX	TAATATGCACATAAACAGTTGCACCAAACACTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-30-FLAPX	GTTTTAAGTATGGCTTCTTATTGTGCCACATCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-31-FLAPX	AAGCTTTTATCTAGCTTTGCACCACTAACCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-32-FLAPX	AGCTTTTTCCGTAATGGTTCTTATCATTCTCTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-33-FLAPX	GGAAGTCTTGCCTGTTTACCATTTAAACAGGGCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-34-FLAPX	ATTTCTAATAATTTGCCATCGCTTATATGGCCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-35-FLAPX	AATGGTTACACCGCTTTTGTCAAATTCGATCGCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-36-FLAPX	ACATCCTCCAGTGTAATCTCATGGTCATTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-37-FLAPX	AATAAATTCGCCTTGCTTGGCCACTGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-38-FLAPX	GAAACAATCTTTTTTAATGTGGCCTTCTCTGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-39-FLAPX	AGTGTTATTATTGTTGTGCACGATCGCGTTCACCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-40-FLAPX	ATGGCTTAATAGTGACATCTCACTCGATAGCCCTCCTAAGTTTCGAGCTGGACTCAGTG

copia-41-FLAPX	CAGAGAAAGCAAACGTTTTCGCAGCGCCAGCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-42-FLAPX	GCGACGCCAAACTTTTTCGTTCATAAACGGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-43-FLAPX	ACGAGTCGCTTAGGTACTCTATTATTGTACTTCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-44-FLAPX	TGCACAACGCTCTGCCTTTTTCCAGGCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-45-FLAPX	ATTTAATTGTTTATTAGGCATGGACTGGGCCCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-46-FLAPX	TTTTAAGTTATTTCAACTGCAACACCAGCACCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-47-FLAPX	TTGTAGGTTGAATAGTATATTCCAACACGCCCCCTCCTAAGTTTCGAGCTGGACTCAGTG
copia-48-FLAPX	GATACGGGGAAAACCCAGAAAAACCCGATCACCTCCTAAGTTTCGAGCTGGACTCAGTG
roo-1-FLAPY	TTTTGGATAAGTCTCCACCTATCCAAATTTCCTTACACTCGGACCTCGTCGACATGCATT
roo-2-FLAPY	CTGTTTTATAGATGCAGTTCGCCGCTTTATCTTACACTCGGACCTCGTCGACATGCATT
roo-3-FLAPY	TCTTCATTTCGAATTGGCACCGGAATAAGCCTTACACTCGGACCTCGTCGACATGCATT
roo-4-FLAPY	GTTAACAGTGTATACACCTCCTTAAGTTCCGTTTACACTCGGACCTCGTCGACATGCATT
roo-5-FLAPY	CTGACCGTTTTCCTGGAATTACTAGGTTCTCTTACACTCGGACCTCGTCGACATGCATT
roo-6-FLAPY	TATAATGCCTGCTTGAATTTTTTCGCCTTCTTTACACTCGGACCTCGTCGACATGCATT
roo-7-FLAPY	TCTTCATATTTTCTTCTAATTGCTCTCTATCATTACACTCGGACCTCGTCGACATGCATT
roo-8-FLAPY	TCTTTTTGGCTCTTGGTCAGCCTCATTTTTTTACACTCGGACCTCGTCGACATGCATT
roo-9-FLAPY	CTATTTGCCCTTCTTCTATATTAGGAATTTCCTTACACTCGGACCTCGTCGACATGCATT
roo-10-FLAPY	TTCCATTTATTCCTTTGCTGGAGCGTATGCTTACACTCGGACCTCGTCGACATGCATT
roo-11-FLAPY	AATGGTATTTCATTGACTTAACTCCAGCTTCCTTACACTCGGACCTCGTCGACATGCATT
roo-12-FLAPY	AATAACTCTTGATCTAATTTTCTTGCAGCTCCTTACACTCGGACCTCGTCGACATGCATT
roo-13-FLAPY	TTGATCTTACTTTGACCACTGTTAATCCATGCTTACACTCGGACCTCGTCGACATGCATT
roo-14-FLAPY	ACACAGCTCGAGTTTGGGAATTGTCTTCCTTTACACTCGGACCTCGTCGACATGCATT
roo-15-FLAPY	AGCATATGCTTTTTCGGAGGCGTCCGCTTACACTCGGACCTCGTCGACATGCATT
roo-16-FLAPY	ATCCATTTCCTTAAGTTGAATCCAACTTTCTGTTACACTCGGACCTCGTCGACATGCATT
roo-17-FLAPY	TAATTTATTAGCTTCTTCTACCGAATCAGCTCTTACACTCGGACCTCGTCGACATGCATT
roo-18-FLAPY	GTAAGGCTTGAATCTCTAATCACTGCCTGTTACACTCGGACCTCGTCGACATGCATT
roo-19-FLAPY	TTTTTCTCCAGATTCATGTAACGAGCTATCGCTTACACTCGGACCTCGTCGACATGCATT
roo-20-FLAPY	GTTTCTTTGAATCTCCTAAGGTGACATCCTCTTACACTCGGACCTCGTCGACATGCATT
roo-21-FLAPY	GTGGCTTCATAATTTTCTCCAGAGCCGAGTTACACTCGGACCTCGTCGACATGCATT
roo-22-FLAPY	GTAAATGAGTAACCACATTTCTGGCTTCTCCTTACACTCGGACCTCGTCGACATGCATT
roo-23-FLAPY	GTTCGCTAATTTCTTCTTCGAATAGCGAACTCTTACACTCGGACCTCGTCGACATGCATT
roo-24-FLAPY	TTGCTGCTTCTGCTGCTGCTGTTGTTTTACACTCGGACCTCGTCGACATGCATT
roo-25-FLAPY	CTTCTGCTGCTGGTAGAGGCTCCTTTTTACACTCGGACCTCGTCGACATGCATT
roo-26-FLAPY	CACATCTGCCTATCTTGAGCGGCGAGTTACACTCGGACCTCGTCGACATGCATT
roo-27-FLAPY	CCTTATCTGTGGTCTCCCACTAAGGGATTACACTCGGACCTCGTCGACATGCATT
roo-28-FLAPY	ATATATTCGTGTTCATGTGTGAACATTCTGCCTTACACTCGGACCTCGTCGACATGCATT
roo-29-FLAPY	ACTAAGATTTCAATGGGCCTAGTTTTTCTGGCTTACACTCGGACCTCGTCGACATGCATT
roo-30-FLAPY	AGGTTATTGCTTGCATTCTTTGTTGCACAGCTTACACTCGGACCTCGTCGACATGCATT
roo-31-FLAPY	ACTCCACTAACTTCTCCTATATAAGGTGTTGTTACACTCGGACCTCGTCGACATGCATT
roo-32-FLAPY	ATTCTTCACTTTCGGACTGAATGTTATGGTGGTTACACTCGGACCTCGTCGACATGCATT
roo-33-FLAPY	ACTAGAATTTTATCGTCGGTTCTTGCCGTACATTACACTCGGACCTCGTCGACATGCATT
roo-34-FLAPY	ATGGCGCAATGCTTAAAACTCCTTGCTGTTACACTCGGACCTCGTCGACATGCATT

