Temporal and spatial niche partitioning in a retrotransposon community of the Drosophila 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 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, Université de 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

Summary

Transposable elements (TEs) are widespread genetic parasites that potentially threaten the stability of their host genome. Hence, the 'peaceful' co-existence observed today between the different TE families and their host genome is likely the result of selection favoring certain features of TE replication cycles that do not harm the host or compete other TEs. Here, using inducible impairment of the ovarian somatic Piwi-dependent silencing of TEs in drosophila, we demonstrate that four Drosophila LTR RetroTransposable Elements (LTR-RTEs), despite sharing the same overall integration mechanism, preferentially integrate into distinct epigenetically marked open chromatin domains of the host germline. Notably, the differential expression of the gtwin and ZAM LTR-RTEs in ovarian and embryonic somatic tissues, respectively, results in the targeting of the distinct accessible chromatin landscapes of early and late embryonic nuclei. These findings highlight connections between the temporal and spatial partitionings of the LTR-RTE niches, that probably allowed their co-existence in the genomic ecosystem.

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, plays an indispensable role. Some sequences, such as transposable elements (TEs) exhibit a self-serving behavior due to their ability to autonomously express and insert in various locations within the genome and can be considered as genomic parasites. Indeed, continuous TE activity may lead to harmful mutations, including disruptions of coding sequences, chromosomal rearrangements facilitated by ectopic recombination, and impairment of gene regulation[1]. These mutations may ultimately jeopardize the integrity of the host genome. Therefore, the present-day host-TE interactions are the only ones that did not lead to extinction of either the host or the TE[2] Unraveling the diversity of such host-TE interactions is a pivotal area of research that provides valuable insights into genome function and disease biology.

TEs generally remain within the colonized host genome, except for rare instances of horizontal transfer[3]. To ensure its survival, each TE must anticipate its progressive mutational decay by inserting new functional copies into the host germline DNA, allowing vertical transmission to next generations. However, the rates of such replicative transpositions have been set, during evolution, at low levels compatible with the survival of both the host and TEs. Some of the mechanisms controlling TE transposition involve 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].

In *Drosophila melanogaster*, comparative analyses of the conserved reverse transcriptase domain of Long Terminal Repeat RetroTransposable Elements (LTR-RTEs), which replicate via an RNA intermediate, revealed that they are distributed into three clades: Copia, BEL and Gypsy[5,6]. LTR-RTEs of the Copia and BEL clades each encode a single open reading frame (ORF) and are represented by species a few species as compared with the many species of the Gypsy clade species also displaying a stronger heterogeneity in coding sequence, with one, two or three ORFs[5,6]. Several TE species of the latter clade are specifically expressed in the somatic cells surrounding the germline, their replication being dependent on the ability to infect germline cells. This infectious capacity, that is linked to the acquisition of an ORF encoding a viral-like envelope in their common ancestor[5,7], is assumed to have preserved the host germline by starting the replication cycle in the adjacent somatic tissue and therefore to have been responsible for the evolutionary success of this clade. Indeed, each ovarian somatic cell type has been exploited as a TE-specific expression niche[7], free of competitive interactions and interference between copies of related TE

58 species. This niche partitioning probably arose after more or less successful attempts by the TE to co-opt
59 endogenous co-factors for transcriptional regulation.

60 At the integration step of the TE replication cycle, co-optation of endogenous proteins as co-factors to
61 specifically insert into either neutral or less essential regions of the genome, is thought to result in tolerant
62 integration niches. Indeed, in organisms with small gene-rich genomes like *S. cerevisiae* and dictyostelids,
63 the insertion into, or close to, coding genes being mutagenic, some TEs preferentially target regions
64 considered as less critical, such as gene-poor heterochromatin or redundant, non-essential multicopy genes
65 like tDNA and rDNA[8]. However, since sharing the same neutral insertion site may lead to competition
66 between TEs, each TE is expected to exhibit its own insertion site preferences. Indeed, recent data suggest
67 that P-elements favor replication origins in Drosophila while LTR-RTEs preferentially integrate near
68 promoters and exons of active genes[9]. Whether integration preferences vary between different LTR-RTE
69 species and/or are influenced by specific cellular contexts requires further investigation.

70 Studying TE ecology regarding not only the interactions between a TE and its host but also between
71 members of the whole community of TE species having colonized the host (the set of TE copies coexisting
72 in the genome of the host) is expected to provide further insights into the ways TE have successfully
73 invaded all present-day eukaryotic genomes.

74 To investigate LTR-RTE ecology, we used a previously constructed *Drosophila melanogaster* line that
75 allows inducible impairment of LTR-RTE repression and then determined the timing and specificities of
76 LTR-RTE expression and integration. Using short-read genome resequencing, we had previously
77 observed a few putative insertions for two LTR-RTE species of the Gypsy clade (gtwin and ZAM)
78 following this de-repression, providing a proof of concept for this approach[10]. At the time, the localization
79 of newly integrated LTR-RTEs was not possible, and the number of integrations was not sufficient to
80 perform a comprehensive comparison of the insertional landscapes.

81 Here, using long-read sequencing to annotate the genomes of flies that had been subjected to LTR-RTE
82 de-repression for many successive generations, we observed the accumulation of LTR-RTEs from
83 different clades. At the 73rd generation, we mapped enough newly inserted copies of four LTR-RTEs (a
84 total of 798 insertions of roo, copia, gtwin or ZAM) to disclose specific biases in their landing site
85 preferences. Indeed, since natural selection had only modestly altered the landscape of these recently
86 integrated LTR-RTEs, it had not yet erased the memory of their initial preferences for insertion sites.
87 Moreover, gtwin and ZAM exhibited differences in expression patterns, but also in the timing of their
88 integration into the embryonic genome, leading to different landing site preferences. Our findings

highlight how, over the course of evolution, the diverse cell identities exploited by various LTR-RTE species for both expression and integration have led to the colonization of TE-specific genomic niches of integration.

Results

Four LTR-RTEs integrate into the Drosophila germline after successive generations of somatic Piwi knockdown

To accumulate a substantial number of LTR-RTEs in the *Drosophila* genome, we used a strain previously constructed in our laboratory[10]. In this strain, a traffic-jam-Gal4/Gal80 inducible driver activates, at high temperature, the expression of a short RNA hairpin construct targeting Piwi (sh-piwi) in the gonadal somatic cells. This somatic knockdown (sKD) alleviates LTR-RTE repression in these cells without causing sterility[10]. When females containing traffic-jam-Gal4/Gal80>sh-piwi are shifted, for a few 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 the ovaries and *de novo* transposition[10]. As previously shown, mobilization of LTR-RTEs already occurs after the induction of Piwi depletion in a single generation[10]. Here, we first confirmed that the fertility of shifted flies and of their progeny were not sufficiently affected to prevent the induction of piwi-sKD during successive generations (Supplementary Fig. 1).

At the 11th (G11), 31st (G31) and 73rd (G73) generations, we isolated a large subset of the shifted population to maintain a stock constantly at 20°C, the non-permissive temperature for piwi-sKD (Fig. 1A). Additionally, as a negative control, we maintained a fraction of the initial parental line G0 constantly at 20°C for 100 generations (G0F100). This experimental scheme provided us with the G0F100 control population that had not undergone any temperature shift, along with the G11, G31 and G73 populations with increasing occurrences of piwi-sKD and hopefully harboring an accumulation of new LTR-RTE insertions in their respective genomes.

