Article

Retrotransposons hijack alt-EJ for DNA replication and eccDNA biogenesis

https://doi.org/10.1038/s41586-023-06327-7

Received: 16 November 2022

Accepted: 14 June 2023

Published online: 12 July 2023

Check for updates

Fu Yang[1,5], Weijia Su[1,5], Oliver W. Chung[1], Lauren Tracy[1], Lu Wang[1,4], Dale A. Ramsden[2] & ZZ Zhao Zhang[1,3]✉

Retrotransposons are highly enriched in the animal genome[1–3]. The activation of retrotransposons can rewrite host DNA information and fundamentally impact host biology[1–3]. Although developmental activation of retrotransposons can offer benefits for the host, such as against virus infection, uncontrolled activation promotes disease or potentially drives ageing[1–5]. After activation, retrotransposons use their mRNA as templates to synthesize double-stranded DNA for making new insertions in the host genome[1–3,6]. Although the reverse transcriptase that they encode can synthesize the first-strand DNA[1–3,6], how the second-strand DNA is generated remains largely unclear. Here we report that retrotransposons hijack the alternative end-joining (alt-EJ) DNA repair process of the host for a circularization step to synthesize their second-strand DNA. We used Nanopore sequencing to examine the fates of replicated retrotransposon DNA, and found that 10% of them achieve new insertions, whereas 90% exist as extrachromosomal circular DNA (eccDNA). Using eccDNA production as a readout, further genetic screens identified factors from alt-EJ as essential for retrotransposon replication. alt-EJ drives the second-strand synthesis of the long terminal repeat retrotransposon DNA through a circularization process and is therefore necessary for eccDNA production and new insertions. Together, our study reveals that alt-EJ is essential in driving the propagation of parasitic genomic retroelements. Our study uncovers a conserved function of this understudied DNA repair process, and provides a new perspective to understand—and potentially control—the retrotransposon life cycle.

Retrotransposons abundantly occupy the genomes of nearly all animals, comprising almost 38% of human DNA[1–3]. During evolution, given their ability to mobilize and rewire gene regulation, retrotransposons bring one source of genome dynamics to rewriting DNA sequences. Within one generation, the nucleic acids generated during their replication cycles can be highly immunogenic and their mobilization step produces DNA breaks and generates mutations[1–3]. In the past, retrotransposon activation was largely considered to be deleterious to the hosts, by causing animal infertility, contributing to diseases, such as cancer, haemophilia or neurodegenerative disorders, and potentially driving ageing[1–3]. Notably, recent studies showed that programmed retrotransposon activation during animal development can offer benefits to the host, such as fending off invading viruses[4,5]. Despite the fundamental impacts brought from these parasitic genomic elements, how retrotransposons fulfil their life cycle to modulate host physiology and pathology remains unclear.

Our previous efforts established *Drosophila* oogenesis as a platform to precisely characterize retrotransposon activity at the mobilization level within an animal[7]. We found that retrotransposons rarely mobilize in germline stem cells[7], which, after differentiation, produce developing oocytes and supporting nurse cells[8]. Instead, retrotransposons use nurse cells as factories to massively manufacture themselves like viruses[7]. These retrotransposons then transport the virus-like particles into the oocyte and mobilize into the genome that will be transmitted to the next generation[7]. Leveraging this unique biological system, which enables us to spatiotemporally follow the retrotransposon activation process, we sought to characterize how retrotransposons generate new copies of integrated DNA and rewrite the host germline genome.

Engineered reporter mainly produces eccDNA

The *Drosophila* genome is enriched with both DNA transposons and retrotransposons, which comprise long terminal repeat (LTR) and non-LTR retrotransposons[9,10]. For these transposon families, very few can achieve mobilization into oocytes[7]. Among them, the LTR-retrotransposon *HMS-Beagle* displays the highest mobilization rate in oocytes[7,11]. To thoroughly investigate its mobilization, we generated a fly strain carrying one copy of eGFP-tagged *HMS-Beagle* (Extended Data Fig. 1a). Landing it into a specific site of the fly genome, this eGFP-tagged *HMS-Beagle* serves as the sole precursor for any newly integrated copies

[1]Department of Pharmacology and Cancer Biology, Duke University School of Medicine, Durham, NC, USA. [2]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [3]Duke Regeneration Center, Duke University School of Medicine, Durham, NC, USA. [4]Present address: State Key Laboratory of Molecular Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai, China. [5]These authors contributed equally: Fu Yang, Weijia Su. ✉e-mail: z.z@duke.edu

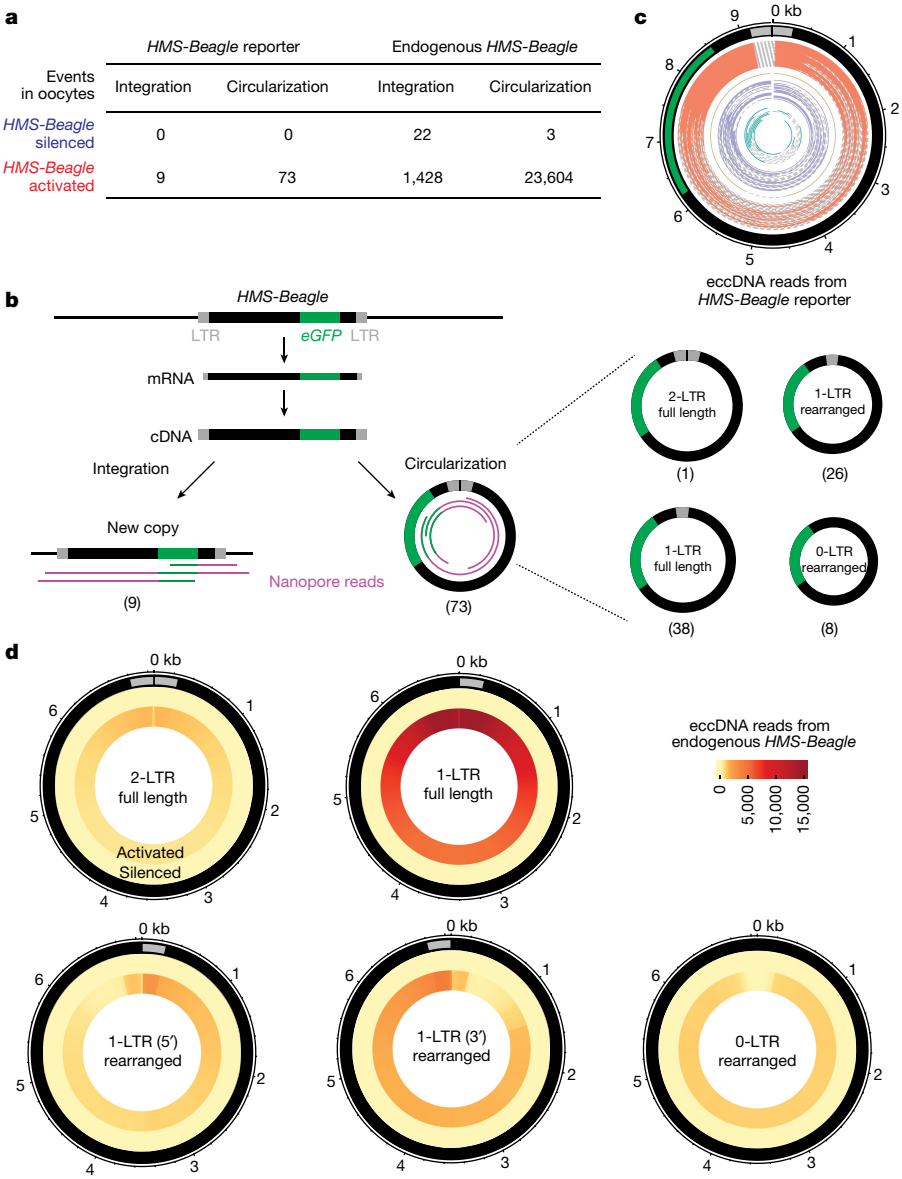


Fig. 1 | *HMS-Beagle* predominantly produces eccDNA after activation. **a**, Summary of the outcomes of replicated *HMS-Beagle* DNA detected in *Drosophila* oocytes. *HMS-Beagle* activation is achieved by suppressing Aub and AGO3 during oogenesis. **b**, The workflow to characterize the integration and circularization (eccDNA) events from an engineered *HMS-Beagle* reporter. The Nanopore sequencing reads were classified as integration when they had flanking sequences mapped to the genome or as eccDNA when they contained end-to-end junction sites. The eccDNA was further classified into four categories on the basis of their structures. The numbers within the parentheses denote the number of reads identified for each type of event. **c**, eccDNA reads from an engineered *HMS-Beagle* reporter. Each circle represents a read: the solid part represents the sequenced region and the dashed line represents the gap filled computationally. Salmon, reads supporting 1-LTR full-length circles; gold, one read supporting 2-LTR full-length circle; purple, reads supporting 1-LTR rearranged circles, probably resulting from autointegration; dark green, eccDNA reads do not contain intact LTR. **d**, Circos plot showing eccDNA reads from endogenous *HMS-Beagle*. The colour scale indicates the mapping coverage throughout the full-length *HMS-Beagle* consensus. The outer layer shows eccDNA reads from *HMS-Beagle*-silenced oocytes. The inner layer shows eccDNA reads from *HMS-Beagle*-activated oocytes. Data for this figure were generated from flies carrying short hairpin RNA targeting *aub* (sh-*aub*) and *ago3* (sh-*ago3*) to trigger transposon activation.

that contain eGFP sequences (Fig. 1a,b). To potentially capture the bona fide mobilization events from this tagged *HMS-Beagle* within oocytes, we sequenced their genome using Nanopore technology, which can directly read DNA up to the megabase scale without PCR amplification.

As expected, in the oocytes with transposons silenced (Extended Data Fig. 1b), the eGFP-tagged *HMS-Beagle* was detected only at its original landing site (Extended Data Fig. 1c). We next triggered transposon activation by depleting *aub* and *ago3* (Extended Data Fig. 1b), which are two key factors from a small RNA (PIWI-interacting RNA (piRNA))-based silencing system that suppresses transposons during *Drosophila* oogenesis[12–14]. Under this condition, we detected nine new insertions from the tagged *HMS-Beagle* with 28× genome coverage (Fig. 1a and Extended Data Fig. 1d), consistent with our previous finding that *HMS-Beagle* preferentially targets the oocyte genome for integration[7].

Notably, manually analysing the eGFP-derived reads that did not support integration indicated the formation of circular DNA. As we prepared genomic libraries using the Tn5 tagmentation method, this linearizes any circular DNA molecules. Sequencing such molecules enabled us to quantify these DNA circles by searching for reads covering the end-to-end junctions. By examining these events, we found that no circles formed when *HMS-Beagle* was silenced (Fig. 1a). However, after triggering its activation (Extended Data Fig. 1b), we

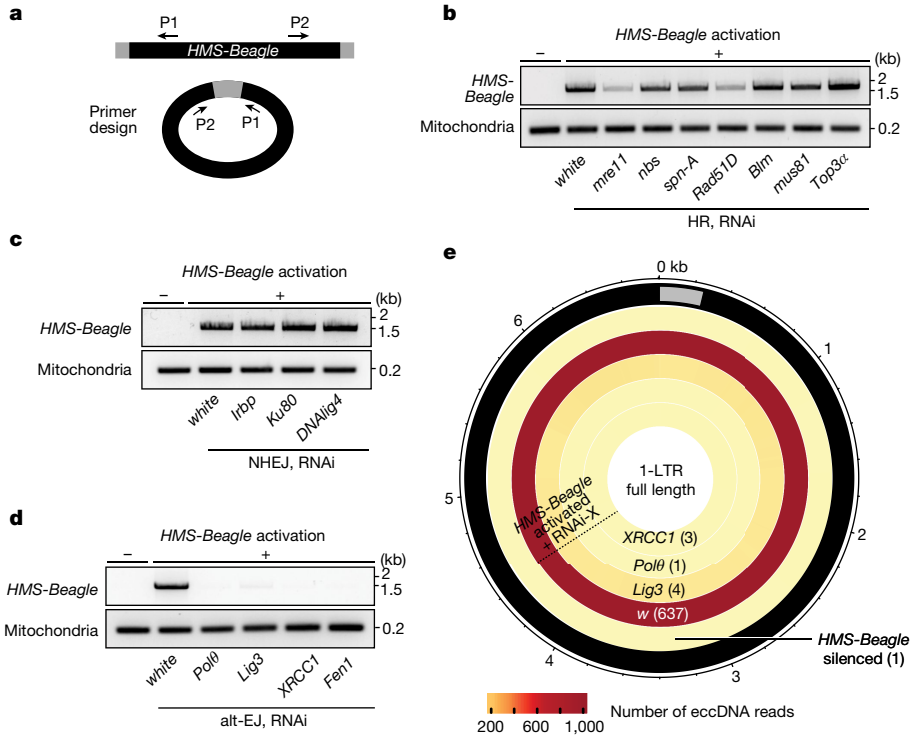


Fig. 2 | Factors from the alt-EJ process drive 1-LTR full-length eccDNA formation. a, Schematic of the design of divergent primers to identify *HMS-Beagle* eccDNA. **b**–**d**, Representative gel image showing whether components from the homologous recombination (HR) (**b**), NHEJ (**c**) or alt-EJ (**d**) process are required for the formation of 1-LTR full-length eccDNA. *white* is also known as *w*. **e**, Circos plot showing 1-LTR full-length eccDNA reads from endogenous *HMS-Beagle*. The colour scale indicates the mapping coverage throughout the full-length *HMS-Beagle* consensus. The numbers within the parentheses denote the number of reads identified from each genotype. Data for this figure were generated from flies carrying sh-*aub* to trigger transposon activation.

observed 73 reads that support the production of circular DNA from eGFP-tagged *HMS-Beagle* with 28-fold genome coverage (Fig. 1a,b), which is 8.1-fold more abundant than observed integration events. Given that *HMS-Beagle* has one LTR at each end, a head-to-tail circle would generate a junction read possessing two LTRs. However, among these 73 circle-supporting reads, only one has a junction with two LTRs (Fig. 1b,c), suggesting that the formation of full-length circles with two LTRs is a rare event. By contrast, 38 reads have junctions covering the start and end positions of the engineered *HMS-Beagle* reporter by containing one LTR (Fig. 1b,c), indicating that full-length circles with one LTR are the dominant circular form. Among the remaining 34 reads that encompass partial *HMS-Beagle* sequences (termed rearranged circles), 26 still have one intact LTR (termed 1-LTR rearranged). The remaining eight reads do not contain intact LTRs (termed 0-LTR rearranged). Given their circular nature, we designated these *HMS-Beagle*-derived circles eccDNA. Collectively, our data suggest that, after activation, our engineered *HMS-Beagle* abundantly produced eccDNA as 1-LTR full-length circles, but achieved far fewer integration events.

Endogenous copies preferentially form eccDNA

Our findings from the engineered *HMS-Beagle* prompted us to examine whether its endogenous copies also form circular DNA after activation. Accordingly, we mined our Nanopore sequencing data to characterize the reads supporting either integrated DNA or eccDNA events from endogenous *HMS-Beagle*. For control oocytes in which transposons are silenced, we detected 0.8 potential integrations and 0.1 potential eccDNA from endogenous *HMS-Beagle* per genome coverage (Fig. 1a). These integration events most likely reflect polymorphisms between the genome of the fly strain used in this study and the *Drosophila* reference genome, therefore defining the false-positive rate of our methodology on probing transposition events.