roo-35-FLAPY	GCTTTAGAGGCTGAATCTCACATGCATTATTTACACTCGGACCTCGTCGACATGCATT
roo-36-FLAPY	TTCGCTGAATTCCAGGGCCAACTTCCTTACACTCGGACCTCGTCGACATGCATT
roo-37-FLAPY	GCATATTCTCTTATTTAGTTGCCATTTCTGACTTACACTCGGACCTCGTCGACATGCATT
roo-38-FLAPY	AATGTGGCTTCCGTCATTATGTGATAGGAATCTTACACTCGGACCTCGTCGACATGCATT
roo-39-FLAPY	ATATTCGGACGTTGTATGCACCATTATTATCCTTACACTCGGACCTCGTCGACATGCATT
roo-40-FLAPY	TGGCAATTTCTTTTTTAAGCTGATTTGGCCTGTTACACTCGGACCTCGTCGACATGCATT
roo-41-FLAPY	ACTTCTGTCATGTTCTCAATTCTTATTTGCATTTACACTCGGACCTCGTCGACATGCATT
roo-42-FLAPY	ACAAGGAACCCATAAATTCGAAAGGAGCACGCTTACACTCGGACCTCGTCGACATGCATT
roo-43-FLAPY	CTCCTAGCGGGTCTAGATATATTGCTGAGTTACACTCGGACCTCGTCGACATGCATT
roo-44-FLAPY	ATATCTTGGTGCTATATCTTTAGGTAATGCGCTTACACTCGGACCTCGTCGACATGCATT
roo-45-FLAPY	CTTGGACATTCCCGATGTCGCTCTAGTTTTACACTCGGACCTCGTCGACATGCATT
roo-46-FLAPY	TTAGTAATGGGTCTAGTGATATATCCTTCCTGTTACACTCGGACCTCGTCGACATGCATT
roo-47-FLAPY	TCTTTACTTTAGCCACTCGGACCTTATCATTACACTCGGACCTCGTCGACATGCATT
roo-48-FLAPY	CTAAAGGCCATCTTGCAGGATGACAATTCTTTACACTCGGACCTCGTCGACATGCATT
gtwin1-FLAPX	TTTATGTCTCTATATGCCGTCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin2-FLAPX	GACAAAAACCTACGGGCTCTCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin3-FLAPX	CTGCAGTGCTATTTGTAGTACCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin4-FLAPX	CTGCAGTGCTATTTGTAGTACCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin5-FLAPX	CAGTCAAAGCCTGGTTAACGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin5-FLAPX	CTCTTAAGTCTGTTTTCCTGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin6-FLAPX	CATCTAATTGCGCCTGGAAACCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin7-FLAPX	AGTGTTGCATCTGTTGTTTTCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin8-FLAPX	TACTTGTGGTGCTTCCACACCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin9-FLAPX	AATGTCAAGCTTGATGTCGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin10-FLAPX	AATGTCAAGCTTGATGTCGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin11-FLAPX	ATGACACATAGTCATCCTGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin12-FLAPX	AAGATCTCGTATGCGTCCACCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin13-FLAPX	TCGAAACGGCCTGGTAATGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin14-FLAPX	TCGAAACGGCCTGGTAATGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin15-FLAPX	TAGGACTGTGTTATGGGCAACCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin16-FLAPX	TGCAGTCTAGTCGAGCAATACCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin7-FLAPX	TAAGTACACGCAGCGACGTTCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin18-FLAPX	TAAGTACACGCAGCGACGTTCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin19-FLAPX	TATTGCATGAGGCTAAGCTCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin20-FLAPX	TGTCAACCGTTTCTCTACATCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin21-FLAPX	GTGAGTCATAACTATCTTGTCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin22-FLAPX	AAAGAATAGCACTGTCCGGTCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin23-FLAPX	AGAGCATCCTCTCTAACTTCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin24-FLAPX	TTCCTTAGACCGGCAATGAACCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin25-FLAPX	GAGGGCAAATCCTTGGGTTGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin26-FLAPX	ATTCCCTAGCCAAAGCTAAACCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin27-FLAPX	GTTAGTGCGTTCTATGGTAGCCTCCTAAGTTTCGAGCTGGACTCAGTG