Using the Oxford Nanopore Technology (ONT) for long-read sequencing of DNA, we sequenced the genomes of 100 pooled males originating from the G0, G0F100, G11, G31, and G73 populations (Fig. 1A). We used the TrEMOLO pipeline[11] to annotate the newly integrated LTR-RTEs. This bioinformatics tool aligns and compares long-read sequences obtained at each defined generation with those of the initial G0 parental genome (Materials and Methods). We identified an increasing number of new LTR-RTE insertions in the successive shifted generations: with 280, 514, and 798 new insertions for the G11, G31, and G73 generations respectively (Fig. 1A). As shown in Table S1, the majority of these newly identified

LTR-RTE insertions, which belong to 43 species, rarely reached 20 new insertions per species. This could result from a relatively low expression of some of these LTR-RTEs together with a low transposition rate and/or a high selective pressure against new insertions. Nevertheless, among the different LTR-RTE species, five (namely roo, copia, rover, ZAM and gtwin), covering the three known clades of LTR-RTEs (Copia; Gypsy; BEL), had equal or more than 40 new insertions in the successive shifted populations (Fig. 1B, Supplementary Table 1).

Through the same bioinformatics analysis of the G0F100 long-read sequences, we detected 263 insertions that were absent from the G0 initial parental genome (Fig. 1A, B). They uniquely corresponded to roo, copia and rover, suggesting that these three LTR-RTEs were able to express and transpose spontaneously (independently of piwi-sKD treatment). To confirm that these three elements were indeed expressed in the G0F100 stock, we assayed their expression patterns using single molecule inexpensive fluorescent *in situ* RNA hybridization (smiRNA-FISH) at 20°C and 25°C (Fig. 1C). These observations confirmed that roo and copia are indeed expressed regardless of temperature and revealed specific expression patterns for these two elements in the ovaries of G0F100 females. More specifically, roo was expressed in the germline and its transcripts accumulated in the oocyte cytoplasm, similar to a LINE element like the I-element[12,13]. We detected transcripts of copia in the nuclei of the follicle cells that surround the germline. The same analysis for rover expression was inconclusive, as we did not detect transcription in shifted or non-shifted ovaries (data not shown). We hypothesized that most of the new rover insertions in the different populations resulted from somatic transposition occurring during development. In agreement with this possibility, we quantified the frequency of the different new rover insertions and found it was always low (Supplementary Table 2). Moreover, we could not obtain any evidence for vertical transmission of these rover insertions. Indeed, in contrast to some copia and roo insertions shared between the successive populations and therefore likely to correspond to germline inherited insertions, we could not detect a single *rover* insertion that was shared at the same genomic position by at least two generations (Supplementary Table 3).

Contrary to the aforementioned LTR-RTEs, ZAM and gtwin, belonging to the Gypsy clade, had exclusively transposed in the shifted generations, as demonstrated by their increased numbers of insertions (Fig. 1B). As this suggests that their expression was regulated by the presence of Piwi protein in the follicle cells, we analyzed their expression by smiRNA-FISH at restrictive and permissive temperatures. These two elements are clearly not expressed in the ovaries at 20°C and start to express in follicle cells of the ovaries at 25°C, when Piwi has been depleted (Fig. 1C). Interestingly, these two elements exhibit specific niches of expression that are selective to the somatic follicle cells. ZAM expression was restricted to

posterior follicle cells as previously reported[7,14], whereas gtwin seemed to have a broader expression pattern throughout the follicle epithelium (Fig. 1C).

With the deep analysis of long reads obtained by ONT sequencing and the assembly of the G0 genome, we could identify three full-length copies of gtwin, two being localized on chromosome Y and one on chromosome 2R. For ZAM, we identified at least two full-length ZAM copies located on chromosome 2R. Each of these species of LTR-RTEs had produced 6 new insertions after 11 generations of Piwi depletion, and 44 (ZAM) and 19 (gtwin) after 31 generations. This trend intensified with successive generations, with 101 and 210 new insertions in the 73rd generation, respectively (Supplementary Table 1).

Altogether, our long-read sequencing analysis of several generations, submitted to successive somatic Piwi depletion in ovaries, revealed that four LTR-RTE species were able to move in the *D. melanogaster* genome and thereby generate a sufficient number of germinal insertions to warrant further studies.

Selection had little impact on the landscape of newly integrated insertions

Using the new insertions annotated in G73, we investigated whether specific insertion sites were favored by each of the four active LTR-RTE species. Aware of the potential biases introduced by selection on the TE insertion landscapes[15], we first tried to assess the extent of such additional biases in the G73 dataset. To do so, we segmented the *D. melanogaster* genome into intergenic regions, introns, and exons, assigned the new insertions found in G73 to either of these three bins, and statistically compared the observed-to-expected numbers under the null hypothesis of random insertion (Supplementary Table 4). For each of the four species of LTR-RTEs (copia, roo, gtwin and ZAM), the observed numbers of intergenic insertions were roughly in line with expectation (Fig. 2A). This suggests that the previously reported purifying selection operating over longer periods against genic insertions[9,16] did not have enough generations to introduce such a bias in our sequencing data. Moreover, we detected only 2 times less insertions in exonic regions than expected in G73, whereas a tenfold reduction compared to expected was observed for much older insertions[9]. Interestingly, we also detected more insertions in intergenic regions, whereas 1.2-fold depletion was noted for older insertions[9].

A second feature of selection on the TE insertion landscape is the elimination of genic insertions that are oriented in the sense direction of transcription. It is believed that insertions oriented in the sense direction are more detrimental than those in the antisense direction, as the latter are less likely to influence gene transcription signals[17]. In G73, we did not notice any significant difference in the number of insertions in these two opposite orientations and rather a higher number of insertions of the different LTR-RTEs in the

sense orientation. Selection over 73 generations had apparently not lasted long enough to create the expected orientation bias in our population (Fig. 2B). Finally, in the G73 population, we did not detect any signs of positive selection for newly integrated LTR-RTEs, as the vast majority were present at very low frequencies within the population (Supplementary Fig. 2A). More particularly, none of the LTR-RTEs newly integrated in piRNA clusters, showed any tendency to invade the G73 population (Supplementary Fig. 2B, 2C). Altogether, we were confident that selection did not have enough time to significantly affect the landscapes of our newly inserted LTR-RTEs.

Each LTR-RTE species has its own specific chromatin states preferences for genomic integration

Eukaryotic genomes are partitioned into chromatin domains containing different epigenetic states that are essential for proper gene regulation[18,19]. So far, insertions of various TEs have been mainly associated with open chromatin structure[20,21], suggesting that a higher accessibility of genomic DNA could help TE integration. To test whether a higher degree of chromatin flexibility is a driver of integration site-specificity, we analyzed whether LTR-RTE insertions were preferentially associated with some of these different chromatin states. We compared the distributions of the LTR-RTE insertions observed in the G73 population with those expected from the relative genomic proportions of the nine distinct chromatin states defined in S2 cells[18]. These cells are derived from late male embryonic tissues (stages 16-17). While we detected insertions of the different species within all nine chromatin states, we mainly observed significant enrichment in chromatin states 1 to 4 characterized by their openness relative to the other chromatin states (Supplementary Table 4) (Fig. 3). Each of the four active LTR-RTE species had its own specific pattern of insertions. This specificity was particularly evident for gtwin insertions that showed a significant enrichment in chromatin states 1 and 3, corresponding to promoter- and enhancer-like chromatin, respectively (Fig. 3). In contrast, copia insertions showed a broader distribution, with significant enrichment in chromatin states 3, and higher-than-expected detection in states 4 and 7. roo and ZAM rather accumulated in the 4th chromatin state defined as open chromatin with enhancer features including H3K36me1 but devoid of H3K27ac (Fig. 3)[22]. Overall, the distinct profiles obtained for LTR-RTE insertions in G73 are not the result of passive processes but could be driven by active integration mechanisms that contribute to target different LTR-RTEs to specific states of open chromatin.

gtwin preferentially inserts into the open chromatin of enhancer and promoter regions before cellularization

We then focused on the timing of the germline integration of gtwin and ZAM, the only two enveloped LTR-RTEs derepressed in the ovarian follicle cells: how long does this integration step occur after the expression step? Indeed, previous genetics studies of gypsy/mdg4, a similar infectious LTR-RTE with an

envelope-coding gene, have indicated that, despite expression during oogenesis, integration occurs only in the genome of the progeny[23]. Such a long delay of the integration step might correlate with the egg laying-induced decondensation of the oocyte karyosome. We therefore expected that ZAM and gtwin would also integrate during embryogenesis.