After transposon activation, we detected 1,428 integrations and 23,604 eccDNAs from endogenous *HMS-Beagle* loci with 28× genome coverage (Fig. 1a,d), highlighting that 94.3% of the replication products from *HMS-Beagle* form circles. Similar to the eGFP-tagged reporter, endogenous *HMS-Beagle* appears to also primarily form 1-LTR full-length circles: 54.7% of eccDNA reads support 1-LTR full-length circles, 6.4% of eccDNA reads indicate 2-LTR full-length circles and 38.9% of eccDNA reads are derived from rearranged *HMS-Beagle* circles (Fig. 1d). We concluded that, consistent with the observations from our reporter, endogenous *HMS-Beagle* also dominantly forms 1-LTR eccDNA.

To validate the formation of eccDNA, we designed a set of divergent PCR primers, which would give a PCR product only after the circularization of *HMS-Beagle* DNA (Fig. 2a and Extended Data Fig. 2). Given that *HMS-Beagle* replication occurs within the ovary during oogenesis[7], we reasoned that their eccDNA is readily generated within the ovaries before egg laying. By using ovary DNA as a template, performing PCR with divergent primers generated two products with distinct sizes (Extended Data Fig. 2). Sanger sequencing revealed that the upper faint band is derived from eccDNA with two LTRs (Extended Data Fig. 2). By contrast, the lower sharp band contains the PCR product from 1-LTR eccDNA (Extended Data Fig. 2). Thus, our data consistently indicate that, after activation, endogenous *HMS-Beagle* preferentially generates eccDNA, especially in the form of 1-LTR circles.

Sequencing genomic DNA by tagmentation can indicate the formation of eccDNA by capturing the end-to-end junctions. However, this method lacks the power to validate their circularity or to reconstruct the complete circular sequences of eccDNA. To obtain strong and direct evidence of circle formation, we established a Nanopore-based

eccDNA sequencing method (eccDNA-seq; Extended Data Fig. 3a). Our method appears to outperform recently published protocols for capturing large eccDNA. Whereas the published protocols produced reads with N50 < 5,000 bp[15,16], our method consistently generated reads with an N50 of around 15,000 bp (Supplementary Table 1). Applying eccDNA-seq to normal fly ovaries generated reads that mainly mapped to the mitochondrial genome (83.8% of total reads; Extended Data Fig. 3b), which is circular. Given that mitochondria take only around 0.6% of the Nanopore sequencing space when the total fly ovarian DNA is sequenced (Extended Data Fig. 3b), we concluded that our eccDNA-seq can enrich circular DNA by 139-fold. This highlights the high efficiency of our method on enriching circular DNA for sequencing. Spiking in a plasmid as an internal control for circular DNA quantification revealed that the amount of mitochondrial DNA remains unchanged after transposon activation (Extended Data Fig. 3c). Thus, the eccDNA-seq reads from transposons were normalized to the mitochondrial DNA reads across samples. In these control samples, *HMS-Beagle* generated very few, if any, eccDNAs (Extended Data Fig. 3d). By contrast, after its activation, we detected 13,362 *HMS-Beagle* eccDNAs (Extended Data Fig. 3d). Consistent with our findings from genomic sequencing and PCR-based methods, 61.32% of *HMS-Beagle* circles detected by eccDNA-seq were 1-LTR full-length circles (Extended Data Fig. 3d). In summary, our eccDNA-seq data provide strong and direct evidence of the formation of eccDNA from *HMS-Beagle* in vivo.

alt-EJ is required for eccDNA production

We next sought to understand the mechanism of eccDNA biogenesis, therefore potentially providing insights into the retrotransposon DNA replication cycle. Retrotransposons use their RNA transcripts as a template for reverse transcription to generate DNA for subsequent integration[17]. This replication intermediate could be the source for eccDNA biogenesis. This hypothesis predicts that depleting transposon RNA during oogenesis would abrogate eccDNA production. Indeed, after depleting *HMS-Beagle* RNA using RNA interference (RNAi; Extended Data Fig. 4a), *HMS-Beagle* eccDNA production was abolished (Extended Data Fig. 4b). Thus, our data indicate that retrotransposon-derived eccDNA is produced from their DNA replication intermediates, leaving their original genomic loci intact. This is different from the previously reported mechanisms on driving eccDNA formation, which involve either genomic DNA fragmentation or recombination within the genome[15,18–20].

For exogenous retroviruses and retroviral elements embedded in the host genome, it has been proposed that the homologous-recombination pathway can mediate the recombination of the two LTRs from the replicated linear DNA for the formation of 1-LTR circles[6,21]. To understand how *HMS-Beagle* forms 1-LTR eccDNA, we first tested the function of homologous-recombination machinery proteins during this process. However, after individually depleting seven key factors linked to this pathway during oogenesis—Nbs, Spn-A (*Drosophila* homologue of mammalian Rad51), Rad51D, Blm, Mre11, Mus81 and Top3a—*HMS-Beagle* still formed 1-LTR circles (Fig. 2b). These data argue against a previously proposed function from the homologous-recombination pathway in eccDNA biogenesis. Moreover, silencing of key factors from the non-homologous end joining (NHEJ) pathway (Irbp, Ku80 or DNA ligase 4) also had no effect on the formation of *HMS-Beagle*-derived eccDNA (Fig. 2c).

To systematically characterize the factors that are essential for *HMS-Beagle* eccDNA formation, we performed a candidate-based RNAi screen to individually deplete 123 factors (from 135 alleles) that are known to function in DNA repair or DNA damage response (Supplementary Table 2). After depleting each factor during oogenesis, we examined the production of *HMS-Beagle* 1-LTR circles using the PCR method that we established (Extended Data Fig. 2). Among these factors, 23 lead to lethality, impeding any further investigation (Supplementary Table 2). From the rest of the candidates, there were four factors screened as essential for *HMS-Beagle* 1-LTR eccDNA production: DNA polymerase θ (also known as Polθ, encoded by *PolQ* (also known as *Polθ*)), XRCC1, Lig3 (encoded by *DNAlig3* (also known as *Lig3*)) and Fen1 (Fig. 2d, Extended Data Fig. 5 and Supplementary Table 2). Notably, all four of these factors have been proposed to work coordinately for the alt-EJ DNA-repair process[22–24] (also known as the microhomology-mediated end-joining).

To further validate the function of alt-EJ factors on driving eccDNA production, we performed eccDNA-seq after individually depleting three of the identified factors: Polθ, XRCC1 and Lig3 (depleting Fen1 leads to semi-lethality, impeding obtaining enough DNA for sequencing). eccDNA-seq generated consistent data with the PCR results: silencing each of these alt-EJ factors completely abolished the biogenesis of 1-LTR eccDNA from *HMS-Beagle* (Fig. 2e).

Circularization for second-strand synthesis

Regarding how the alt-EJ process licenses eccDNA production from *HMS-Beagle*, it is possible that alt-EJ is required for transposon activation and, accordingly, is necessary for eccDNA production. To test this possibility, we performed Nanopore RNA sequencing (RNA-seq) and found the expression of *HMS-Beagle* transcripts is unaltered after depletion of alt-EJ factors (Extended Data Fig. 6). This suggests that the alt-EJ factors are not required for retrotransposon transcription. Instead, we provide evidence that alt-EJ is essential for the synthesis of *HMS-Beagle* second-strand DNA through a circularization process, and is therefore necessary for eccDNA biogenesis (Fig. 3a).

Previous research on yeast LTR-retrotransposons and retroviruses has laid the groundwork on the replication cycle of these elements[6,17,25,26]. They first transcribe genomic RNA[6]. Using the 3′ end of a tRNA to pair with its primer-binding site (PBS) sequence, the RNA transcripts are reverse transcribed into the first-strand DNA[6] (Fig. 3a (step 1)). After finishing the first-strand DNA synthesis, the reverse transcriptase uses its RNase H activity to digest most of the retroviral RNA from the DNA–RNA hybrid except a short RNA sequence at the 3′ end——the polypurine tract[6,27,28] (PPT; Fig. 3a (step 2)). The PPT functions as a primer to synthesize the 3′ LTR and the PBS sequence (using the tRNA as a template) for the second-strand DNA[6,28] (Fig. 3a (step 3)). To replicate the rest of the second-strand sequence that is upstream of the PPT site, it has been proposed that the PBS sequences between the first- and second-strand retroviral DNA can anneal with each other[6,29]. Known as the second-strand transfer (Fig. 3a (step 4)), this step converts the 3′ end of the second-strand DNA to the priming site for the synthesis of the rest of the strand[6,29]. Despite the essentiality of this step during the life cycle of retrotransposons and retroviruses, what mediates this process remains unclear.

Different from other DNA repair pathways, alt-EJ primes DNA synthesis by annealing a short homology (3–25 bp)[22,24]. Notably, the PBSs for retroviral elements are, in general, ≤18 nucleotides[6,30]. The apparent function of alt-EJ in mediating microhomology formation led us hypothesize that alt-EJ circularizes the two DNA strands by annealing their PBS homology (Fig. 3a). This circularization step initiates the subsequent second-strand DNA synthesis. This would produce a non-covalent circle with two fates: either fill the nick to dominantly generate covalent 1-LTR eccDNA or convert it into linear DNA with 2 LTRs (Fig. 3a). The linear DNA can serve as precursors for three subsequent outcomes: forming 2-LTR circles; using its LTR to attack its own interstitial sequences in *cis* to generate rearranged circles (known as autointegration); and inserting into host genomic DNA in *trans* for integration (Fig. 3a). Our model predicts that impeding the alt-EJ process would halt second-strand replication. This would lead to a higher single-stranded DNA ratio and the abolished biogenesis of the linear full-length double-stranded DNA, which serves as precursors for all downstream outcomes. To rigorously test our model, we correspondingly quantified the production of single-stranded DNA (Fig. 3b and Extended Data Fig. 7),

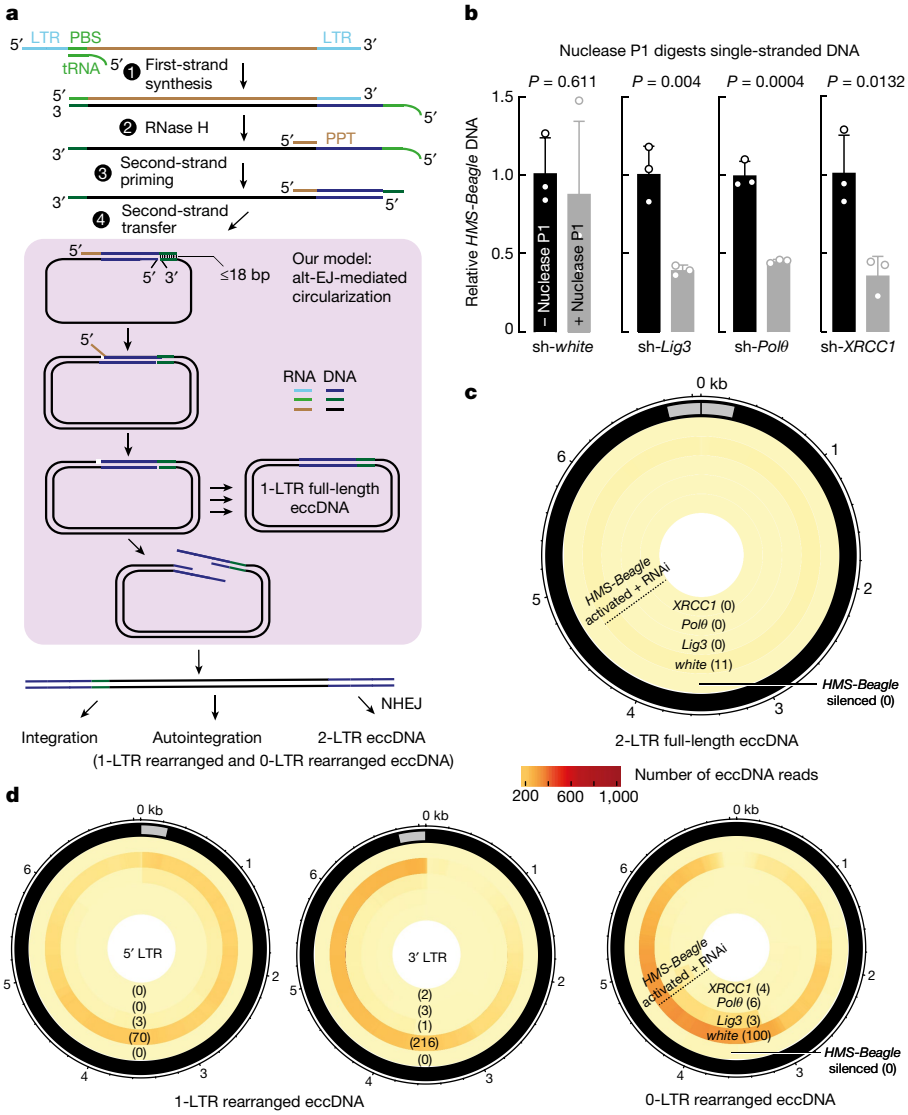


Fig. 3 | Blocking the alt-EJ process abrogates DNA synthesis and all eccDNA production from *HMS-Beagle*. **a**, A model depicting how alt-EJ-mediated circularization drives DNA synthesis and is therefore essential for eccDNA production and mobilization. Step 1: a tRNA fragment pairs with the primer binding site (PBS) to initiate the first-strand DNA synthesis through reverse transcription to form a RNA–DNA hybrid. Step 2: RNase H activity removes RNA from the RNA–DNA hybrid, but leaves a polypurine tract (PPT). Step 3: the PPT initiates the second-strand DNA synthesis for the 3′ LTR. Step 4: alt-EJ-mediated circularization drives the synthesis of the remaining of the second-strand DNA. **b**, Quantitative PCR (qPCR) analysis of the relative abundance of single-stranded *HMS-Beagle* DNA. Data are mean ± s.d. from three biological replicates ($n = 3$). *P* values were calculated using two-tailed two-sample unequal-variance *t*-tests. **c**, Circos plot showing 2-LTR full-length eccDNA reads from endogenous *HMS-Beagle*. The colour scale indicates the mapping coverage throughout the full-length *HMS-Beagle* consensus. The numbers within the parentheses denote the number of reads identified from each genotype. These circles are probably generated by joining the two LTRs together through NHEJ. **d**, Circos plot showing rearranged eccDNA reads from endogenous *HMS-Beagle*. The colour scale indicates the mapping coverage throughout the full-length *HMS-Beagle* consensus. The numbers within the parentheses denote the number of reads identified for each genotype. 1-LTR rearranged circles are probably generated by autointegration events, with the LTR attacking its own interstitial sequences in *cis*. 0-LTR rearranged circles are possibly the by-products of autointegration. Data for this figure were generated from flies carrying sh-*aub* to trigger transposon activation.

2-LTR eccDNA (Fig. 3c), rearranged eccDNA (Fig. 3d), linear full-length double-stranded DNA (Fig. 4a) and integration events after the depletion of alt-EJ factors (Fig. 4b).

First, we examined whether *HMS-Beagle* DNA extracted from the alt-EJ-perturbated ovaries is more sensitive to the treatment of nuclease P1, an endonuclease that digests single-stranded DNA. Indeed, after treatment with nuclease P1, whereas the levels of *HMS-Beagle* DNA from the control ovaries remained unchanged, silencing Polθ, Lig3 or XRCC1 led to a greater than twofold reduction in *HMS-Beagle* DNA (Fig. 3b). To further measure the amount of single-stranded DNA, we performed immunoprecipitation using Mab3034 antibodies, which preferentially bind to single-stranded DNA[31]. After individually depleting alt-EJ factors, the amount of *HMS-Beagle* single-stranded DNA increased significantly. Together, these data suggest that the alt-EJ process is essential for the completion of the second-strand synthesis to produce double-stranded DNA.