gtwin28-FLAPX	TAGAGCCTTTGCGTATGTAGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin29-FLAPX	TCATAGGTGATTGCCTTGTCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin30-FLAPX	TTGAGTGGTTCCGATCTTTTCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin31-FLAPX	AAATGAGGGTTTTTCCCTTGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin32-FLAPX	TTATTTTGTCTACCCTGTTTCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin33-FLAPX	TAGCTGTTATTATTCTGGGCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin34-FLAPX	TAGGCTCAGGTGACTGATTCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin35-FLAPX	GTTACAGTGGAGCCTAACTCCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin36-FLAPX	AAGGAGGTCGGTTGCCTATGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin37-FLAPX	GAATCTCTTCTGGCCACGAGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin38-FLAPX	TGTTGGATGGTGTTCTGCACCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin39-FLAPX	ATTCGCTGTCGTTTTTATCGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin40-FLAPX	CGTCATCTATCTCGGTAACGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin41-FLAPX	GCGCTCCCTAAAAAATTGATCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin42-FLAPX	GTTCAATGAACGGCAGTCGGCCTCCTAAGTTTCGAGCTGGACTCAGTG
gtwin43-FLAPX	AATTTCCTCCCATGGAACTGCCTCCTAAGTTTCGAGCTGGACTCAGTG
zam1-FLAPY	ATTGCCAACATTTCTTTTTCTTTACACTCGGACCTCGTCGACATGCATT
zam2-FLAPY	ACGCCATAAATATATGGTCTTTACACTCGGACCTCGTCGACATGCATT
zam3-FLAPY	TTAGTGGCTTGTGATCTGTATTACACTCGGACCTCGTCGACATGCATT
zam4-FLAPY	GTTGGGTTCTTTGAAATTCATTTACACTCGGACCTCGTCGACATGCATT
zam5-FLAPY	TTACACTCGGACCTCGTCGACATGCATTGTCTCCAACGAATTATTTTT
zam6-FLAPY	TTTGTGAACCTTTCTTGTGATTACACTCGGACCTCGTCGACATGCATT
zam7-FLAPY	TTACACTCGGACCTCGTCGACATGCATTTTTAAATTTGGGTCCGCTCTAC
zam8-FLAPY	ATGTGGGGCAAGGCTTAACATTACACTCGGACCTCGTCGACATGCATT
zam9-FLAPY	TTACACTCGGACCTCGTCGACATGCATTTTCGTTAATAGGTTTTTCGG
zam10-FLAPY	TCTAGTATGAGTTGCGTGTTTTACACTCGGACCTCGTCGACATGCATT
zam11-FLAPY	GGTGTTGTAGTTTGGTAAGATTACACTCGGACCTCGTCGACATGCATT
zam12-FLAPY	TTACACTCGGACCTCGTCGACATGCATTGGCTGTGAATATTTTTTCCT
zam13-FLAPY	AGGGTTCCTTTCAAGATTTTTTACACTCGGACCTCGTCGACATGCATT
zam14-FLAPY	ATTATCGTCCGCCAAGAATGTTACACTCGGACCTCGTCGACATGCATT
zam15-FLAPY	GCTTTTTCGATTAATGCGGATTACACTCGGACCTCGTCGACATGCATT
zam16-FLAPY	TCCTACGATTTCGTGGAGAATTACACTCGGACCTCGTCGACATGCATT
zam17-FLAPY	TTACACTCGGACCTCGTCGACATGCATTTGAATTTGTTTTGTTCGGGGT
zam18-FLAPY	TGCTGTTAGTGTGATATTCCTTACACTCGGACCTCGTCGACATGCATT
zam19-FLAPY	GAATGTTTCGTCTATGCCTCTTACACTCGGACCTCGTCGACATGCATT
zam20-FLAPY	GTTTTTCATATTGGGGAAGTTTACACTCGGACCTCGTCGACATGCATT
zam21-FLAPY	AGGTTTCGCAATTCCTAATTTTACACTCGGACCTCGTCGACATGCATT
zam22-FLAPY	GTCGTATTTGAGTTTTAGACTTACACTCGGACCTCGTCGACATGCATT
zam23-FLAPY	GGGTTTCTGGGGTTTCAAATTTACACTCGGACCTCGTCGACATGCATT
zam24-FLAPY	GTATTATGTCGAGGGGTTTCTTACACTCGGACCTCGTCGACATGCATT
zam25-FLAPY	TTTGTCTATGATTGTCAGGTTTACACTCGGACCTCGTCGACATGCATT
zam26-FLAPY	TTTCTATTTGGGATGGGGTATTACACTCGGACCTCGTCGACATGCATT