Given our findings that gtwin insertions were significantly enriched in promoter- and enhancer-like open chromatin states 1 and 3[18], we focused on the corresponding embryonic epigenetic marks. We leveraged available modENCODE Chromatin ImmunoPrecipitation data of the relevant histone marks coupled with high-throughput sequencing (ChIP-seq) at late (16 to 20 hours after egg laying (AEL))[24] and early (~2.5 hours AEL, corresponding to stage 5)[25] embryogenesis. In this analysis, ZAM was used as a negative control as it is rather associated with chromatin state 4, an open chromatin state that lacks the four tested histone modifications[18,22]. This analysis suggested that gtwin insertion sites tend to be enriched in H3K4me3/H3K9ac and/or H3K4me1/H3K27ac epigenetic histone marks already deposited in 2.5-h embryos and likely corresponding to promoter and enhancer regions, respectively (Fig. 4A).

Interestingly, these regions have been described as among the first to become open and accessible from stage 3-4 (1 to 2 hours AEL)[26]. Focusing on chromatin accessibility, by using previously published ATAC-seq data from whole early embryos[27], we confirmed that gtwin insertion sites were preferentially localized in the open chromatin peaks of stage 3-4 and stage 5 embryos (1-2 and ~2.5 hours AEL, respectively) (Fig. 4B). By contrast, ZAM insertion sites did not correspond to chromatin regions that are open at these early stages of embryogenesis. This open chromatin state, with low nucleosome occupancy and few higher-order chromatin structures[28], is believed to result from the binding of a special class of transcription factors known as pioneer factors. These pioneer factors are unique in their ability to overcome nucleosomal barriers that establish accessibility of cis-regulatory elements required for further DNA transactions[29,30]. We wondered whether the binding of a specific pioneer factor could be correlated with the gtwin insertion sites. To investigate this, we identified the gtwin insertions located in open chromatin of stage 5 embryos (n=84) and then analyzed the stage 5 embryonic ChIP-seq data of four pioneer factors: GAGA Factor (GAF), Opa, Chromatin-linked Adaptor for MSL proteins (CLAMP), and Zelda [31–34]. As expected, most of the stage 5 open chromatin regions containing a gtwin insertion (69 out of 84) were bound by at least one of these pioneer factors. However, it appears that no specific pioneer factor was responsible for gtwin binding: only 32, 24, 6 and 7 of these regions correlated with GAF, Opa, CLAMP and Zelda binding sites, respectively (Fig. 4C). Altogether, our analyses revealed that chromatin accessibility, independently of the pioneer factor associated, is the most significant feature determining the choice of gtwin insertion sites in early embryos.

Using this preferential chromatin feature associated with gtwin integration, we wondered whether it would be possible to determine its timing of integration during embryogenesis. Notably, we wanted to determine if germline gtwin integrations had occurred before or after germline specification, a process that naturally initiates around 1.30h AEL[35]. To do so, we took advantage of single cell indexing sci-ATAC-seq data which, during successive two-hour intervals of Drosophila embryogenesis, enables clustering of individual nuclei on the basis of similarity of their chromatin accessibility patterns[36]. We determined the averaged chromatin accessibility within 4 kb windows around gtwin insertion sites in the four first ATAC-seq clusters of nuclei identified at 0-2 hours AEL. For each cluster, gtwin insertion sites tended to be localized in open chromatin (Fig. 4D). From peak calling we estimated that 76 gtwin insertions were present in regions open in at least one of the four clusters studied at 0-2 hours AEL. Interestingly, 56 of these insertion sites were present in at least two clusters (Fig. 4E). This suggests that most gtwin insertions occurred during very early syncytial stages of Drosophila embryogenesis, before the differentiation of these first clusters and probably also before the specification of germ cells, the nuclei of which being the first to become cellularized at this stage[35].

Chromatin accessibility of ZAM insertion sites correlates with ZAM late embryonic expression

Our analyses of chromatin accessibility did not reveal any preference for ZAM integration in open chromatin regions during early embryonic development (Fig. 4B, D). To investigate the chromatin accessibility of ZAM insertion sites in later developmental stages, we used time-course sci-ATAC-seq data available at 2-hour intervals spanning the entire embryogenesis[36]. Across eight intervals from 0 to 16 h AEL, we calculated the average sci-ATAC-seq signal for 101 ZAM insertions and 210 gtwin insertions, normalized with signals from 100 randomly selected insertions (Materials and Methods) (Fig. 5A). This analysis confirmed that gtwin insertion sites already exhibited an accessible chromatin status in the very early stages of embryogenesis that was then maintained at that high level throughout embryonic development. Concerning the status of the ZAM insertion sites, the chromatin gradually became more accessible during embryogenesis, particularly in 10-12 h, 12-14 h, and 14-16 h AEL embryos (Fig. 5A). Altogether this temporal analysis suggested that ZAM might integrate later during development.

Two non-exclusive hypotheses may explain the delayed timing of ZAM integration into the germline. The first possibility is that ZAM viral particles infecting the germline during oogenesis remain dormant during the early stages of embryogenesis. Alternatively, ZAM might undergo a second wave of expression at later stages of embryogenesis. Since the eggs were collected at 25°C, a temperature that permits piwi-sKD, a second wave of ZAM de-repression could occur. To test this possibility, we combined ZAM smiFISH and germline-specific vasa immunostaining in late embryos. As shown in Fig. 5B, we observed

a high expression of ZAM in the gonads of the embryos laid at 25°C but not at 20°C. At these late stages of embryogenesis, the primordial germ cells (PGCs) have migrated away from the midgut toward the adjacent mesoderm and have become associated with somatic gonadal precursors (SGP)[35,37] that express traffic jam (tj)[38]. As expected, we observed ZAM expression in the tj-positive cells of the late embryonic gonads specifically at 25°C (Fig. 5C). Moreover, the expression of ZAM in SGPs at 25°C is consistent with the permissive temperature for piwi-sKD, as piwi-sKD is driven by traffic-jam-Gal4 at 25°C and not at 20°C due to the presence of Gal80ts. Overall, these analyses show that ZAM is expressed at later stages of embryonic development, in the SGP cells positive for tj and sensitive to Piwi depletion when embryos are laid at 25°C.

To determine whether this later timing of ZAM expression in embryos could result in germline-specific insertions, we decided to determine whether ZAM insertion sites correlate with chromatin regions that are accessible in the late embryonic germ cells. To do so, we took advantage of a *Drosophila* line expressing a GFP tagged version of the germline-specific gene *vasa*[39] to perform GFP cell sorting using FACS[40] (Materials and Methods) from cell extracts originating from overnight egg collections. After cell sorting, ATAC-seq experiments were performed in duplicates on embryonic GFP-positive cells corresponding to PGCs. We compared the averaged profiles of the ATAC-seq signals, performed in duplicate, in the 2kb windows centered around the 101 ZAM insertions sites, with those obtained for 100 randomly chosen 2kb windows in the same cells. We observed that in the PGCs, the averaged ATAC-seq profile reached a maximum centered on the ZAM insertion sites (Fig. 5D). Altogether these experiments show that ZAM insertion sites correlate with a specific chromatin accessibility landscape in late PGCs that was not observed in the four early embryonic clusters before cell differentiation (Fig. 4D). Overall, we propose that the late wave of ZAM somatic expression could lead to PGC infection, and the subsequent targeting of ZAM integration machinery to genomic regions that are accessible in late PGCs.