We next used the Nanopore ligation-based sequencing method to directly examine the biogenesis of linear full-length double-stranded DNA. Whereas triggering transposon activation resulted in the production of double-stranded DNA with two intact LTRs flanking each end (Fig. 4a), individually silencing Polθ or XRCC1 completely abolished the formation of this essential precursor for all downstream events (Fig. 4a),

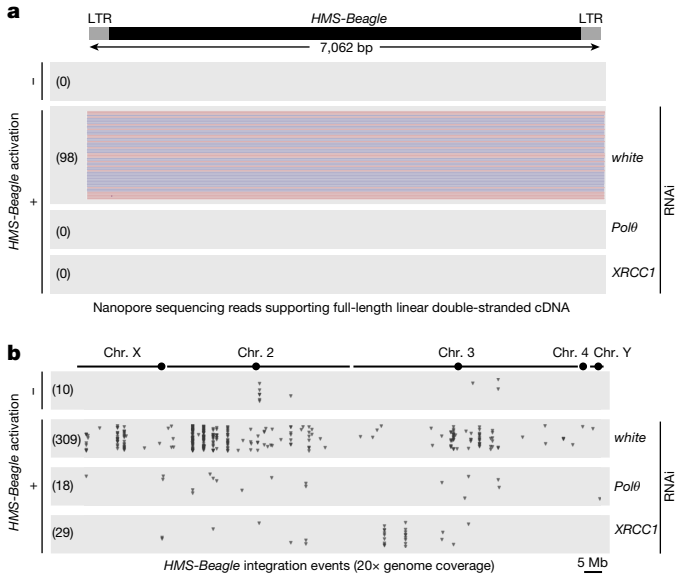


Fig. 4 | Blocking alt-EJ process abrogates *HMS-Beagle* mobilization. **a**, Nanopore sequencing analysis of the biogenesis of full-length double-stranded linear DNA from replicated *HMS-Beagle*. Red and blue reads reflect the sequenced plus and minus strands, respectively. **b**, The new integrations from *HMS-Beagle*. Each triangle represents an integration event detected by Nanopore genome sequencing. The numbers in the parentheses present the total amounts of integration events detected. The numbers of integrations detected under the transposon-silenced condition probably represent the false-positive rates from our methodology. Data for this figure were generated from flies carrying sh-*aub* to trigger transposon activation.

such as the biogenesis of 2-LTR and rearranged eccDNA or integration events. Our data further suggest that alt-EJ drives the conversion process from single-stranded to double-stranded DNA.

Furthermore, we quantified the production of 2-LTR and rearranged eccDNA after silencing alt-EJ factors (Fig. 3c,d). Here we report the eccDNA-seq reads by normalizing them to mitochondrial genome coverage. For 2-LTR full-length circles, we detected 11 of them from *HMS-Beagle* after triggering its activation (Fig. 3c). However, individually silencing Polθ, Lig3 or XRCC1 completely abolished their biogenesis (Fig. 3c). Similarly, the number of rearranged circles also dropped to the background level after suppressing the alt-EJ process (Fig. 3d). Triggering *HMS-Beagle* activation generated 286 rearranged eccDNAs containing one intact LTR (1-LTR rearranged circles), indicative of autointegration events. Meanwhile, there were 100 rearranged *HMS-Beagle* circles that did not contain intact LTRs (0-LTR rearranged circles; Fig. 3d), which were probably generated as the by-products of autointegration events. However, for both categories of rearranged circles, we detected ≤6 circles with Polθ, Lig3 or XRCC1 silenced (Fig. 3d). Our data therefore further support a function of the alt-EJ pathway in the DNA replication process of LTR retrotransposons.

Finally, we tested whether impeding the alt-EJ process would abrogate not only eccDNA production, but also *HMS-Beagle* mobilization. To test this, we individually depleted Polθ or XRCC1 (flies with Lig3 silenced lay very few mature oocytes), and then examined the transposon mobilization rates in oocytes. For each genotype, DNA from somatic carcasses was sequenced to construct individual genomes, which serve as the reference to precisely define new transposon integrations in oocytes. With the same genome coverage (20×), we detected 309 new insertion events from *HMS-Beagle* when the alt-EJ process was undisturbed (Fig. 4b). However, once this process was blocked to abolish eccDNA biogenesis, the new insertion events also substantially decreased: 18 insertions after Polθ depletion and 29 insertions after XRCC1 depletion (Fig. 4b). Collectively, our data indicate that alt-EJ is essential for the generation of double-stranded *HMS-Beagle* DNA that serves as a precursor for both circularization and integration.

Conserved function of alt-EJ

Besides using piRNA pathway perturbation to study the function of alt-EJ on retrotransposon DNA replication during oogenesis, we sought to further investigate its role under normal developmental

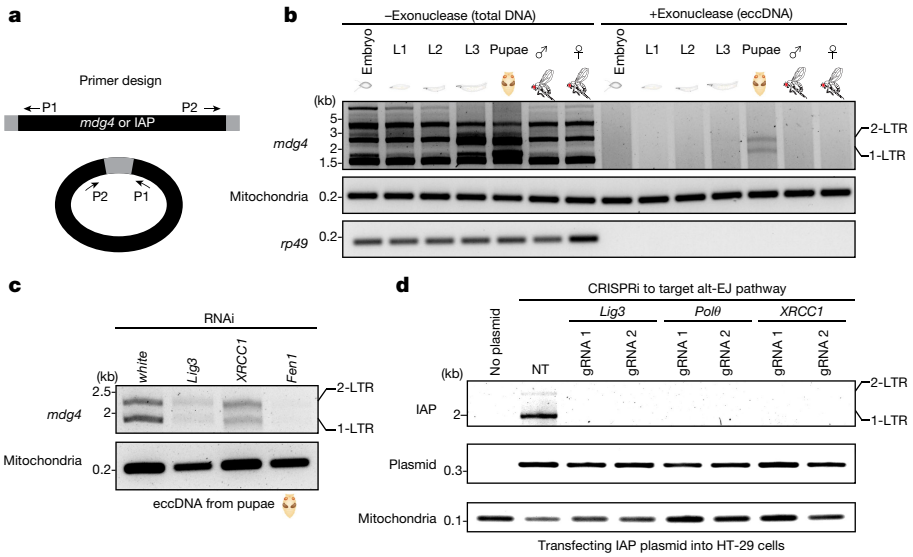


Fig. 5 | *mdg4* and mammalian IAP form eccDNA through the alt-EJ factors. **a**, Schematic of the divergent primers to detect *mdg4* or IAP eccDNA. **b**, The *mdg4* retrotransposon produces both 1-LTR and 2-LTR eccDNA at the pupal stage, the time window during which mobilization occurs. Performing PCR using total DNA as template produced non-specific bands. Using exonuclease to enrich eccDNA generated two PCR products corresponding to 1-LTR and 2-LTR eccDNA, as confirmed by Sanger sequencing. L1–L3, larval stages 1–3. **c**, eccDNA production from *mdg4* depends on alt-EJ factors. **d**, Suppressing the factors from alt-EJ repair process blocks IAP eccDNA biogenesis. The PCR products corresponding to 1-LTR and 2-LTR eccDNA were confirmed by Sanger sequencing. The non-targeting (NT) gRNA is a random gRNA without a targeting site in the human genome.

conditions. We recently found that the retrotransposon *mdg4* (also known as *Gypsy*), naturally mobilizes in somatic tissues[5]. Particularly, *mdg4* appears to mobilize only at the pupal stage[5], when flies are regenerating new somatic tissues for adulthood. Accordingly, as an indication of the completion of the *mdg4* DNA replication, we monitored *mdg4* eccDNA production at different developmental stages. We found that *mdg4* specifically generated eccDNA at the pupal stage, but not other developmental stages (Fig. 5a,b), consistent with the time window during which mobilization events are detected. Notably, silencing the alt-EJ factors suppressed *mdg4* eccDNA production (Fig. 5c). These results suggest that the alt-EJ pathway is also essential for retrotransposon replication in somatic tissues.

Next, we examined whether mobile elements from different species also use the alt-EJ process for their DNA replication. By using eccDNA production as a readout, we investigated the function of alt-EJ in the replication cycle of intracisternal A-particle (IAP), a mouse LTR retrotransposon. IAP presents around 2,800 full-length copies in the mouse genome and its activation contributes to about 6% of all pathogenic mutations[32]. To unambiguously examine IAP activity, previous studies established a procedure to introduce IAP into cultured human cells[33,34], which do not contain IAP in their own genome. According to this procedure, we monitored IAP eccDNA formation in human cells and found that IAP indeed generated circular DNA, including both 1-LTR and 2-LTR circles (Fig. 5d). Notably, disrupting the reverse transcriptase activity, but not the integrase function, leads to abolished biogenesis of both 1-LTR and 2-LTR eccDNA (Extended Data Fig. 8a–c). Using eccDNA production as a readout, we next examined whether alt-EJ is essential for IAP DNA replication. Notably, after individually depleting the human orthologues of the factors identified in *Drosophila* that function in the retrotransposon life cycle (*Polθ*, *XRCC1* and *DNAlig3*; Extended Data Fig. 8d–f), IAP did not manufacture both 1-LTR and 2-LTR eccDNA (Fig. 5d). Meanwhile, suppressing NHEJ blocks only 2-LTR eccDNA production (Extended Data Fig. 9), suggesting that the 2-LTR eccDNA is formed by joining the two ends of the replicated linear double-stranded precursors. Together, these findings suggest a conserved function of alt-EJ in driving the retrotransposon replication cycle in Metazoa.

Discussion

The current view posits that retroviral elements synthesize their DNA largely for integration, and the remaining unintegrated ones can form 1-LTR eccDNA through homologous recombination between the 2 LTRs[2,3,6,21,35]. Although these models were proposed more than four decades ago and have since been extensively cited[16,17,28], the direct evidence to support them is still lacking. Our data do not support these models. Instead, by combining the power of Nanopore long-read sequencing with our robust genetic system, we report that retrotransposons hijack alt-EJ-mediated circularization to dominantly produce 1-LTR eccDNA, but achieve far fewer integrations (Extended Data Fig. 10). Instead of merely being produced as the dead-end by-products, these circles can potentially serve some biological purposes, such as transcribing mRNA to initiate a new round of replication cycle or breaking internally and then inserting into the genome to rewrite the genetic information. Moreover, given that circular DNA is highly potent for inducing innate immunity[15], it is possible that retrotransposon-derived eccDNAs can function as immune regulators. Notably, we recently found that, during the developmental time window during which the *mdg4* eccDNA is manufactured, *mdg4* activation licences the host's immune system for future antiviral responses[5].

alt-EJ was initially viewed as merely a backup pathway for canonical DNA repair[22,24]. By performing genetic screens in vivo, here we uncovered its conserved function in the replication cycle of parasitic genetic mobile elements. Expression of both alt-EJ factors and retrotransposons is tightly controlled. Notably, both of them appear to maintain high activity during embryogenesis, or under ageing or pathological progression, such as cancer[2,22,24,36–40]. Under these conditions, alt-EJ probably drives the replication of retrotransposon DNA, enabling eccDNA production and mobilization from retrotransposons, thereby ultimately contributing to disease progression or driving evolution.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06327-7.

1. Wells, J. N. & Feschotte, C. A field guide to eukaryotic transposable elements. *Ann. Rev. Genet.* **54**, 539–561 (2020).
2. Kazazian, H. H. Jr. & Moran, J. V. Mobile DNA in health and disease. *New Engl. J. Med.* **377**, 361–370 (2017).
3. Fueyo, R., Judd, J., Feschotte, C. & Wysocka, J. Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.* **23**, 481–497 (2022).
4. Frank, J. A. et al. Evolution and antiviral activity of a human protein of retroviral origin. *Science* **378**, 422–428 (2022).
5. Wang, L. et al. Retrotransposon activation during *Drosophila* metamorphosis conditions adult antiviral responses. *Nat. Genet.* **54**, 1933–1945 (2022).
6. Telesnitsky, A. & Goff, S. P. in *Retroviruses* (eds Coffin, J. M. et al.) 121–160 (Cold Spring Harbor Laboratory Press, 1997).
7. Wang, L., Dou, K., Moon, S., Tan, F. J. & Zhang, Z. Z. Hijacking oogenesis enables massive propagation of LINE and retroviral transposons. *Cell* **174**, 1082–1094 (2018).
8. Xie, T. & Spradling, A. C. A niche maintaining germ line stem cells in the *Drosophila* ovary. *Science* **290**, 328–330 (2000).
9. Kaminker, J. S. et al. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.* **3**, research0084.1 (2002).
10. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
11. Lammel, U. & Klambt, C. Specific expression of the *Drosophila* midline-jumper retrotransposon in embryonic CNS midline cells. *Mech. Dev.* **100**, 339–342 (2001).
12. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.* **12**, 246–258 (2011).
13. Li, C. et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137**, 509–521 (2009).
14. Vagin, V. V. et al. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**, 320–324 (2006).
15. Wang, Y. et al. eccDNAs are apoptotic products with high innate immunostimulatory activity. *Nature* **599**, 308–314 (2021).
16. Henriksen, R. A. et al. Circular DNA in the human germline and its association with recombination. *Mol. Cell* **82**, 209–217 (2022).
17. Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40**, 491–500 (1985).
18. Shoshani, O. et al. Chromothripsis drives the evolution of gene amplification in cancer. *Nature* **591**, 137–141 (2021).
19. Libuda, D. E. & Winston, F. Amplification of histone genes by circular chromosome formation in *Saccharomyces cerevisiae*. *Nature* **443**, 1003–1007 (2006).
20. Moller, H. D. et al. Formation of extrachromosomal circular DNA from long terminal repeats of retrotransposons in *Saccharomyces cerevisiae*. *G3* **6**, 453–462 (2015).
21. Brown, P. O. in *Retroviruses* (eds Coffin, J. M. et al.) 161–204 (Cold Spring Harbor Laboratory Press, 1997).
22. Brambati, A., Barry, R. M. & Sfeir, A. DNA polymerase theta (Polθ)—an error-prone polymerase necessary for genome stability. *Curr. Opin. Genet. Dev.* **60**, 119–126 (2020).
23. Mateos-Gomez, P. A. et al. Mammalian polymerase θ promotes alternative NHEJ and suppresses recombination. *Nature* **518**, 254–257 (2015).
24. Ramsden, D. A., Carvajal-Garcia, J. & Gupta, G. P. Mechanism, cellular functions and cancer roles of polymerase-theta-mediated DNA end joining. *Nat. Rev. Mol. Cell Biol.* **23**, 125–140 (2022).
25. Lauermann, V. & Boeke, J. D. Plus-strand strong-stop DNA transfer in yeast Ty retrotransposons. *EMBO J.* **16**, 6603–6612 (1997).
26. Heyman, T., Agoutin, B., Friant, S., Wilhelm, F. X. & Wilhelm, M. L. Plus-strand DNA synthesis of the yeast retrotransposon Ty1 is initiated at two sites, PPT1 next to the 3′ LTR and PPT2 within the *pol* gene. PPT1 is sufficient for Ty1 transposition. *J. Mol. Biol.* **253**, 291–303 (1995).
27. Tanese, N., Telesnitsky, A. & Goff, S. P. Abortive reverse transcription by mutants of Moloney murine leukemia virus deficient in the reverse transcriptase-associated RNase H function. *J. Virol.* **65**, 4387–4397 (1991).
28. Finston, W. I. & Champoux, J. J. RNA-primed initiation of Moloney murine leukemia virus plus strands by reverse transcriptase in vitro. *J. Virol.* **51**, 26–33 (1984).
29. Rhim, H., Park, J. & Morrow, C. D. Deletions in the tRNA(Lys) primer-binding site of human immunodeficiency virus type 1 identify essential regions for reverse transcription. *J. Virol.* **65**, 4555–4564 (1991).
30. Le Grice, S. F. "In the beginning": initiation of minus strand DNA synthesis in retroviruses and LTR-containing retrotransposons. *Biochemistry* **42**, 14349–14355 (2003).