zam27-FLAPY	TTACACTCGGACCTCGTCGACATGCATTCTTTGATGCAATTGATGCCG
zam28-FLAPY	GGTAGTCCGAATTGACTGAATTACACTCGGACCTCGTCGACATGCATT
zam29-FLAPY	TTACACTCGGACCTCGTCGACATGCATTTCGTTGCAAAATTCTACTCCT
zam30-FLAPY	GCTGGAAGGATGTGACATGGTTACACTCGGACCTCGTCGACATGCATT
zam31-FLAPY	CGTTCTACTGGAGAATTACTTTTACACTCGGACCTCGTCGACATGCATT
zam32-FLAPY	TCTCTGTCAAAGAGGAGTGTTTACACTCGGACCTCGTCGACATGCATT
zam33-FLAPY	TTTTCCGTGTGTCTAGTATTTTACACTCGGACCTCGTCGACATGCATT
zam34-FLAPY	TTCTGACATTATTTCTTCGTTTACACTCGGACCTCGTCGACATGCATT
zam35-FLAPY	GGAGTGGATTGCGTTGTTATTTACACTCGGACCTCGTCGACATGCATT
zam36-FLAPY	TTACACTCGGACCTCGTCGACATGCATTAAAAGGGGTGTGTTTGGTGG
zam37-FLAPY	TTACACTCGGACCTCGTCGACATGCATTCGAATAAATGGGTCCTACCA
zam38-FLAPY	TTACACTCGGACCTCGTCGACATGCATTTCATGCTCATTATTGGGTAT

Supplementary Table S7: List of primers used for smiFISH experiments

		gtwin			ZAM		
	Proportion	Obs	Exp	P.adj	Obs	Exp	P.adj
Sci-ATAC-seq 0-2h	0.084992	49	17.85	9.25694e-11	9	8.58	0.857736
ATAC-seq PGC	0.017645	2	3.71	0.595270	5	1.78	0.134952
Overlapping	0.024339	27	5.11	5.67542e-12	3	2.46	0.857736
Other	0.873024	132	183.34	8.91007e-19	84	88.18	0.459302
Sci-ATAC-seq 0-2h	0.093613	58	19.66	6.59140e-14	10	9.45	0.863642
ATAC-seq PGC unique	0.010916	2	2.29	1	5	1.10	0.020664
Overlapping	0.015718	18	3.30	1.20960e-08	2	1.59	0.863642
Other	0.879753	132	184.75	3.44636e-20	84	88.86	0.599892

Supplementary Table S8: Distribution of gtwin and ZAM new insertions based on chromatin accessibility in 0-2h embryos (sci-ATAC-seq) and PGC of late embryos (ATAC-seq) at their respective landing sites. « Obs » and « Exp » refer to the observed and expected numbers of LTR-RTE insertions for each genomic and chromatin feature. Adjusted p-values ("P.adj") were calculated using binomial tests corrected with the Benjamini–Hochberg step-up procedure to control the false discovery rate.