Discussion

Transposable elements may alternate rapid bursts of activity and prolonged phases of repression during which these genomic parasites do not replicate efficiently within the host genome[41]. Present-day TE-TE and TE-host genome interactions are probably the outcome of multiple co-evolutions, allowing the peaceful coexistence of different TE species within the same genome. In an attempt to describe the diversity of these interactions, we took advantage of a particular *Drosophila melanogaster* laboratory strain[10] to simultaneously impair the repression of, and therefore awaken, several LTR-RTEs. We observed that four active elements (roo, copia, gtwin and ZAM) had been able to efficiently transpose into

the Drosophila germline, after several dozens of successive generations of LTR-RTE de-repression. We found that they all differ in several steps of their replication cycles.

Regarding the expression step, even though they were all transcribed at the end of oogenesis, the RNAs of one of them, roo, were specifically expressed in the germinal nurse cells, while those of the three others were found in different cell-types/compartments of the somatic follicular epithelium, either at the posterior pole, for ZAM, or ubiquitously for gtwin and copia (but, in the latter case, sequestered into the nuclei). We also disclosed a second window of ZAM somatic expression in the SGPs of the late embryonic gonad.

Concerning the integration step, our approach was based on characterizing the overall epigenetic specificity of their genomic insertion sites. We thus found that, although each of the four landing site landscapes correlated with open chromatin, they all seemed to display specific epigenetic preferences. Copia and roo insertions were found predominantly in chromatin states 3 and 4, respectively, while gtwin appeared to preferentially choose promoter- and enhancer-like landing domains, enriched in chromatin states 1 and 3, and ZAM would rather favor chromatin state 4. Moreover, our data indicated that the preference for specific genomic insertion sites may follow the differentiation of the chromatin landscape of the cells that are invaded by gtwin and ZAM at different stages of the embryonic development. Indeed, for maternally deposited gtwin, a significant proportion of the insertions seemed to have occurred as soon as their landing sites had begun to be accessible, at the very beginning of embryogenesis. Conversely, consistent with the late embryonic wave of ZAM somatic expression, several ZAM insertions were located within different open chromatin regions, accessible only in late embryonic germ cells. However, although the insertion of a LTR-RTE into closed chromatin is generally considered unlikely, it cannot be entirely ruled out. Therefore, we cannot exclude the possibility that some maternally deposited ZAM virus-like particles would have driven integration at early stages into close chromatin landing sites that would open later in the gonadic PGCs. Altogether, our findings disclose a novel level of LTR-RTE niche partitioning linking temporal and spatial features of the integration step of the replication cycle.

Diversity of expression and integration niches

The diversity of ovarian expression patterns reported here (Fig. 1) has also been observed recently for 16 species of evolutionarily related LTR-RTEs[7]. These different patterns in the onset of the replication cycles of this TE class, indicate that each LTR-RTE species has evolved specific interactions hijacking tissue-specific transcription factors to adapt their proper expression niche to a specific cell type. 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[42,43], 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[8,21,44]. For example, DNA transposons, like the P-element, manage to create new copies by integrating near the replication origins of the Drosophila genome[9,21], whereas retrotransposons Ty1 and Ty3 specifically insert into Pol III promoters of *S. cerevisiae*[8]. Recently, it has been suggested that LTR-RTEs are rather attracted to open chromatin of active genes, whereas LINE elements, such as the I-element, target AT-rich sites and tend to integrate near telomeres[9]. These TE-specific host affinities have been described to depend on the enzymes driving their integration such as transposases or 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 (Fig. 3). 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 example of decompacted chromatin targeting is the histone H4 tail that can no longer be targeted by HIV when embedded in closed chromatin[45,46].

Our data 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 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.

Putative selective forces leading 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 niche partitioning.

First, we can notice that the expression of the four LTR-RTEs species appears to be restricted to 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[47,48]. A second type of selective pressure might have prevented toxic TE expression[48] 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 insert 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[49] that is followed by larval stages where germ cell division may compensate for previous cell death[50]. Fourth, further research is needed to characterize the phenotypic effects of the TE insertions we studied and determine if their preferred integration sites correspond to TE-specific safe havens within the host genome. Similarly, it is unknown whether the TE-specificity of these integration niches results from detrimental fitness effects of competition between different TE species at the insertion sites. Finally, our non-overlapping TE expression patterns are in agreement with previous observations[7] 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 and leading to their stable coexistence.

Materials and Methods

Drosophila stocks

Fly stocks (G0, G11, G31, G73), used to determine LTR-RTE mobilization and integration, shared the same genotype: *w ; tj-GAL4 ; tubP-GAL80ts, sh-piwi*, as previously described[10]. These stocks have been initially shifted, at every generation, from 20°C to 25°C during a 5-day period, at the adult stage (Fig. 1A). At the 11th (G11), 31th (G31) and 73rd (G73) generation, a large subset of the shifted population was isolated to maintain a stock constantly at 20°C, the non-permissive temperature for piwi-sKD. A fraction

of the initial parental line G0 was also kept constantly at 20°C for 100 generations (G0F100). At each generation, the strains were maintained with a large progenitor population of more than 500 flies. A Drosophila line harboring the genotype: *w* ; *vas::EGFP* was used[39].

Analysis of female fecundity

For the control (Charolle) and the tested (G0F100) strain, 20 freshly hatched females were mated with 10 males at 20°C for 3 days while 10 freshly hatched females were mated with 5 males for 4 days at 25°C. The flies were then let to lay at 20°C and 25°C, respectively, and the number of F1 eggs was counted every 24 hours for 3 days. All egg collections were then left to develop at 20°C. The number of F1 pupae was counted for each condition. The number of F2 eggs laid by 20 3-day-old F1 mated females was counted at 20°C for both conditions. Three biological replicates were analyzed.

Oxford Nanopore Technology (ONT) Sequencing Data Analysis

DNA was extracted from 100 males as previously described[51] and long-read sequencing data were analyzed using the TrEMOLO software[11] 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 was analyzed in Table S5. Frequency estimation was conducted using the TE analysis module of TrEMOLO 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 shifted libraries were established from populations that independently evolved from a shared G0 ancestor line. Consequently, any insertions found in both the G0F100 and any shifted library were 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-RTE 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.

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[52]. 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 6). Annealing between specific probes and their respective FLAP was performed as previously described[52] 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 PBS-Triton 0,3% 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 (Table S6) and either an anti-Rat Vasa antibody

(DHSB, 1:120) or a Guinea Pig traffic jam antibody (gift from D. Godt[53], 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 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.

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

Using the chromosomal gene and exons annotations of *Drosophila melanogaster* genome (BDGP6.46) available on Ensembl Biomart[54] 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[54]. 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 4). 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[55] default parameters, defining the relative size of intergenic regions (pig=0.314359), introns (pin=0.418317) and exons (pex=0.267325) (Supplementary Table 4). 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[18] based on dm3 genome version were transformed to the latest version (Dmel_R6.32) using liftOver tool[56]. As for genomic structures, the proportion of each chromatin state relative to the genome was computed using bedtools genomecov[55] default parameters

(Supplementary Table 4). Significant enrichment or depletion of LTR-RTE insertions in the different chromatin states were calculated using bilateral binomial statistical test considering the null hypothesis is the probability of insertion in a given state for each LTR-RTE species is 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).

Relationship between the orientation of LTR-RTE and gene expression direction

The significance of the LTR-RTEs orientation within genes, according to gene annotation of the Drosophila genome (BDGP6.46) available on Ensembl Biomart[54], was evaluated using binomial tests corrected by Benjamini–Hochberg with equal probabilities for insertions to occur in either orientation.

Analysis of ATAC-seq and ChIP-seq available datasets

Raw data from published ATAC-seq and ChIP-seq experiments (Supplementary Table 7) 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[57]. 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[58]. macs2 (v2.2.7.1) (bdgpeakcall package[59] 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 7), a consensus bed file was created by keeping only overlapping regions from the different replicates using bedtools intersect (v2.27.1)[55]. 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 from the sci-ATAC-seq atlas previously published[36] 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, 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 averaged 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[40]. Briefly, overnight laid embryos from *vas*::EGFP line[39] 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 sucessive 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

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).