31. Hu, Z., Leppla, S. H., Li, B. & Elkins, C. A. Antibodies specific for nucleic acids and applications in genomic detection and clinical diagnostics. *Expert Rev. Mol. Diagn.* **14**, 895–916 (2014).
32. Gagnier, L., Belancio, V. P. & Mager, D. L. Mouse germ line mutations due to retrotransposon insertions. *Mobile DNA* **10**, 15 (2019).
33. Dewannieux, M., Dupressoir, A., Harper, F., Pierron, G. & Heidmann, T. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nat. Genet.* **36**, 534–539 (2004).
34. Schorn, A. J., Gutbrod, M. J., LeBlanc, C. & Martienssen, R. LTR-retrotransposon control by tRNA-derived small RNAs. *Cell* **170**, 61–71 (2017).
35. Shank, P. R. et al. Mapping unintegrated avian sarcoma virus DNA: termini of linear DNA bear 300 nucleotides present once or twice in two species of circular DNA. *Cell* **15**, 1383–1395 (1978).
36. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).
37. Wang, J. et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**, 405–409 (2014).
38. Macfarlan, T. S. et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
39. Pang, M., McConnell, M. & Fisher, P. A. The *Drosophila mus308* gene product, implicated in tolerance of DNA interstrand crosslinks, is a nuclear protein found in both ovaries and embryos. *DNA Repair* **4**, 971–982 (2005).
40. Vaidya, A. et al. Knock-in reporter mice demonstrate that DNA repair by non-homologous end joining declines with age. *PLoS Genet.* **10**, e1004511 (2014).

Methods

Fly strains, housing and husbandry conditions

All flies were grown on standard agar-corn medium. Female flies aged 3–7 days were selected for experiments unless otherwise noted. Fly alleles used for the genetic screens are listed in Supplementary Table 2 and the rest of the alleles used in this study are listed in Supplementary Table 3. Flies carrying *vas-Gal4* were used in this study to achieve germline-specific gene silencing. For Fig. 1 and Extended Data Figs. 1–4, sh-*aub* and sh-*ago3* double RNAi flies were used for genome sequencing (genome-seq) and eccDNA-seq. These flies were maintained at 25 °C. For Figs. 2–4 and Extended Data Figs. 5 and 6, only sh-*aub* was used to block the piRNA pathway. Meanwhile, to facilitate the genetic screen, *tub-Gal80^{ts}* was introduced into the genetic background to achieve conditional RNAi silencing. For the targeted screening, crosses were set at 18 °C for 3 days and then shifted to 29 °C to activate the RNAi constructs. For the Oxford Nanopore cDNA-seq and eccDNA-seq experiments, crosses were set and kept at 18 °C for 9 days then shifted to 29 °C for 7 days. For the genome-seq analysis of oocytes, F_1 virgins with the desired genotypes were collected and crossed with *w^{1118}* males. F_2 embryos laid within 6 h were collected for DNA extraction. To detect *mdg4* eccDNA from different developmental stages, the mixed genders of *ac5c-Gal4* > sh-*white* flies from embryo to pupa stages and 2–5 days old adult male and female flies were collected, respectively (Fig. 5b). For Fig. 5c, flies carrying *ac5c*-Gal4 were used to silence the indicated factors.

eGFP reporter and shHMS-Beagle transgenic flies

The construct of the *HMS-Beagle* transposition reporter was generated using the Counter-Selection BAC Modification Kit (Gene Bridges, K002). The BAC clone p[acman]-CH322-33A08 was used in this study as a template. The eGFP reporter was landed into position 6242 of *HMS-Beagle*.

To construct the plasmid for *HMS-Beagle* silencing, DNA fragments of short hairpin RNA were synthesized (a list of the sequences is provided in Supplementary Table 4) and cloned into the NheI and EcoRI sites of VALIUM20. All of the constructs were verified by colony PCR and Sanger sequencing. All of the plasmids were site-specifically landed into the fly genome at the *attP2* site.

RNA-FISH

The Stellaris RNA FISH probe set for *HMS-Beagle* was from a previous study and RNA-FISH experiments were performed as described previously[7]. In brief, three pairs of ovaries were dissected in cold PBS and fixed for 20 min in 4% formaldehyde. Ovaries were washed once with PBST and twice with PBS, and then immersed in 70% (v/v) ethanol for 8 h at 4 °C. The ovaries were then washed once with wash buffer A (LGC Biosearch Tech, SMF-WA1-60) at room temperature for 5 min, then incubated with 50 ml hybridization buffer (LGC Biosearch Tech, SMF-HB1-10) containing the probe set (125 nM) for hybridization overnight at 37 °C. Next, the ovaries were washed twice with wash buffer A for 30 min at 37 °C and once with wash buffer B (LGC Biosearch Tech, SMF- WB1-20) for 5 min at room temperature. The samples were mounted with 20 µl Vectashield Mounting Medium (Vector Laboratories, 101098-042). Images were acquired on the Leica SP5 inverted microscope. All of the images were assembled in Adobe Photoshop and Illustrator.

eccDNA-seq library preparation and Oxford Nanopore sequencing

For eccDNA sequencing, total DNA from ovaries was extracted using the Quick-gDNA MicroPrep Kit (Zymo Research, D3021). After removing linear DNA and performing rolling-circle amplification and debranching, the library was prepared using the Ligation Sequencing Kit (Oxford Nanopore, SQK-LSK109). In detail, 2 µg of total DNA was mixed with 2 µl Plasmid-Safe DNase (Lucigen, E3110K), 5 µl 10× Plasmid-Safe buffer, 2 µl 100 mM ATP (Thermo Fisher Scientific, R0441) and ultrapure water (Thermo Fisher Scientific, 10977023) to 50 µl. On a thermocycler machine, the mixture was incubated at 37 °C for 3 h. Then, 2 µl Plasmid-Safe DNase and 1 µl ATP were added to the mixture. The mixture was further incubated at 37 °C for 16 h and 70 °C for 30 min on a thermocycler machine. Then, 50 µl AMPure XP beads (Beckman Coulter, A63881) was used to purify the DNA. The concentration of the purified circular DNA was measured using the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific, Q33231). The RCA reaction was set as follows: 2 ng circular DNA, 5 µl 10× Phi29 DNA polymerase buffer, 1 µl Phi29 DNA Polymerase (New England Biolabs, M0269L), 2.5 µl 10 mM dNTP (Qiagen, 201901), 2.5 µl Exo Resistant Random Primer (Thermo Fisher Scientific, SO181) and ultrapure water to 50 µl. The mixture was incubated at 30 °C for 16 h and 65 °C for 10 min on a thermocycler machine. The RCA product was purified by iso-propanol precipitation and debranched by T7 Endonuclease I (New England Biolabs, 0302L). The short fragments were eliminated using the Short-Read Eliminator XL kit (Circulomics, SS-100-111-01). The library was then constructed according to the Oxford Nanopore SQK-LSK109 protocol. All of the libraries were sequenced in R9.4 flow cells on a GridION instrument according to the manufacturer's instructions.

AFM imaging

To prepare the sample for AFM imaging, 2–5 ng DNA was diluted in low-salt buffer (25 mM HEPES pH 7.5, 10 mM $MgCl_2$ and 50 mM NaCl) to a total volume of 10 µl. The entire mixture was deposited onto a freshly cleaved mica surface (Ted Pella, 50) and incubated for 1 min before being rinsed three times with 30 µl ultrapure water. The mica surface was then dried using compressed air. Imaging was performed using the Asylum Cypher Atomic Force Microscope equipped with an AC240TS-R3 probe (Oxford Instruments, 803.OLY.AC240TS-R3) in ACMoleculeAir mode, and images were processed using Gwyddion (v.2.52).

Nanopore RNA-seq library preparation

For RNA-seq, fly ovaries were dissected on ice-cold PBS. The poly(A)$^+$ RNA was extracted using the Magnetic mRNA Isolation Kit (New England Biolabs, S1550S) according to the manufacturer's instructions with the following small modification: the LBB was incubated with 100 µl 1× Turbo DNase buffer and 3 µl Turbo DNase (Thermo Fisher Scientific, AM2239) at 37 °C for 30 min and RNA was eluted by 100 µl elution buffer. The RNA was purified again using the RNA Clean & Concentrator-5 kit (Zymo Research, R1016) and the RNA concentration was measured using the Qubit RNA HS Assay kit (Thermo Fisher Scientific, Q32852). Next, 700 ng of RNA was used to prepare a cDNA library according to the Oxford Nanopore SQK-DCS109 protocol. All of the libraries were sequenced in R9.4 flow cells on the GridION instrument according to the manufacturer's instructions.

Nanopore genome-seq library preparation

For genome-seq, DNA from F_2 embryos was extracted using the Quick-gDNA MicroPrep Kit (Zymo Research, D3021). Then, 400–700 ng DNA was used to prepare the library with either the Ligation Sequencing Kit (Fig. 4; Oxford Nanopore, SQK-LSK109) or the Rapid Sequencing Kit (Fig. 1; Oxford Nanopore, SQK-RAD004). All of the libraries were sequenced in R9.4 flow cells on the GridION instrument according to the manufacturer's instructions.

Divergent PCR

Total DNA (100 ng in 10 µl volume) was mixed with 1 µl 10× Plasmid-Safe DNase buffer, 0.5 µl Plasmid-Safe DNase and 0.5 µl 100 mM ATP. The mixture was incubated at 37 °C for 16 h on thermocycler, followed by at 70 °C for 30 min. A total of 1 µl of the digested DNA was used for divergent PCR using CloneAMP HiFi PCR Premix (Takara Bio, 639398), Gotaq Green Master Mix (Promega, M7123) or 2× Phanta Max Master

Mix (Vazyme, P515). A list of the primer sequences is provided in Supplementary Table 4.

Cell culture and IAP plasmid transfection

HT-29 cells were cultured in RPMI 1640 (Thermo Fisher Scientific, 11875093), 10% FBS (Cytvia, SH30396.03) and 1% penicillin–streptomycin (Thermo Fisher Scientific, 15140122) at 37 °C and 5% CO_2. HT-29 cells were seeded at 300,000 cells per well in six-well plates with complete medium and allowed to grow overnight. A total of 2 ml of culture medium was replaced the next day before transfection. IAP 440N1 WT plasmid was a gift from M. Dewannieux. In total, 5 μg of plasmids was delivered to cells using the Lipofectamine 3000 Transfection Reagent (Thermo Fisher Scientific, L3000001) according to the manufacturer's recommended protocol. Cells were incubated with transfection mixture for 24 h and then incubated with fresh medium. Total DNA was collected 48 h after transfection, and eccDNA was isolated according to procedure described above. To remove the remaining transfected plasmid, 100 ng total DNA (in 10 μl volume) was mixed with 0.5 μl DpnI and 1 μl of CutSmart Buffer (New England Biolabs, R0176S). The mixture was incubated at 37 °C for 30 min and 70 °C for 20 min on thermocycler. For transfection control amplification, DpnI digestion was excluded. Then, 1 μl 10× Plasmid-Safe DNase buffer, 0.5 μl Plasmid-Safe DNase and 0.5 μl 100 mM ATP were added to the mixture. The mixture was incubated at 37 °C for 2 h. An additional 0.3 μl 10× Plasmid-Safe DNase buffer, 0.5 μl Plasmid-Safe DNase and 0.5 μl 100 mM ATP were added and incubated for 16 h on the thermocycler, followed by 70 °C for 30 min. Divergent PCR products were analysed with 0.8% agarose gel and imaged using the Bio-Rad ChemiDoc XRS System (Bio-Rad, 1708265).

CRISPRi depletion of proteins in HT-29 cells

To construct the hUbC-dCas9-ZIM3-KRAB-hU6-sgRNA-PuroR (ZIM3-One) plasmid, dCas9-ZIM3-KRAB was cloned from pLX303-ZIM3-KRAB-dCas9 (Addgene, 154472) into a hU6-sgRNA-PuroR vector (a gift from K. Wood). sgRNAs targeting the promoter region were designed using an online tool (http://chopchop.cbu.uib.no/) and are listed in Supplementary Table 4. An additional guanine was appended to the sgRNAs that do not start with a guanine. Each sgRNA was cloned into the ZIM3-One plasmid and lentiviruses were produced. HT-29 cells were transduced with lentiviruses for 24 h and selected with 2 μg ml^{-1} puromycin for 10–14 days. The bulk cells were collected to evaluate the depletion efficiency by either western blotting or qPCR with reverse transcription (RT–qPCR).

Gene mutation by CRISPR–Cas9 in HEK293T cells

The sgRNAs were derived from the MinLib CRISPR guide RNA library and are listed in Supplementary Table 4. To ensure efficiency, an additional guanine was appended to the sgRNAs that do not start with a guanine. Each sgRNA was cloned into the pU6-(BbsI)-CBh-Cas9-T2A-BFP plasmid (Addgene, 64323) and verified by Sanger sequencing.

HEK293T cells were cultured in DMEM supplemented with GlutaMAX (Thermo Fisher Scientific, 10569-010), 10% FBS (Cytvia SH30396.03) and 1% penicillin–streptomycin (Thermo Fisher Scientific, 15140122) at 37 °C with 5% CO_2. Cells were seeded at 300,000 cells per well in six-well plates with complete medium and allowed to grow overnight. The next day, 2 ml of culture medium was replaced before transfection. Transfection was carried out by delivering 3 μg of plasmid to cells using the Lipofectamine 3000 Transfection Reagent (Thermo Fisher Scientific, L3000001) according to the manufacturer's recommended protocol. Cells were incubated with transfection mixture for 24 h and then incubated with fresh medium for an additional 48 h. Cells were then sorted for BFP signal using the Beckman Coulter Astrios EQ High-Speed Cell Sorter. The collected cells were plated into 6 cm dishes and allowed to recover for 48 h after sorting. Next, the cells were dissociated and diluted to 30 cells per ml. In total, 100 μl of the cell suspension was distributed into 96-well plates per well. Single clones were allowed to grow into stable colonies over a period of approximately 10–14 days. Finally, the colonies were validated for mutation by Sanger sequencing analysis of the site of the sgRNA-directed mutation.

Western blot analysis

Cells were lysed in RIPA buffer (Thermo Fisher Scientific, 89900) with 1× complete protease inhibitor cocktail (Roche, 4693159001). The lysate was resolved by SDS–PAGE gels and analysed by immunoblotting with the indicated primary antibodies. The following primary antibodies were used: anti-DNA ligase 3 (Proteintech, 26583-1-AP; 1:1,000), anti-XRCC1 (Proteintech, 21468-1-AP; 1:1,000), and anti-β-actin (Proteintech, 66009-1-Ig; 1:10,000). Secondary antibodies include: anti-mouse and anti-rabbit IgG-HRP (Thermo Fisher Scientific, G-21040 and G-21234; 1:5,000). The membrane was developed by SuperSignal West Pico PLUS Chemiluminescent Substrate Kit (Thermo Fisher Scientific, 34577) according to the manufacturer's instructions.