ATAC-Seq analysis

After initial quality checks of the data 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). Duplicates alignements 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. The chromatin accessibility in a region spanning 1kb 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 were then computed using Matlab (R2024).

Statistical analyses and visualization

Statistics and data visualization were performed using the ggplot2 library (v3.4.3) (https://ggplot2.tidyverse.org) on R (v4.3.1) (https://www.R-project.org/) and Matlab (R2024). Cytoscape (v3.10.1)[60] was used to create a graph.

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 raw data are available on GEO under accession number: GSE274394.

Acknowledgments

We thank Bernd Schuettengruber for the comments and the editing of 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 and Drosophila core facilities. 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.

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. Payer, L.M., and Burns, K.H. (2019). Transposable elements in human genetic disease. Nat. Rev. Genet. *20*, 760–772. https://doi.org/10.1038/s41576-019-0165-8.

2. 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. https://doi.org/10.1016/j.gde.2023.102092.

3. Gilbert, C., and Feschotte, C. (2018). Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr. Opin. Genet. Dev. *49*, 15–24. https://doi.org/10.1016/j.gde.2018.02.007.

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. https://doi.org/10.1146/annurev-genet-120417-031441.

5. 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. https://doi.org/10.1101/gr.145000.

6. Nefedova, L., and Kim, A. (2017). Mechanisms of LTR-Retroelement Transposition: Lessons from Drosophila melanogaster. Viruses *9*, 81. https://doi.org/10.3390/v9040081.

7. 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. Preprint at bioRxiv, https://doi.org/10.1101/2023.08.03.551782 https://doi.org/10.1101/2023.08.03.551782.

8. 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. https://doi.org/10.1038/nrg.2017.7.

9. 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., gkad054. https://doi.org/10.1093/nar/gkad054.

10. 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. https://doi.org/10.1093/nar/gky761.

11. 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. https://doi.org/10.1186/s13059-023-02911-2.

12. Brennecke, J., Malone, C.D., Aravin, A.A., Sachidanandam, R., Stark, A., and Hannon, G.J. (2008). An epigenetic role for maternally inherited piRNAs in transposon silencing. Science *322*, 1387–1392. https://doi.org/10.1126/science.1165171.

13. Chambeyron, S., Popkova, A., Payen-Groschene, G., Brun, C., Laouini, D., Pelisson, A., and Bucheton, A. (2008). piRNA-mediated nuclear accumulation of retrotransposon transcripts in the Drosophila female germline. Proc. Natl. Acad. Sci. U. S. A. *105*, 14964–14969. https://doi.org/10.1073/pnas.0805943105.

14. Yoth, M., Maupetit-Méhouas, S., Akkouche, A., Gueguen, N., Bertin, B., Jensen, S., and Brasset, E. (2023). Reactivation of a somatic errantivirus and germline invasion in Drosophila ovaries. Nat. Commun. *14*, 6096. https://doi.org/10.1038/s41467-023-41733-5.

15. Langmüller, A.M., Nolte, V., Dolezal, M., and Schlötterer, C. (2023). The genomic distribution of transposable elements is driven by spatially variable purifying selection. Nucleic Acids Res., gkad635. https://doi.org/10.1093/nar/gkad635.

16. 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. https://doi.org/10.1093/molbev/msq337.

17. Sultana, T., van Essen, D., Siol, O., Bailly-Bechet, M., Philippe, C., Zine El Aabidine, A., Pioger, L., Nigumann, P., Saccani, S., Andrau, J.-C., et al. (2019). The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. Mol. Cell *74*, 555-570.e7. https://doi.org/10.1016/j.molcel.2019.02.036.

18. Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T., et al. (2011). Comprehensive analysis of the chromatin landscape in Drosophila. Nature *471*, 480–485. https://doi.org/10.1038/nature09725.

19. Gorkin, D.U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A.Y., Li, B., Chiou, J., Wildberg, A., Ding, B., et al. (2020). An atlas of dynamic chromatin landscapes in mouse fetal development. Nature *583*, 744–751. https://doi.org/10.1038/s41586-020-2093-3.

20. Zhang, X., Zhao, M., McCarty, D.R., and Lisch, D. (2020). Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. Nucleic Acids Res. *48*, 6685–6698. https://doi.org/10.1093/nar/gkaa370.

21. 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. https://doi.org/10.1073/pnas.1112960108.

22. Jayakrishnan, M., Havlová, M., Veverka, V., Regnard, C., and Becker, P.B. (2024). Genomic context-dependent histone H3K36 methylation by three Drosophila methyltransferases and implications for dedicated chromatin readers. Preprint at bioRxiv, https://doi.org/10.1101/2024.02.06.577191

23. 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.

24. Nègre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R., et al. (2011). A cis-regulatory map of the Drosophila genome. Nature *471*, 527–531. https://doi.org/10.1038/nature09990.

25. Li, X.-Y., Harrison, M.M., Villalta, J.E., Kaplan, T., and Eisen, M.B. (2014). Establishment of regions of genomic activity during the Drosophila maternal to zygotic transition. eLife *3*. https://doi.org/10.7554/eLife.03737.

26. Blythe, S.A., and Wieschaus, E.F. Establishment and maintenance of heritable chromatin structure during early Drosophila embryogenesis. eLife *5*, e20148. https://doi.org/10.7554/eLife.20148.

27. Samata, M., Alexiadis, A., Richard, G., Georgiev, P., Nuebler, J., Kulkarni, T., Renschler, G., Basilicata, M.F., Zenk, F.L., Shvedunova, M., et al. (2020). Intergenerationally Maintained Histone H4 Lysine 16 Acetylation Is Instructive for Future Gene Activation. Cell *182*, 127-144.e23. https://doi.org/10.1016/j.cell.2020.05.026.

28. Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187. https://doi.org/10.1016/j.cell.2016.09.018.

29. Barral, A., and Zaret, K.S. (2024). Pioneer factors: roles and their regulation in development. Trends Genet. TIG *40*, 134–148. https://doi.org/10.1016/j.tig.2023.10.007.

30. Zaret, K.S., and Mango, S.E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. Curr. Opin. Genet. Dev. *37*, 76–81. https://doi.org/10.1016/j.gde.2015.12.003.

31. 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. https://doi.org/10.1371/journal.pgen.1002266.

32. 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. https://doi.org/10.7554/eLife.66668.

33. 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. https://doi.org/10.1371/journal.pone.0186855.

34. Koromila, T., Gao, F., Iwasaki, Y., He, P., Pachter, L., Gergen, J.P., and Stathopoulos, A. (2020). Odd-paired is a pioneer-like factor that coordinates with Zelda to control gene expression in embryos. eLife *9*, e59610. https://doi.org/10.7554/eLife.59610.

35. Starz-Gaiano, M., and Lehmann, R. (2001). Moving towards the next generation. Mech. Dev. *105*, 5–18. https://doi.org/10.1016/S0925-4773(01)00392-6.

36. Calderon, D., Blecher-Gonen, R., Huang, X., Secchia, S., Kentro, J., Daza, R.M., Martin, B., Dulja, A., Schaub, C., Trapnell, C., et al. (2022). The continuum of Drosophila embryonic development at single-cell resolution. Science *377*, eabn5800. https://doi.org/10.1126/science.abn5800.

37. Santos, A.C., and Lehmann, R. (2004). Germ Cell Specification and Migration in Drosophila and beyond. Curr. Biol. *14*, R578–R589. https://doi.org/10.1016/j.cub.2004.07.018.

38. 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. https://doi.org/10.1038/ncb1058.

39. 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. https://doi.org/10.1111/dgd.12607.

40. Shigenobu, S., Arita, K., Kitadate, Y., Noda, C., and Kobayashi, S. (2006). Isolation of germline cells from Drosophila embryos by flow cytometry. Dev. Growth Differ. *48*, 49–57. https://doi.org/10.1111/j.1440-169X.2006.00845.x.

41. Kofler, R., Nolte, V., and Schlötterer, C. (2015). Tempo and Mode of Transposable Element Activity in Drosophila. PLOS Genet. *11*, e1005406. https://doi.org/10.1371/journal.pgen.1005406.

42. 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. https://doi.org/10.1128/JVI.74.22.10658-10669.2000.

43. 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. https://doi.org/10.1093/nar/gkh708.

44. 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. https://doi.org/10.1126/science.1259114.

45. 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. https://doi.org/10.1186/s12977-017-0378-x.

46. 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. https://doi.org/10.1128/mbio.01733-21.

47. Burns, K.H. (2017). Transposable elements in cancer. Nat. Rev. Cancer *17*, 415–424. https://doi.org/10.1038/nrc.2017.35.

48. Dubnau, J. (2018). The Retrotransposon storm and the dangers of a Collyer's genome. Curr. Opin. Genet. Dev. *49*, 95–105. https://doi.org/10.1016/j.gde.2018.04.004.

49. 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. https://doi.org/10.1091/mbc.4.9.885.

50. Gilboa, L., and Lehmann, R. (2006). Soma-germline interactions coordinate homeostasis and growth in the Drosophila gonad. Nature *443*, 97–100. https://doi.org/10.1038/nature05068.

51. 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. https://doi.org/10.3390/cells9081776.

52. 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. https://doi.org/10.1093/nar/gkw784.

53. Gunawan, F., Arandjelovic, M., and Godt, D. (2013). The Maf factor Traffic jam both enables and inhibits collective cell migration in Drosophila oogenesis. Development *140*, 2808–2817. https://doi.org/10.1242/dev.089896.

54. Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al. (2023). Ensembl 2023. Nucleic Acids Res. *51*, D933–D941. https://doi.org/10.1093/nar/gkac958.

55. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

56. Hinrichs, A.S. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598. https://doi.org/10.1093/nar/gkj144.

57. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354. https://doi.org/10.1038/s41598-019-45839-z.

58. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics *26*, 2204–2207. https://doi.org/10.1093/bioinformatics/btq351.

59. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

60. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

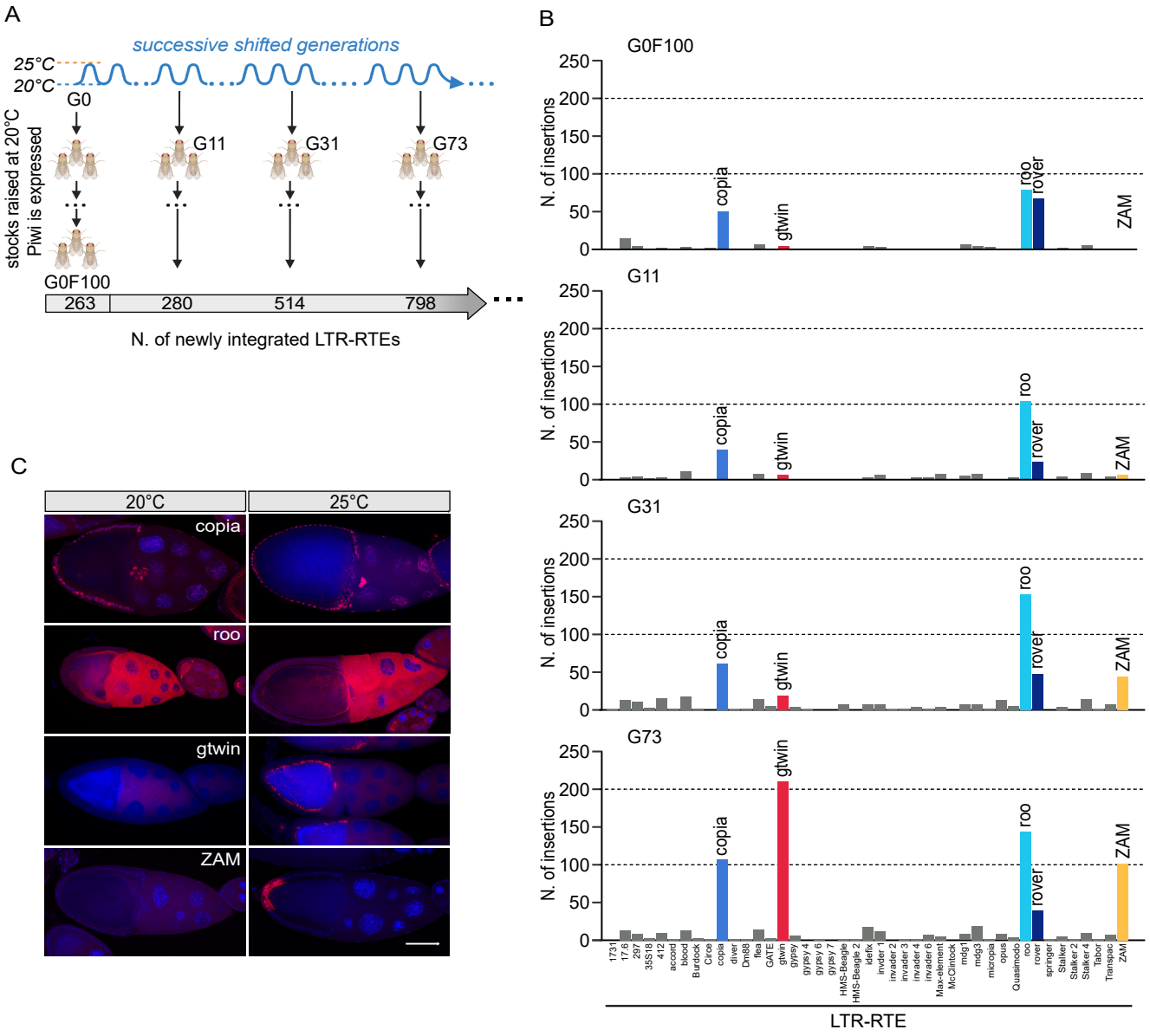

Fig.1: **Four LTR-RTEs integrate into the Drosophila germline after successive generations of somatic Piwi knockdown**

(A) Schematic representation of the successive transient Piwi somatic knockdowns (Piwi-sKD), induced by shifting adults from 20°C to 25°C for 5 days in each generation, followed by constant maintenance at 20°C of three large populations isolated from generations 11, 31 and 73. G0F100 corresponds to the initial G0 parental line that was constantly kept at 20°C for 100 generations. Bottom: The grey arrow contains the numbers of new LTR-RTE insertions detected in the different isolated populations as compared to the initial G0 parental line.

(B) Quantification of the new LTR-RTE insertions annotated in the control generation (G0F100) and the successive shifted generations G11, G31 and G73, as compared to the initial parental G0 line. (C) Representative images of stage 10 ovarian expression patterns obtained for copia, roo, gtwin and ZAM LTR-RTEs by smiFISH (in red) at the 20°C restrictive temperature and after a 5-day-adult shift to the 25°C permissive temperature.

DNA was stained with with 4,6-diamidino-2-phenylindole (DAPI; blue). Bar represents 50 µm.