RNA purification and RT–qPCR

Total RNA of fly embryos or cells was extracted using the mirVana miRNA Isolation Kit (Thermo Fisher Scientific, AM1560). Total RNA (10 μg) was treated with 2 μl Turbo DNase (Thermo Fisher Scientific, AM2238) at 37 °C for 30 min. After DNase treatment, the RNA was purified using the RNA Clean & Concentrator-5 kit (Zymo Research, R1016). The cDNA was synthesized using the iScript cDNA Synthesis Kit (Bio-Rad, 1708890). RT–qPCR was performed with two technical replicates using SsoFast EvaGreen (Bio-Rad, 1725204) on the CFX96 Real-time System (Bio-Rad). Fold changes in mRNA were calculated using the $\Delta\Delta C_t$ method. *Rp49* was used as the internal control for quantifying fly gene expression. *RR18S* was used as the internal control for quantifying human gene expression. The primer sequences for qPCR of each gene were listed in Supplementary Table 4. *P* values were calculated from at least three independent biological replicates using two-tailed two-sample unequal-variance *t*-tests (Excel, Microsoft).

ssDNA immunoprecipitation and qPCR

Total DNA (2,000 ng) in a 50 μl volume was mixed with 5 μl 10× CutSmart buffer and 1 μl XhoI (New England Biolabs, R0146S). The mixture was incubated at 37 °C for 4 h and heat-inactivated at 65 °C for 20 min. In total, 5 μl of the mixture was set aside as the input DNA and the remaining mixture was split into 1 ml PBST (10 mM NaH_2PO_4, 175 mM NaCl, pH 7.4, 0.1% Triton X-100) and incubated overnight at 4 °C with 2 μg anti-ssDNA antibody (Sigma-Aldrich, MAB3034, 16-19) or 2 μg normal mouse IgG (Sigma-Aldrich, 12-371). The DNA–antibody complex was captured using 20 μl Dynabeads Protein G (Thermo Fisher Scientific, 10004D) for 2 h at 4 °C. The beads were sequentially washed three times with PBST. The DNA was eluted with 80 μl elution buffer (10 mM Tris-HCl, 300 mM NaCl, 5 mM EDTA, 0.5% SDS, pH 8.0) containing 5 μg proteinase K (Zymo Research, D3001-2-20) at 55 °C for 1 h with vigorous vortexing. The eluted DNA was then purified by phenol–chloroform (Thermo Fisher Scientific, 15593-049) and qPCR was performed using SsoFast EvaGreen (Bio-Rad, 1725204) on the CFX96 Real-time System. Fold changes were calculated using the $\Delta\Delta C_t$ method, with input DNA used for normalization with the shWhite and shAub control set as 1.

Nuclease P1 treatment and qPCR

Total DNA (40 ng in 10 μl volume) was mixed with 1 μl 10× NEBuffer r1.1, 0.5 μl XhoI (New England Biolabs, R0146S), and with or without 0.1 μl nuclease P1 (New England Biolabs, M0660S). The mixture was incubated at 37 °C for 4 h on a thermocycler, followed by 75 °C for 20 min. A total of 1 μl of the digested DNA was used for qPCR, which was performed with two technical replicates using SsoFast EvaGreen (Bio-Rad, 1725204) on the CFX96 Real-time System (Bio-Rad). Fold changes in mRNA were calculated using the $\Delta\Delta C_t$ method. *Rp49* was used as the internal control for quantifying DNA. The primer sequences for qPCR of each gene were listed in Supplementary Table 4. *P* values were calculated from at least

three independent biological replicates using two-tailed two-sample unequal-variance *t*-tests (Excel, Microsoft).

Spike-in and qPCR

The total DNA from ovary was extracted by using the Quick-gDNA Micro-Prep Kit (Zymo Research, D3021). Total DNA (100 ng in 50 µl volume) was mixed with 5 ng plasmid (contains CopGFP), 2 µl Plasmid-Safe DNase (Lucigen, E3110K), 5 µl 10× Plasmid-Safe buffer, 2 µl 100 mM ATP (Thermo Fisher Scientific, R0441), and ultrapure water (Thermo Fisher Scientific, 10977023). The mixture was incubated at 37 °C for 16 h and 70 °C for 30 min on a thermocycler. Then, 45 µl AMPure XP beads (Beckman Coulter, A63881) was used to purify DNA and eluted with 10 µl water. Purified DNA (1 µl) was used to perform qPCR using SsoFast EvaGreen (Bio-Rad, 1725204) on the CFX96 Real-time System (Bio-Rad). CopGFP was used as the internal control for quantifying mitochondrial DNA. A list of the primer sequences for qPCR analysis of each gene is provided in Supplementary Table 4. *P* values were calculated from at least three independent biological replicates using two-tailed two-sample unequal-variance *t*-tests (Excel, Microsoft).

Nanopore sequencing read preprocessing and mapping

The fast5 files generated by the Nanopore GridION machine were used as input in MinKNOW v.21.05.25 (MinKNOW core v.4.3.12). Guppy v.5.0.16 is integrated into MinKNOW. The basic data preprocessing parameters are as follows: basecall model = high-accuracy base-calling; read filtering = 9; The passed fastq files produced by MinKNOW were used for further quality control. Adapter sequences were detected and trimmed using porechop (v.0.2.4) with the parameters: --extra_end_trim 0 --discard_middle. This setting removes only the adapter sequencing detected at the beginning and the end of the reads; if the adapter sequence is detected in the middle of the reads, the reads were filtered out. Output files of the porechop were used for further analysis. Reads were mapped to the reference genome of *Drosophila melanogaster* version dm6 (GCA_000001215.4) and the transposon consensus sequences. The transposon consensus reference used in this study is available at GitHub (https://github.com/ZhaoZhangZZlab/eccDNA_formation_2021/tree/main/Reference/). This transposon dataset contains 121 transposons, classified at the subfamily level. Read mapping was performed using minimap2 (v.2.17-r941)[41] with the parameter settings -ax map-ont -Y -t 16 to retain the soft clipping sequences for all supplementary alignments in the SAM output. Mapped results were converted to the .bam format, sorted by reference coordinates, and indexed using samtools (v.1.12)[42]. Data visualization was performed using R (v.4.1.2) and Python (v.3.9.12). IGV (v.2.12.0)[43] was used to visualize the mapping results.

Read identification from the engineered and the endogenous *HMS-Beagle*

Sequencing reads were first mapped to *HMS-Beagle* consensus sequences. Mapped reads were further mapped to the GFP sequences, which can be found at GitHub (https://github.com/ZhaoZhangZZlab/eccDNA_formation_2021/tree/main/Reference/). Reads mapped to both *HMS-Beagle* consensus and the GFP sequences were considered from engineered *HMS-Beagle* products. Reads mapped only to *HMS-Beagle* consensus but not to the GFP sequence were considered from the endogenous *HMS-Beagle*.

eccDNA read selection from the genomic libraries prepared by Tn5 method

To detect transposon-derived eccDNA, we first mapped the reads to the *HMS-Beagle* consensus sequence and the dm6 genome. As *HMS-Beagle* from the linear DNA is typically encompassed by genomic sequences, reads carrying the *HMS-Beagle* sequence flanked by genomic sequences were filtered out. Only reads entirely mapped to *HMS-Beagle* were considered as candidates for eccDNA construction and classification.

eccDNA identification and classification

For both the reads selected from the genomic-seq or eccDNA-seq, chimeric alignments are suggestive of structural variation in genomic DNA sequencing. Reads with supplementary alignments were therefore used to identify the junction–junction site of the circular DNA. In general, the two alignments from the same read were compared as a pair. The reads were identified as transposon-derived circular DNA reads if the pair of alignments met the following conditions: (1) they mapped to the same strand; (2) the spliced sites of the two alignments were adjacent to each other, overlapping with each other, or closer than 100 bp; (3) the two alignments were in convergent configuration. To further classify the transposon-derived eccDNA on the basis of their structures, the following strategies were applied: (1) the reads were classified as 1-LTR full-length eccDNA if both mapping sites on the transposons were at the start and end of the transposon sequences, and the supplementary alignments overlapped with each other by the length of the LTR; (2) the reads were classified as 2-LTR full-length eccDNA if both mapping sites on the transposon were at the start and end of the transposon sequence, and the supplementary alignments did not overlap with each other. The reads were classified as 1-LTR rearranged eccDNA if only one mapping site was at either the start or the end of the transposon, and the other mapping site was in the middle of the transposon. The reads were classified as non-LTR rearranged eccDNA if both mapping sites on the transposon were in the middle of the transposon. Notably, by following this criterion, reads that had only partial LTR sequences were classified as non-LTR fragments. Considering the higher sequencing error rate of the Nanopore technology, we allowed 100 bp flexibility when setting the coordinates cut-offs. The abundance of each eccDNA type is represented by the number of reads that are classified into each type. In the figures, the circos (circlize v.0.4.14)[44] plot density indicates the mapping coverage of each eccDNA type. The coverage was generated using the bedtools (v2.29.2)[45] suite, with the parameters bedtools genomecov -bga.

Integration events detection

To identify the integration events, reads were first mapped to the transposon consensus. Candidate reads supporting integrations were selected if they met the following criteria: (1) reads mapped to transposon by at least 500 bp; (2) reads mapped to either the start or the end of the transposons. Next, the selected reads were mapped to the dm6 reference genome to determine the junction sites. Reads carrying transposon–genome chimeric structures that are not present in the reference genome were considered to be potential integration events. Reads aligned to a single transposon but multiple genomic regions were probably due to the repetitive nature of the landing sites. For these reads, the genomic landing locations were assigned on the basis of the best mapping results. Reads with only one transposon–genome chimeric configuration were probably from the insertions with the sequences only spanning partial of the transposons. Reads with multiple transposon–genome chimeric configurations were probably from insertions with the sequences spanning the entire transposon and the flanking sequences from both sides. These reads were minorities in the population because it requires reads long enough to cover the full-length transposons. These reads were further examined by their flanking sequence features. If the flanking genomic sequences were from the same chromosome, and the breakpoints were adjacent to each other, the reads were selected as insertions. Otherwise, the reads were excluded. Moreover, reads were filtered out if they carry structures that two transposons join from the ends because these reads are unlikely generated from the insertion events. To characterize the transposon insertion loci, the insertion events were clustered on the basis of the genome coordinates. The events that had breakpoints closer than 100 bp were grouped into one cluster. Any insertions shared between the oocytes and their corresponding somatic carcasses or shared by two or more datasets were removed because these probably resulted from

the polymorphisms between the fly genomes used in this study and the reference dm6 genome. Final integration events were represented in a bed file with the information including transposon name, insertion location and the number of reads that support the insertion cluster.

Normalization

Genomic sequencing data were normalized to the genome coverage. The size factors were calculated by the total number of bases in the library divided by the effective genome size of dm6 (142,573,017 bp). eccDNA-seq data were normalized by the coverage of the mitochondrial genome. The size factors were calculated by the total number of reads mapped to the mitochondrial genome divided by the size of the dm6 chrM (19,524 bp). The number of eccDNA was normalized using the equation:

$$C = \frac{\sum_{n=1}^{N} \frac{c_n}{s_n}}{N} \times \bar{S},$$

where N is the number of replicates, c_n and s_n are the raw eccDNA and the size factor in each replicate, respectively, and $\bar{S}$ is the average sequencing depth of all datasets. Silencing *DNAlig3* probably affects the mitochondria numbers within the ovary. Thus, the eccDNA data from shDNAlig3 flies were further normalized to the number of reads from the inactive transposons.

Statistics and reproducibility

Statistical tests were performed using GraphPad Prism (v.8) and Microsoft Excel (v.16.67). The significance test comparing different groups was determined using two-tailed two-sample unequal-variance *t*-tests. The experiments in Fig. 2b–d and Extended Data Figs. 2c and 8c were repeated at least three times independently with similar results. Similarly, the experiments in Fig. 5b–d and Extended Data Figs. 8d–f and 9a,b were independently repeated twice, yielding consistent results. Biological replicates were used for all of the independent experiments. The micrographs shown in Extended Data Figs. 2b and 4a are representative of two independently conducted experiments, with similar results obtained.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The sequencing data were deposited to the National Center for Biotechnology Information (NCBI) under accession number PRJNA794176. The sequence of the eGFP-tagged *HMS-Beagle* is available at GitHub (https://github.com/ZhaoZhangZZlab/eccDNA_formation_2021/tree/main/Reference). Source data are provided with this paper.

Code availability

All related code is available at GitHub (https://github.com/ZhaoZhangZZlab/eccDNA_formation_2021).

41. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
42. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
43. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
44. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparinggenomic features. *Bioinformatics* **26**, 841–842 (2010).
45. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

Acknowledgements We thank M. Dewannieux and K. Wood for providing plasmids; J. Brennecke, X. Chen, J. Sekelsky and the members of the BDSC for providing fly stocks; the members of the Z.Z.Z. laboratory and D. MacAlpine for suggestions; D. Fox, X.-F. Wang, B. Cullen, L. Lin and D. MacAlpine for reading the manuscript; and D. Erie and P. Marszalek for suggestions on AFM sample preparation. This work was supported by grants to Z.Z.Z. from the Pew Biomedical Scholars Program and the National Institutes of Health (DP5 OD021355 and R01 GM141018); and to D.A.R. from the National Cancer Institute (P01CA247773).

Author contributions Z.Z.Z., F.Y. and W.S. conceived the project. All of the authors designed the experiments. O.W.C. performed experiments for Fig. 5d and Extended Data Fig. 9c,d. L.T. performed experiments for Extended Data Fig. 9e. L.W. generated data for Fig. 1. F.Y. performed the rest of the experiments. W.S. performed all of the bioinformatics. Z.Z.Z., F.Y. and W.S. wrote the manuscript. All of the authors read and approved the manuscript.

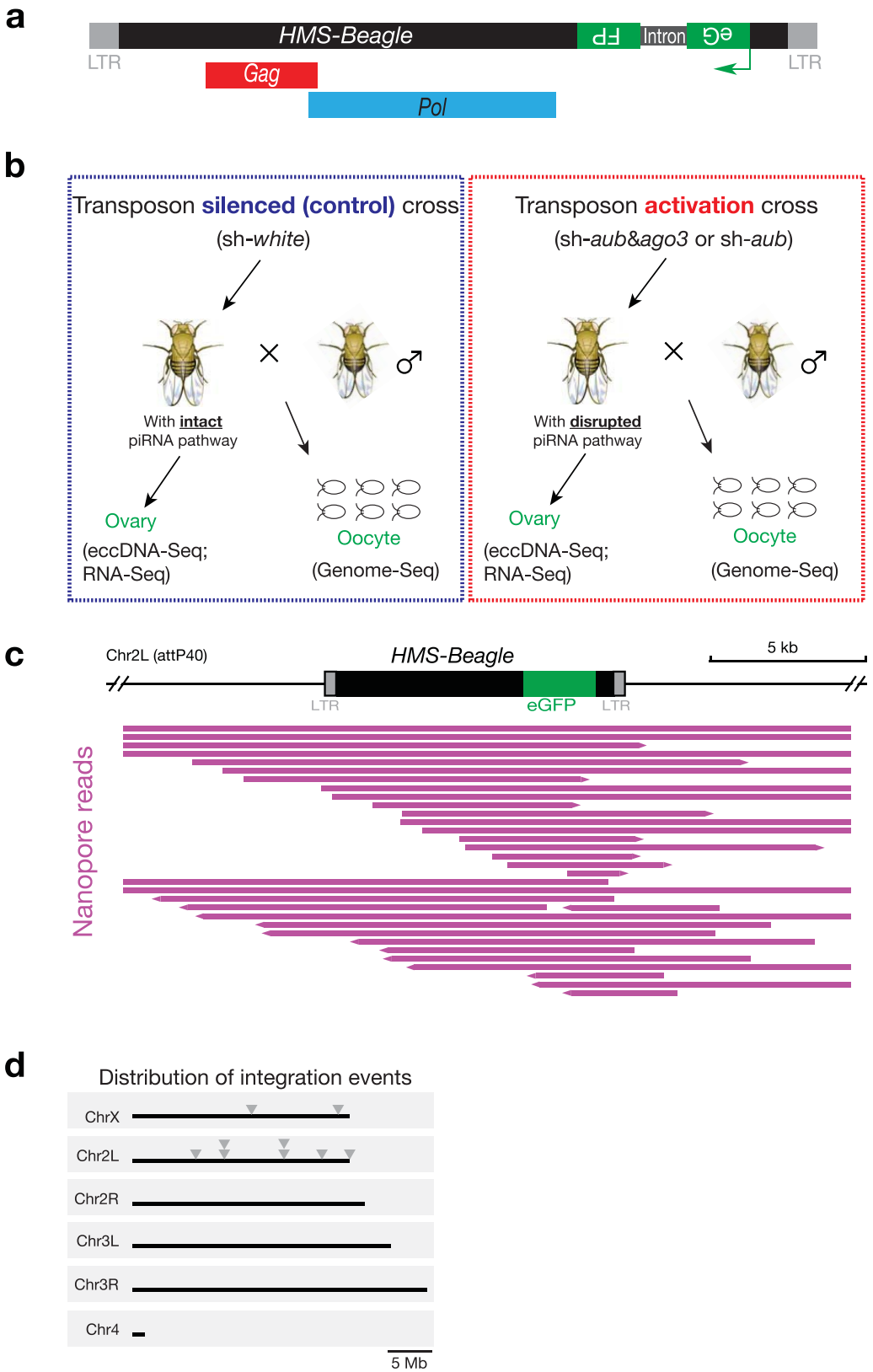
Competing interests Z.Z.Z., F.Y. and W.S. are listed co-inventors on a US provisional patent application (no. 63/309,136) filed by Duke University related to this work.