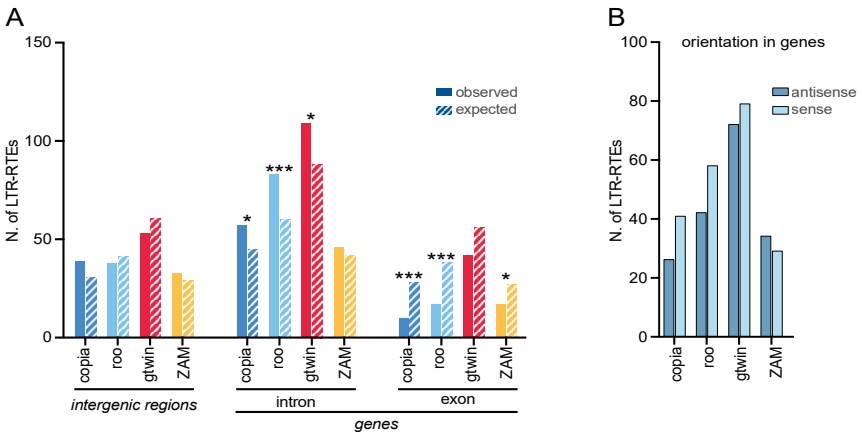

Fig. 2: **Selection had little impact on the landscape of newly integrated insertions**
(A) Barplots depicting the observed (filled) and expected (dashed) numbers of LTR-RTEs in the intergenic, intronic and exonic regions of the genome. Expected values were calculated according to the size of each genomic region considered and the total number of new insertions obtained for each LTR-RTE. 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. (B) Numbers of insertions for each LTR-RTE that had occurred into the antisense strand (dark blue) or sense strand (light blue) of protein-coding genes. No significative differences were observed between insertions in these two orientations using binomial tests corrected by Benjamini–Hochberg.

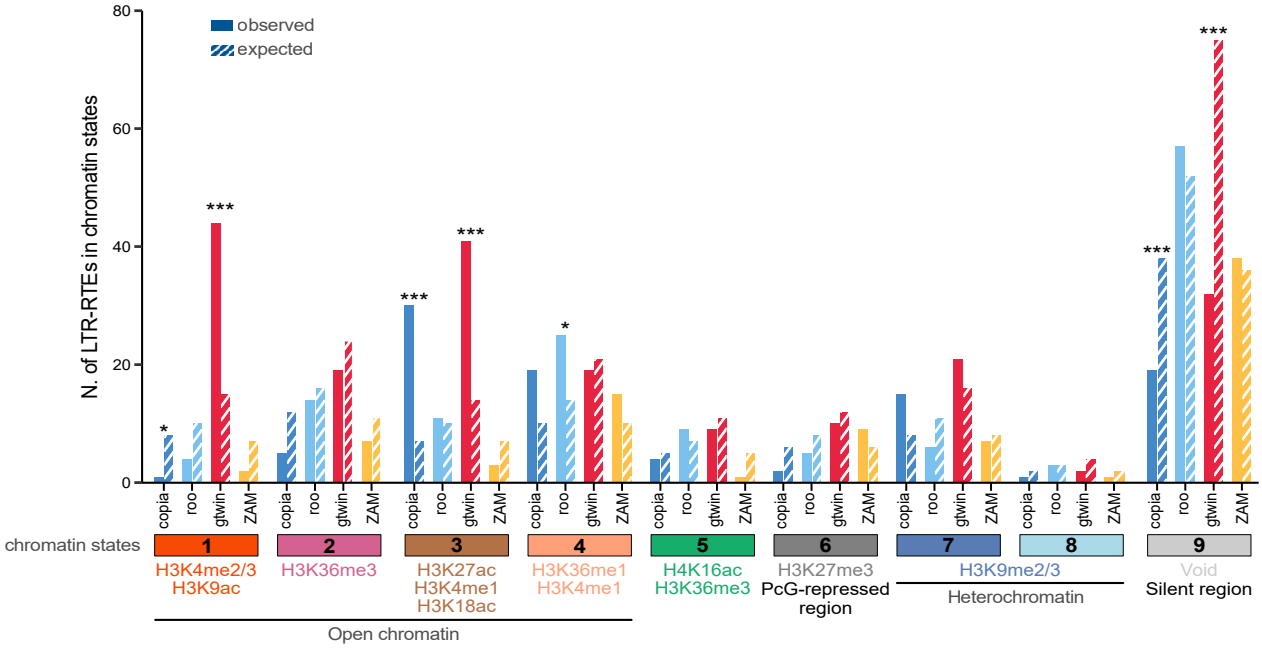

Fig. 3: **Each LTR-RTE species has its own specific chromatin states preferences for genomic integration**
Barplots depicting the observed (filled) and expected (dashed) numbers of LTR-RTEs in the nine described chromatin states. Typically enriched post-translational histone marks in each chromatin state are indicated below the corresponding 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.

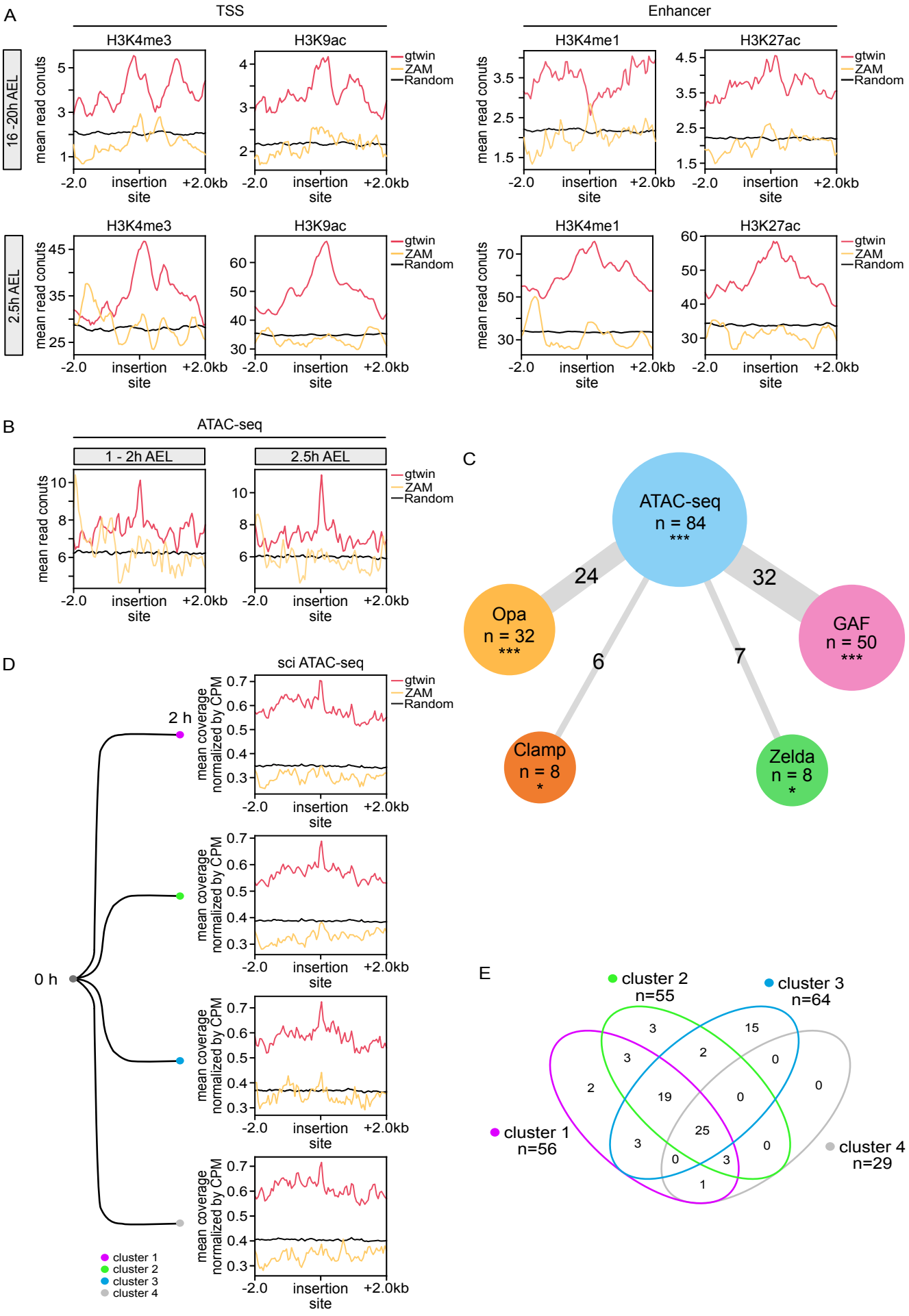
A
TSS
Enhancer
H3K4me3
H3K9ac
H3K4me1
H3K27ac
16 -20h AEL
2.5h AEL
mean read conuts
gtwin
ZAM
Random
-2.0
insertion site
+2.0kb
B
ATAC-seq
1 - 2h AEL
2.5h AEL
C
ATAC-seq
n = 84
***
Opa
n = 32
***
GAF
n = 50
***
Clamp
n = 8
*
Zelda
n = 8
*
24
32
6
7
D
sci ATAC-seq
2 h
0 h
mean coverage normalized by CPM
cluster 1
cluster 2
cluster 3
cluster 4
E
cluster 1
n=56
cluster 2
n=55
cluster 3
n=64
cluster 4
n=29
3
15
2
3
0
0
2
19
0
25
3
0
0
3
1

Fig. 4: **gtwin preferentially inserts into the open chromatin of enhancer and promoter regions before cellularization**

(A) Metaplots depicting the mean read counts of ChIP-seq data for H3K4me3, H3K9ac, H3K4me1 and H3K27ac obtained in embryos 16-20 h (top panels) and 2.5 h (bottom panels) after egg laying (AEL). 4-kb windows centered either on the 210 gtwin (red curve) or on the 101 ZAM (yellow curve) novel insertions found in G73 population.
Black curve (random) was obtained by averaging the mean read count values of 100 positions randomly selected 100 times on the same genome.
(B) Metaplots depicting the mean read counts obtained by ATAC-seq on embryos 1-2 h (left) and 2.5 h (right) AEL, considering the same 4 kb windows as in A.
(C) Overlap of pioneer-factor-rich and 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.
(D) Metaplots depicting in 4 kb window centered on 210 gtwin (red), 101 ZAM (yellow) or 100 random insertion sites (black), single-cell ATAC-seq (sci ATAC-seq) mean normalized coverage (by Counts Per Million unique mapped reads) of the 4 nuclei clusters identified in 0-2 h embryos.
(E) Venn diagram highlighting the distribution of the 76 gtwin insertions sites enriched in sci-ATAC-seq signal among the 4 clusters.

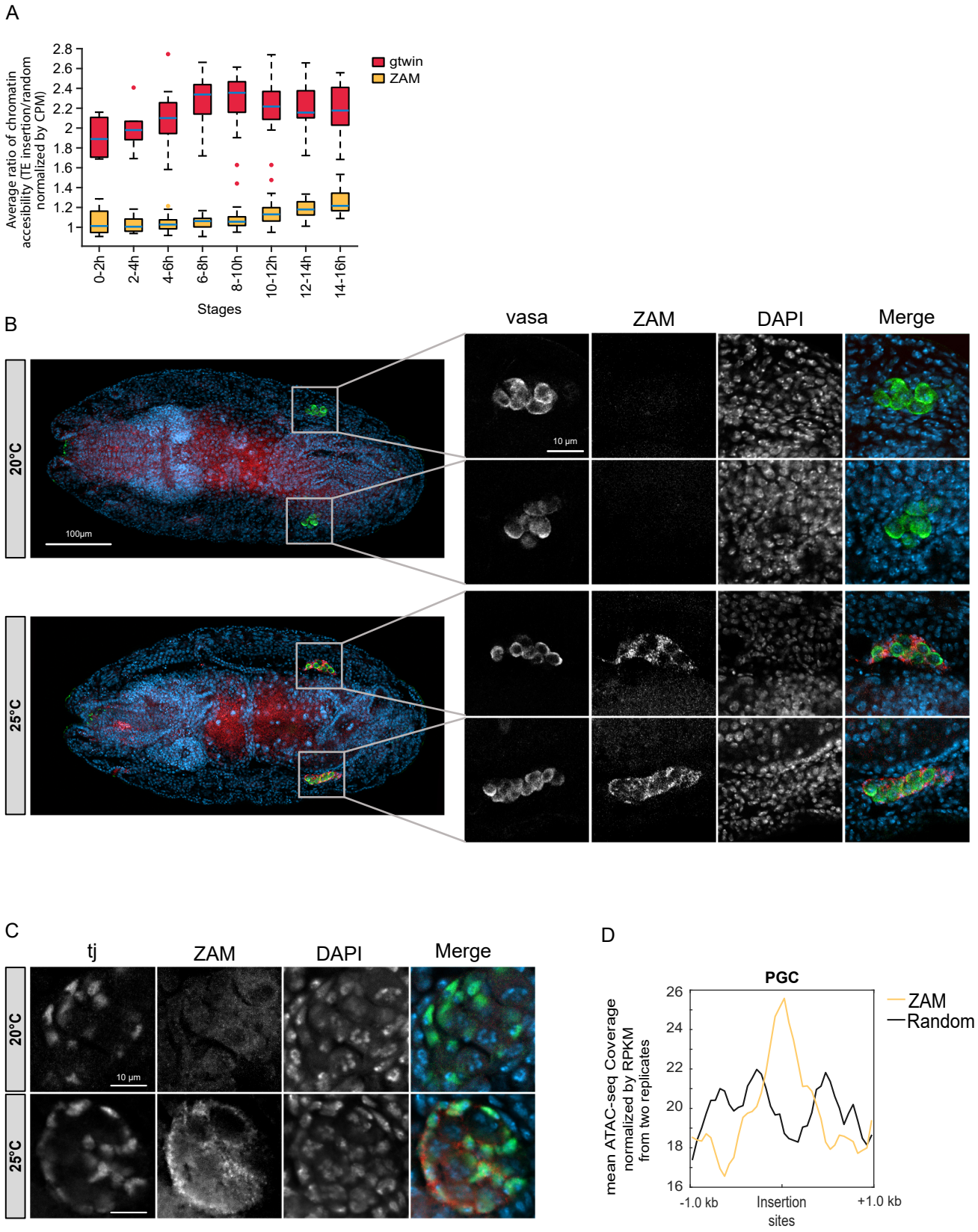
A
Average ratio of chromatin accesibility (TE insertion/random normalized by CPM)
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
gtwin
ZAM
0-2h
2-4h
4-6h
6-8h
8-10h
10-12h
12-14h
14-16h
Stages
B
vasa
ZAM
DAPI
Merge
20°C
25°C
100μm
10 μm
C
tj
ZAM
DAPI
Merge
20°C
25°C
10 μm
D
PGC
mean ATAC-seq Coverage normalized by RPKM from two replicates
26
24
22
20
18
16
ZAM
Random
-1.0 kb
Insertion sites
+1.0 kb

Fig. 5: **Chromatin accessibility of ZAM insertion sites correlates with ZAM late embryonic expression** (A) Boxplots representing the temporal kinetics of the average chromatin accessibility around gtwin (red) and ZAM (orange) insertions, relative to random profiles. At each time window is presented the distribution of the relative chromatin accessibility computed for the set of defined cell clusters. At each time window and for each cell cluster, this accessibility is defined as the average sci-ATAC-seq signal of the 200 bp windows centered on LTR-RTE insertions divided by the average signal of 100 randomly selected 200 bp windows. (B) Vasa immunostaining combined with ZAM smiFISH on 12- to 16-h whole-mount 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). (C) Traffic jam immunostaining combined with ZAM smiFISH on gonadal cells of 12- to 16-h embryos at 20°C (upper panel), 25°C (lower panel). The somatic primordial gonadal cells (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). (D) Metaplot showing, in 2kb windows centered on 101 ZAM (yellow) or 100 random insertion sites (black), the mean ATAC-seq signals of two replicates, normalized by coverage (by Counts Per Million unique mapped reads: RPKM), of PGCs purified from overnight embryos.