Additional information

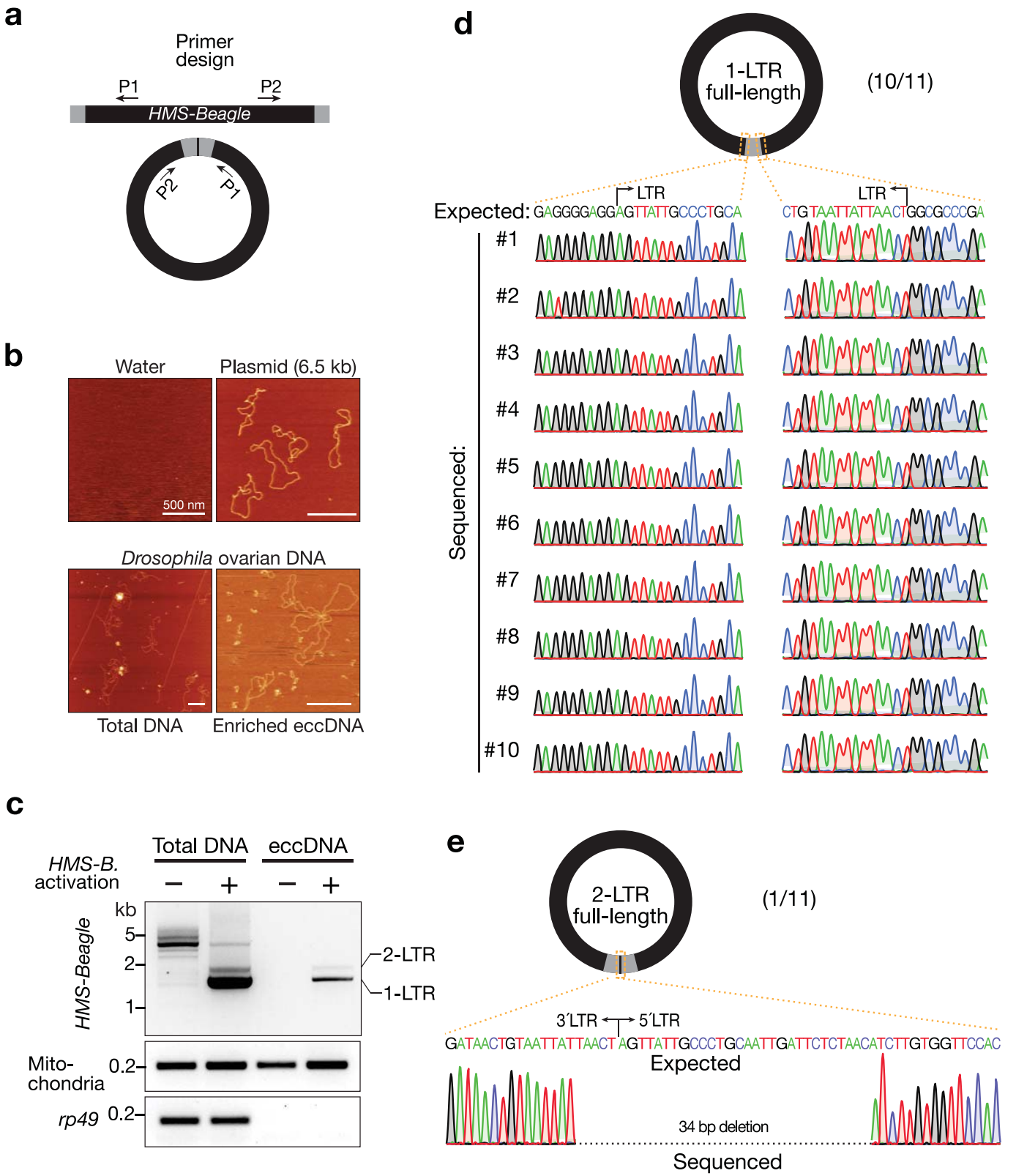
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06327-7.

Correspondence and requests for materials should be addressed to ZZ Zhao Zhang.

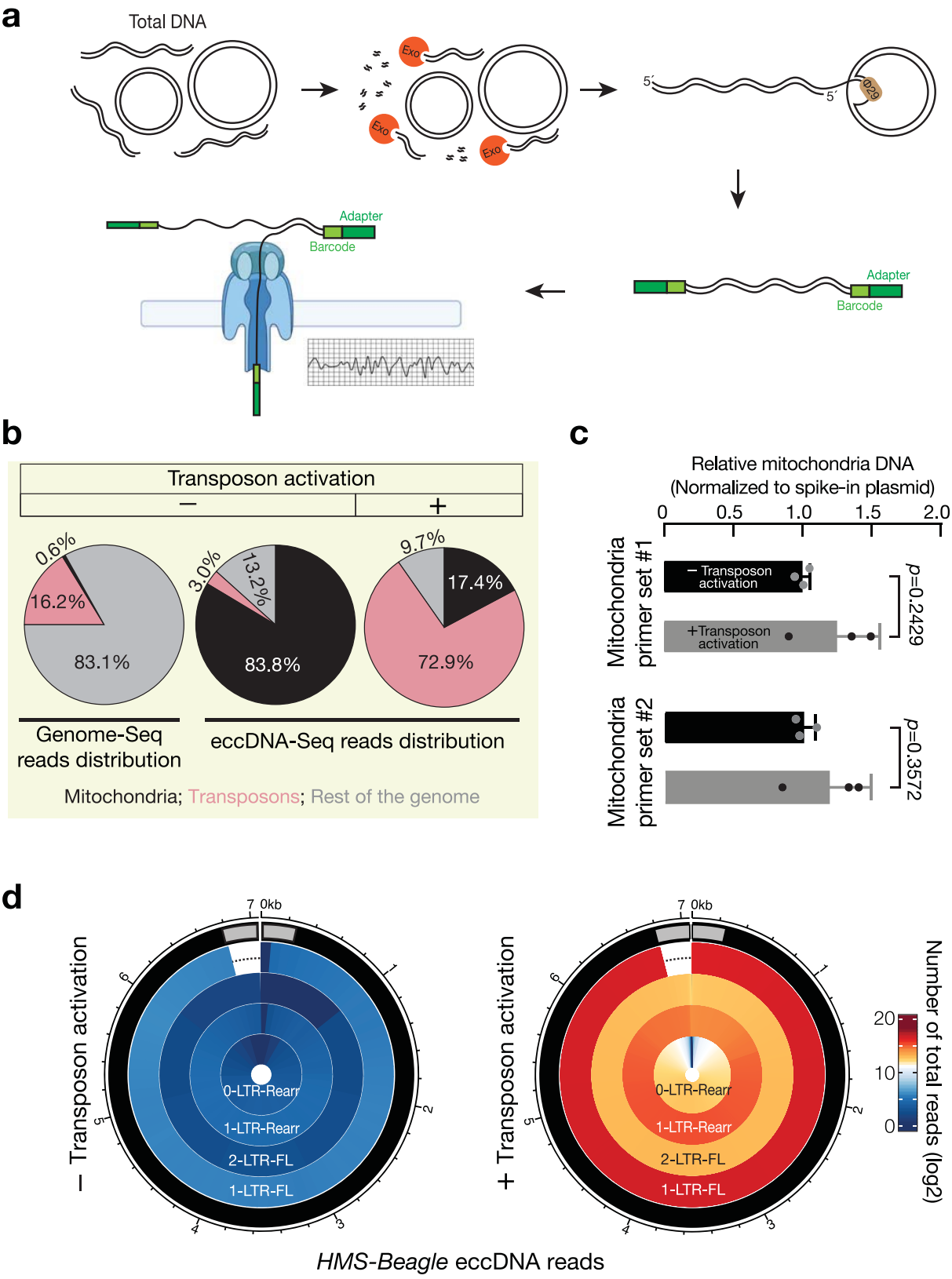
Peer review information *Nature* thanks Todd Macfarlan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | The engineered *HMS-Beagle* reporter dominantly forms eccDNA. a, Schematic design of the *HMS-Beagle* reporter. An eGFP reporter is inserted into the 3′ UTR of *HMS-Beagle* sequence in an antisense direction. **b**, Fly cross scheme to collect samples for measuring the potential integration and eccDNA events from transposon-silenced and transposon-activated flies. **c**, Integrative Genomics Viewer (IGV) alignments showing reads mapped to *HMS-Beagle* reporter locus in the genome from embryos laid by transposon-silenced females. Individual purple horizonal bar represents a unique Nanopore read containing eGFP sequence. All of the reads contained at least one of the LTRs and extended to the adjacent region, indicating they were aligned to the original genomic locus of the reporter. **d**, The distribution of new integrations from engineered *HMS-Beagle* reporter on *Drosophila* genome. Each triangle represents a new integration event.

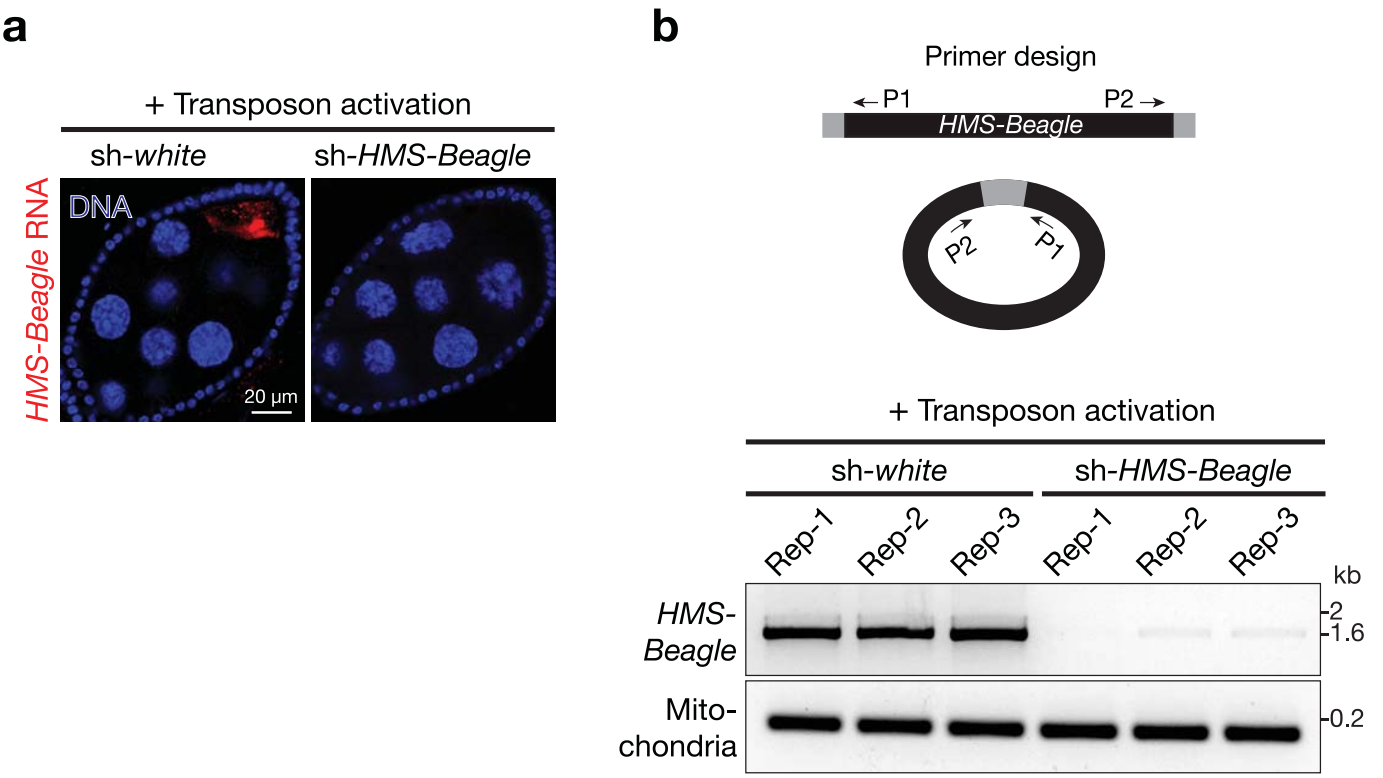


Extended Data Fig. 2 | PCR based assay to measure *HMS-Beagle* eccDNA. **a**, schematic of the design of divergent primers to identify retrotransposon eccDNA. **b**, AFM imaging to visualize the shapes of DNA. Exonuclease digestion significantly enrich eccDNA for detection in panel **c**. Scale bar, 500 nm. **c**, the representative gel image showing retrotransposons predominantly form 1-LTR circles. Performing PCR using total DNA as template produced non-specific bands, likely resulting from the nested transposon fragments resided within the linear genome. Using exonuclease to enrich eccDNA generated two PCR products corresponding to 1-LTR and 2-LTR eccDNA respectively. **d** and **e**, Sanger sequencing to validate the formation of *HMS-Beagle* eccDNA. The PCR products for the very right lane of panel c were cloned into plasmid vector and 11 corresponding colonies were sequenced. Ten of the 11 colonies are from 1-LTR eccDNA (**d**). One colony is from 2-LTR eccDNA (**e**). Notably, this 2-LTR eccDNA has 34 bp deletion at the end-end junction site, indicating it is formed by the error-prone NHEJ pathway. This conclusion is further supported by Extended Data Fig. 9.

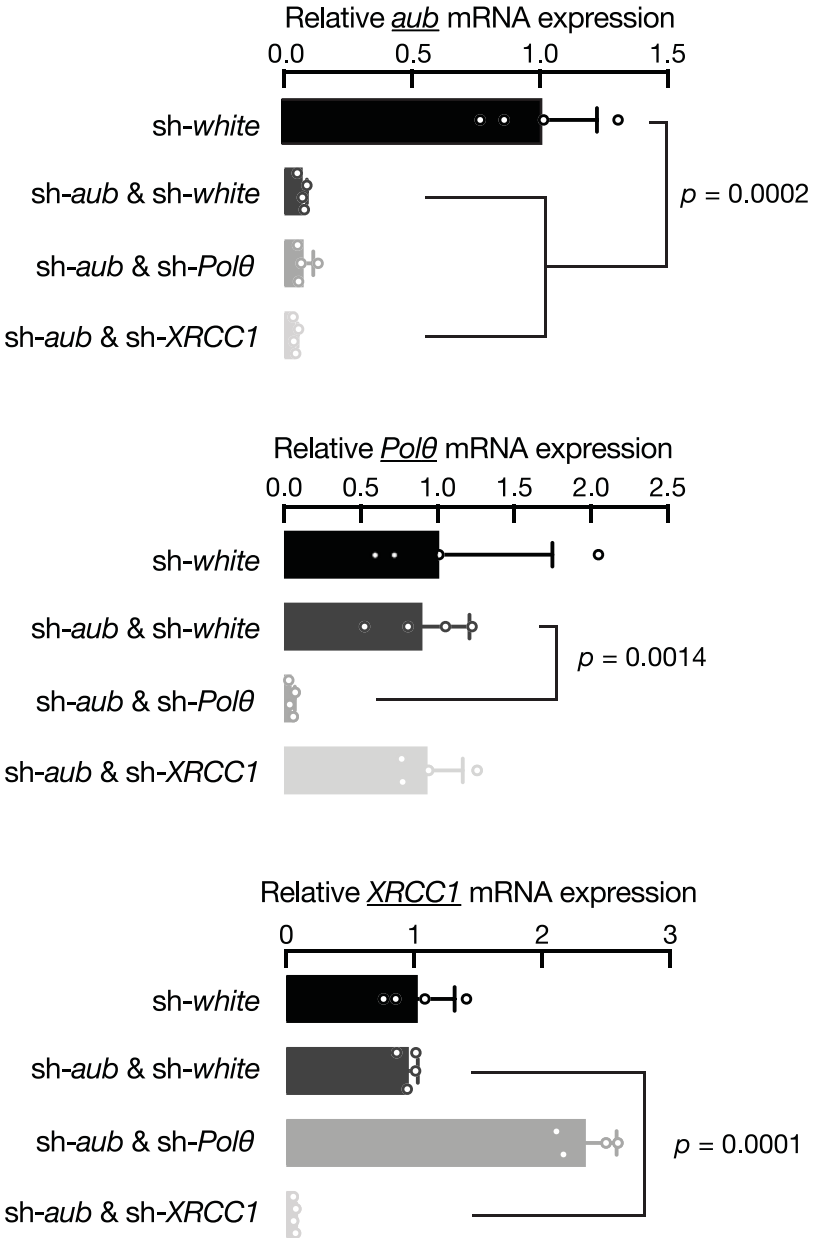


Extended Data Fig. 3 | See next page for caption.

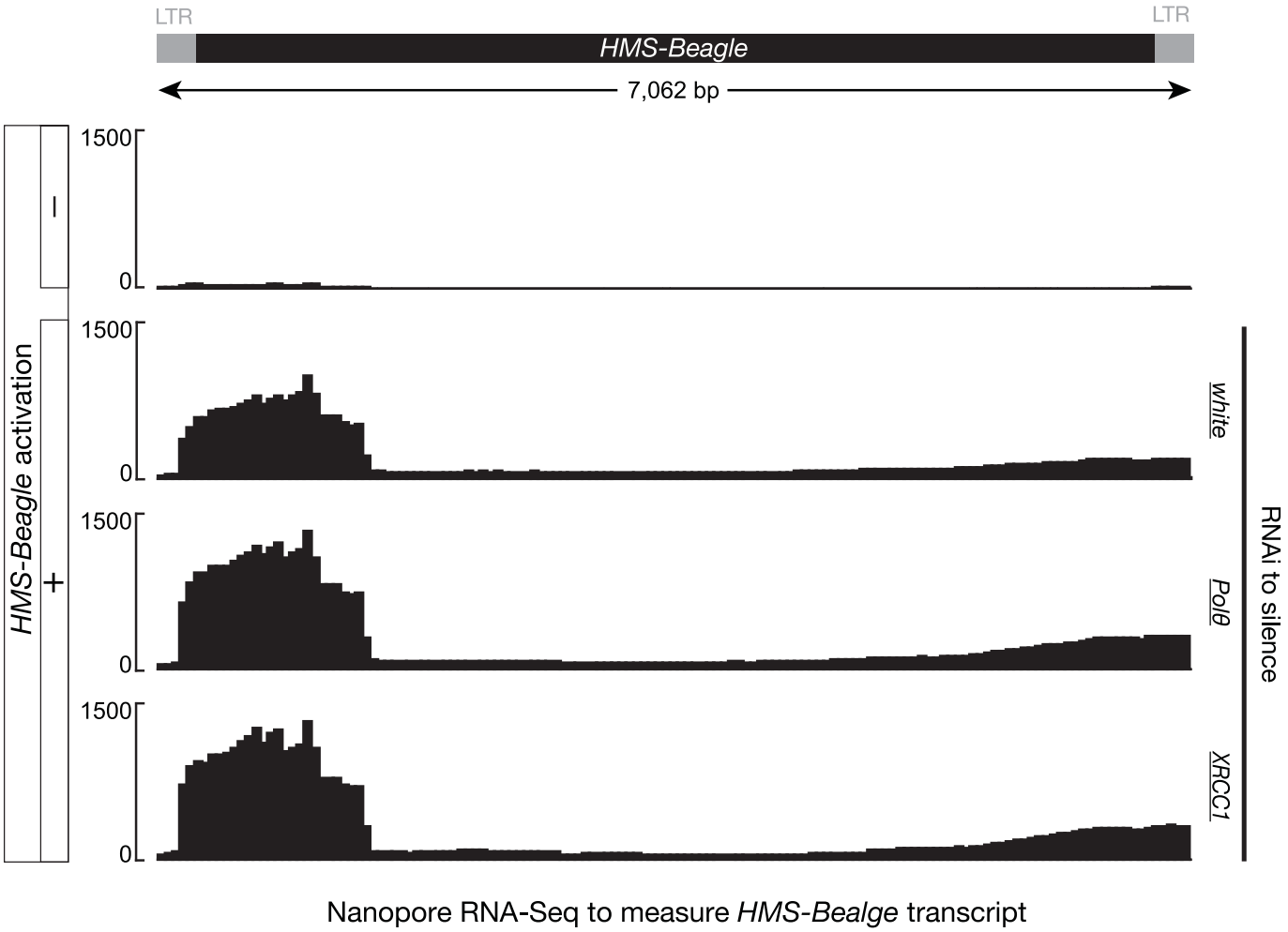
Extended Data Fig. 3 | eccDNA-seq to provide direct evidence of circle formation. a, Schematic of the eccDNA-seq workflow. After extracting total DNA, linear DNA was removed by Plasmid-Safe DNase digestion. eccDNA was amplified by Phi29 DNA polymerase through rolling circle amplification. And the sequencing libraries were prepared and sequenced on a Nanopore instrument. **b**, The proportion of eccDNA-seq and genome-seq reads mapped to mitochondria (black regions), transposons (pink regions), and the rest of the genome (grey regions). All samples were from fly ovaries. The genome-seq libraries were made by the tagmentation method and sequenced by the Nanopore platform to capture circular DNA, such as the mitochondrial genome. **c**, Bar graph showing qPCR results of mitochondrial DNA copies detected by two sets of primers respectively. The relative abundances are normalized to the spike-in plasmid. The mitochondrial DNA copies are essentially unchanged upon transposon activation in *Drosophila* ovary. The bars report mean ± standard deviation from three biological replicates (n = 3). *p* values were calculated with a two-tailed, two-sample unequal variance *t* test. **d**, Circos plots showing the number of the eccDNA-seq reads for the four classes of *HMS-Beagle* circles. From the outer layer to inward: 1-LTR full-length circles, 2-LTR full-length circles, 1-LTR-rearranged circle, and non-LTR rearranged circles.



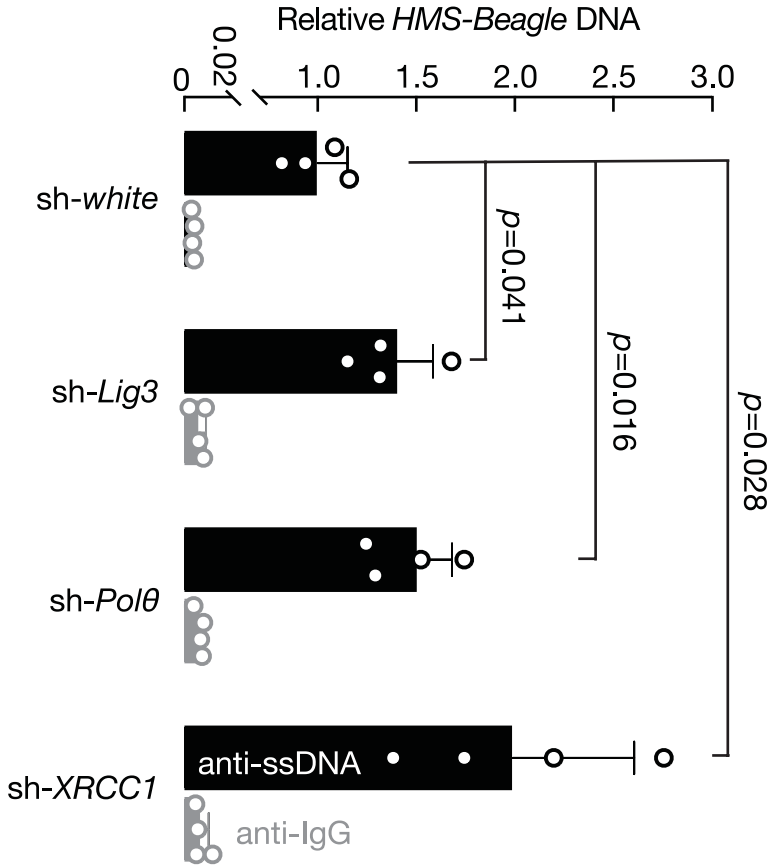
Extended Data Fig. 4 | eccDNA production from *HMS-Beagle* requires its mRNA intermediates. a, RNA-FISH to detect *HMS-Beagle* mRNA. All flies carrying sh-*aub* to activate transposons in germline cells. Further introducing sh-*white* (serving as a control) into the animals does not change transposon activity: *HMS-Beagle* remains activated. Upon introducing sh*HMS-Beagle* construct to silence it, its RNA was undetectable by RNA-FISH. Scale bar, 20 μm, **b**, Top: primer design to detect *HMS-Beagle* eccDNA (Extended Data Fig. 2). Bottom: The representative gel image of PCR products showing that *HMS-Beagle* eccDNA production was abolished when its mRNA production was suppressed by RNAi. Each genotype has three biological replicates.



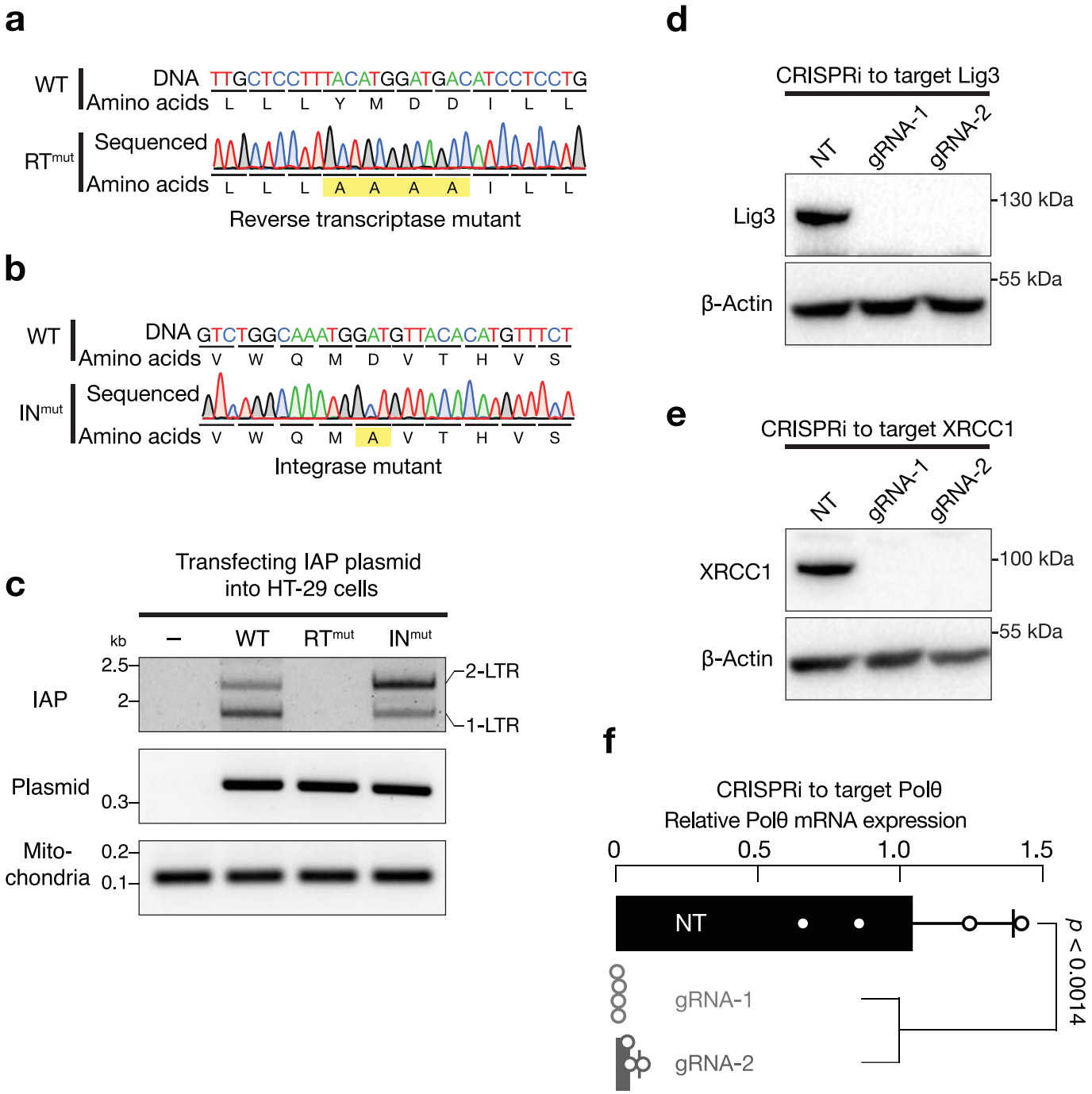
Extended Data Fig. 5 | Confirmation of the RNAi silencing efficiency in oocytes. RT-qPCR showing the depletion efficiency of indicated genes by germline-specific RNAi. Relative mRNA levels were normalized to *rp49* gene. The bars report mean ± standard deviation from four biological replicates (n = 4). *p* values were calculated with a two-tailed, two-sample unequal variance *t* test. Silencing Lig3 or Fen1 made flies barely lay eggs/oocytes, impeding a validation of the RNAi silencing efficiency for them.



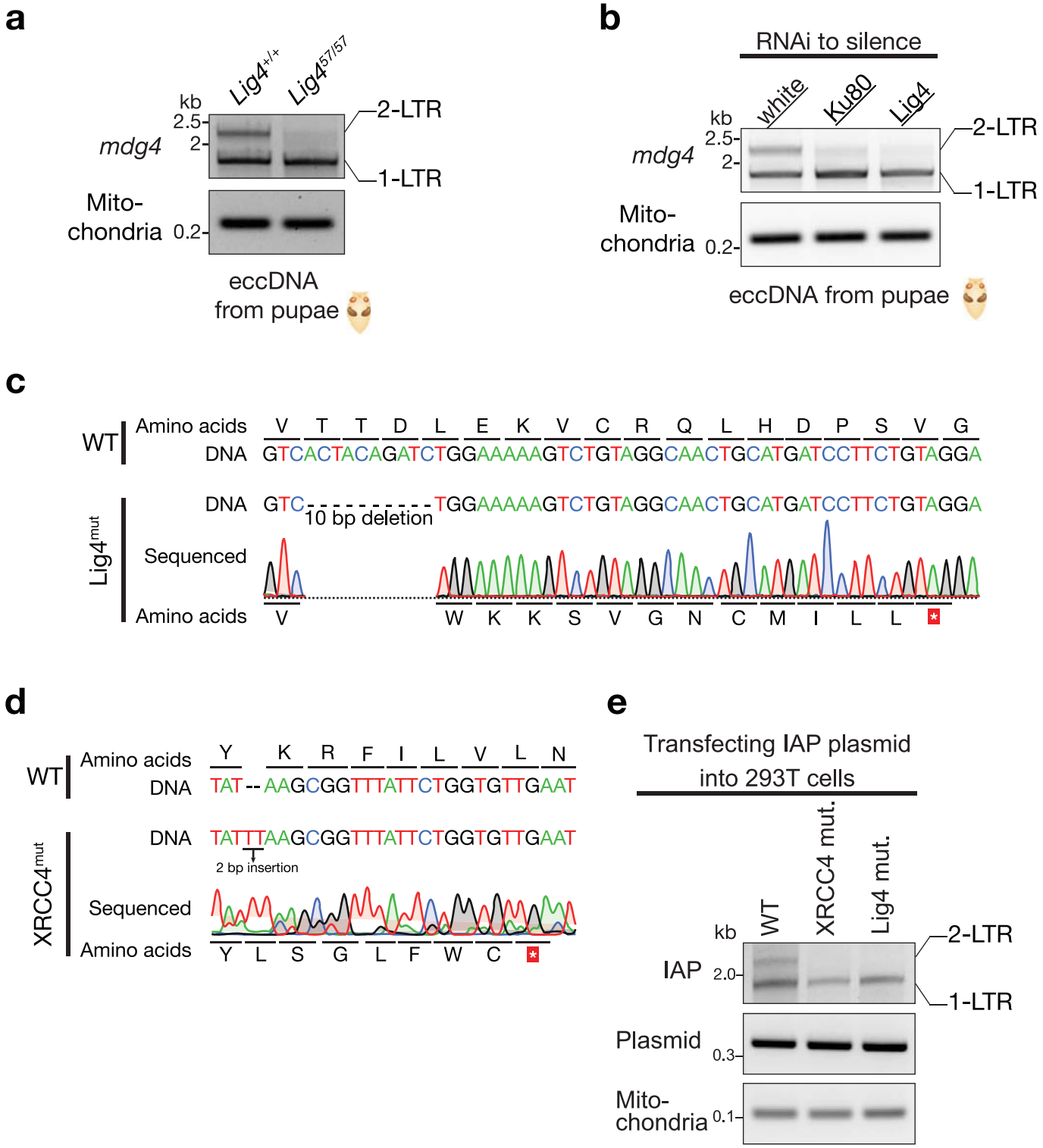
Extended Data Fig. 6 | *HMS-Beagle* mRNA remains unchanged upon depletion of the components from alt-EJ process. Transposon activation was achieved by silencing Aub in germline cells. The Y-axis is normalized reads count.



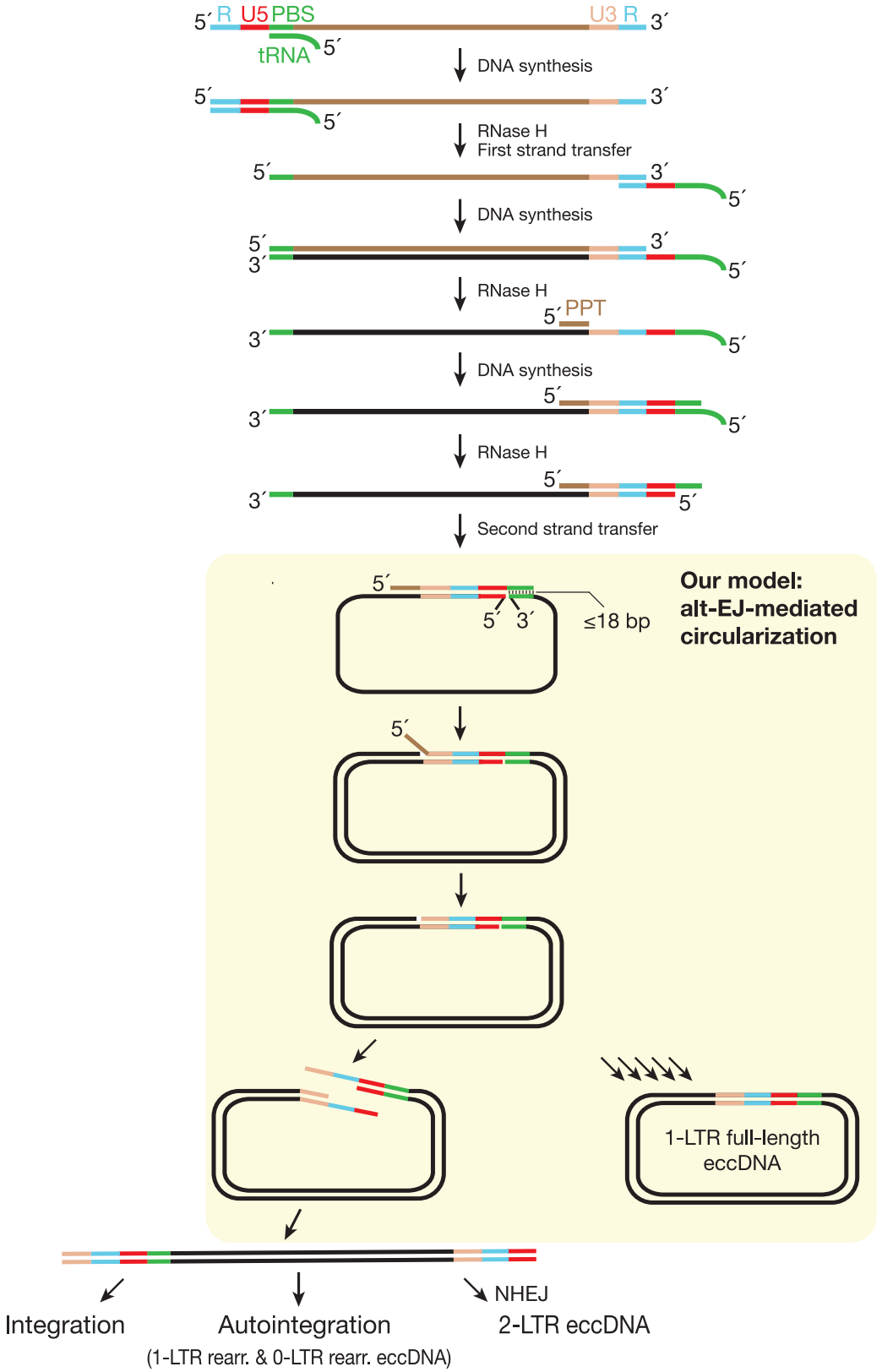
Extended Data Fig. 7 | Immunoprecipitation assay to measure the accumulation of *HMS-Beagle* single-stranded DNA upon alt-EJ suppression. The bars report mean ± standard deviation from four biological replicates (n = 4). *P* values were calculated with a two-tailed, two-sample unequal variance *t* test. Although Mab3034 antibodies used in this experiment have 10-fold higher affinity for single-stranded DNA than double-stranded DNA[31], they still can bind *HMS-Beagle* genomic double-stranded DNA across all samples. This would mask the difference of single-stranded DNA across samples and lead to underestimation of the amount of accumulated single-stranded DNA upon alt-EJ inhibition.



Extended Data Fig. 8 | IAP needs its reverse transcriptase, but not integrase, activity for eccDNA biogenesis. a, Sanger sequencing to validate the IAP reverse transcriptase mutant. **b**, Sanger sequencing to validate the IAP integrase mutant. **c**, PCR based assay to measure the production of IAP eccDNA. The very left lane was the condition without introducing IAP plasmid. **d**–**e**, Either immunoblotting (**d** and **e**) or RT-qPCR (**f**) to test the silencing efficiency of CRISPRi on depleting the alt-EJ factors. For each gene, two gRNAs were designed. NT (non-targeting) is a random gRNA without a targeting site. For RT-qPCR, relative mRNA levels were normalized to the RR18S gene. The bars report mean ± standard deviation from four biological replicates (n = 4). Statistical significance were calculated with a two-tailed, two-sample unequal variance *t* test. The *p* value for gRNA-1 is 0.0011, while for gRNA-2 is 0.0014.



Extended Data Fig. 9 | NHEJ pathway is essential for 2-LTR eccDNA biogenesis. a, Mutating *Lig4* abolishes 2-LTR eccDNA production for *mdg4* retrotransposon. **b**, Silencing *Ku80* or *Lig4* by RNAi reduces *mdg4* 2-LTR eccDNA formation. **c**, Sanger sequencing to validate lig4 mutation of the 293T cells. **d**, Sanger sequencing to validate XRCC4 mutation of the 293T cells. **e**, Mutating either Lig4 or XRCC4 abolishes 2-LTR eccDNA production for IAP retrotransposon.



Extended Data Fig. 10 | Detailed model of the replication cycle of LTR-retrotransposons supported by our study. Our data support alt-EJ factors mediate a circularization step for retrotransposon 2-nd strand DNA synthesis. While this step can generate full-length linear double-stranded DNA for integration, it appears to dominantly produce 1-LTR eccDNA.

nature portfolio

Corresponding author(s): Zhao Zhang

Last updated by author(s): May 18, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Nanopore sequencing data was obtained by GridION machine, fast5 files were processed by MinKNOW version 21.05.25 (MinKNOW core 4.3.12). Guppy 5.0.16 is integrated into the MinKNOW. Confocal images were acquired by Leica SP5 microscope. qPCR was performed by CFX96 Real-time System (Bio-Rad). AFM images were acquired by Asylum Cypher Atomic Force Microscope and processed using Gwyddion 2.52.
Data analysis	Nanopore sequencing data were processed by porechop (0.2.4) (https://github.com/rrwick/Porechop). Reads mapping was performed using the minimap2 (2.17-r941) software (https://github.com/lh3/minimap2) with parameter settings -ax map-ont -Y -t 16 to keep the soft clipping sequences for all supplementary alignments in the SAM output. Mapped results were converted to the bam format, sorted by reference coordinates, and indexed by samtools (1.12) (http://www.htslib.org/doc/samtools.html). bedtools (v2.29.2) (https://bedtools.readthedocs.io/en/latest/) was used to process the alignment results. RNA sequences were analyzed by the software DEseq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html). Data visualization was achieved by R (4.1.2) (https://www.r-project.org/) and python (3.9.12) (https://www.python.org/). The software IGV (2.12.0) was used to visualize mapping results (https://software.broadinstitute.org/software/igv/). Code was deposited to https://github.com/ZhaoZhangZZlab/eccDNA_formation_2021. Statistical tests were calculated by GraphPad Prism (v8) and Microsoft Excel (v16.67).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The sequencing data were deposited to the National Center for Biotechnology Information (NCBI) with accession number PRJNA794176.
All related code is available at https://github.com/ZhaoZhangZZlab/eccDNA_formation_2021

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	not applicable.
Population characteristics	not applicable.
Recruitment	not applicable.
Ethics oversight	not applicable.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Statistical methods were not used to predetermine the sample size. Instead, the sample sizes were chosen based on previous experience and established practices within the field (doi: 10.1101/pdb.prot5198; DOI: https://doi.org/10.7554/eLife.66405). We chose about 10 flies for the divergent PCR (Fig. 2b-d, Fig. 5b,c, Extended Data Fig. 2c, Extended Data Fig. 4b, Extended Data Fig. 10a, b), qPCR (Fig. 3b, Extended Data Fig. 3c, Extended Data Fig. 5), and mRNA-seq (Extended Data Fig. 6). We chose a sample size of >30 flies for the eccDNA-seq of ovaries (Fig. 2e, Fig. 3c, d, Extended Data Fig. 3b, d), ssDNA immunoprecipitation (Extended Data Fig.7) and >500 embryos for genome-seq (Fig.1 and Fig. 4).
Data exclusions	No data were excluded from analysis.
Replication	All experiments were independently repeated as indicated in the respective figure legends. The Nanopore sequencing experiments were repeated twice with no problems in reproducibility and the representative sequencing results were shown.
Randomization	All the flies with indicated genotypes were randomly collected, thus, they were not allocated into specific subgroups.
Blinding	The investigators were not blinding to the flies' genotypes. Because no subjective analyses were performed in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

rabbit anti-DNA ligase 3 (#26583-1-AP, Proteintech, 1:1000)
rabbit anti-XRCC1 (#21468-1-AP, Proteintech, 1:1000)
mouse anti-β-Actin (#66009-1-Ig, Proteintech, 1:10000)
mouse anti-ssDNA (#MAB3034, Sigma, 2 µg per precipitation)
anti-mouse IgG-HRP (#G-21040, Thermo Scientific, 1:5000)
anti-rabbit IgG-HRP (#G-21234, Thermo Scientific, 1:5000)
normal mouse IgG (#12-371, Sigma, 2 µg per precipitation)

Validation

All antibodies were purchased from commercial suppliers and the corresponding validation studies can be found on their website:
rabbit anti-DNA ligase 3 (#26583-1-AP, Proteintech)
https://www.ptglab.com/products/LIG3-Antibody-26583-1-AP.htm
rabbit anti-XRCC1 (#21468-1-AP, Proteintech)
https://www.ptglab.com/products/XRCC1-Antibody-21468-1-AP.htm
mouse ant-β-Actin (#66009-1-Ig, Proteintech)
https://www.ptglab.com/products/Pan-Actin-Antibody-66009-1-Ig.htm
mouse anti-ssDNA (#MAB3034, Sigma)
https://www.sigmaaldrich.com/US/en/product/mm/mab3034?gclid=CjwKCAjw8-OhBhB5EiwADyoY1dfPC_c9lJPIz2Qf78spsXsETtE_PHRz-wWoYfy6ip84HzyIJo4schoC7lYQAvD_BwE&gclsrc=aw.ds

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

HT-29 cells were from Duke cell culture facility (#CL-3404-FV).
HEK293T cells were from Duke cell culture facility (#CL-3395-FV).

Authentication

Cell lines were not additionally authenticated.

Mycoplasma contamination

All cell lines were tested as negative of mycoplasma at the beginning of the study, but not routinely tested thereafter.

Commonly misidentified lines
(See ICLAC register)

None.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals

sh-aub and sh-ago3 flies are a gift from Dr. Julius Brennecke. vas-Gal4 is a gift from Dr. Xin Chen. Lig4[57] is a gift from Dr. Jeff Sekelsky. sh-white, Tub-Gal80[ts], and RNAi stains for the screen are obtained from the Bloomington Drosophila Stock Center. The HMS-Beagle-TR and sh-HMS-Beagle transgenic flies were generated by landing the plasmids into a fly carrying attP2 site. All the female flies used in this study were 3-7 days old and the embryos were collected within 6 hours after being laid.

Wild animals

No wild animals were used in this study.

Reporting on sex

All the experiments by using adult flies were carried out on 3-7 days old female flies. The 0-to-6 hours old fly embryos were randomly collected and are the mixture of both male and female.

Field-collected samples

No field collected samples were used in this study.

Ethics oversight

No ethical approval was required for this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.