# Transposable element and host silencing activity in gigantic genomes

**Jie Wang[1*], Liang Yuan[2], Jiaxing Tang[1,3], Jiongyu Liu[1], Cheng Sun[4], Michael W. Itgen[5], Guiying Chen[3], Stanley K. Sessions[6], Guangpu Zhang[1,3*], Rachel Lockridge Mueller[5*]**

[1]CAS Key Laboratory of Mountain Ecological Restoration and Bioresource Utilization & Ecological Restoration and Biodiversity Conservation Key Laboratory of Sichuan Province, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, Sichuan 610041, China

[2]School of Life Sciences, Xinjiang Normal University, Urumqi, China

[3]College of Life Sciences, Sichuan Normal University, Chengdu, China

[4]College of Life Sciences, Capital Normal University, Beijing, China

[5]Department of Biology, Colorado State University, Fort Collins, CO, USA

[6]Biology Department, Hartwick College, Oneonta, NY, USA

*** Correspondence:**
Jie Wang, wangjie@cib.ac.cn
Guangpu Zhang, zgp1122@outlook.com
Rachel Lockridge Mueller, rlm@colostate.edu

**Keywords: TE expression, piRNA pathway, genome size evolution, salamander, KRAB-ZFP, *Ranodon sibiricus***

## Abstract

Transposable elements (TEs) and the silencing machinery of their hosts are engaged in a germline arms-race dynamic that shapes TE accumulation and, therefore, genome size. In animal species with extremely large genomes (>10 Gb), TE accumulation has been pushed to the extreme, prompting the question of whether TE silencing also deviates from typical conditions. To address this question, we characterize TE silencing via two pathways — the piRNA pathway and KRAB-ZFP transcriptional repression — in the male and female gonads of *Ranodon sibiricus*, a salamander species with a ~21 Gb genome. We quantify 1) genomic TE diversity, 2) TE expression, and 3) small RNA expression and find a significant relationship between the expression of piRNAs and TEs they target for silencing in both sexes. We also quantified TE silencing pathway gene expression in *R. sibiricus* and 14 other vertebrates with genome sizes ranging from 1 – 130 Gb and find no association between pathway expression and genome size. Taken together, our results reveal that the gigantic *R. sibiricus* genome includes at least 19 putatively active TE superfamilies, all of which are targeted by the piRNA pathway in proportion to their expression levels, suggesting comprehensive piRNA-mediated silencing. Males have higher TE expression than females, suggesting that they may contribute more to the species' high genomic TE load. We posit that apparently conflicting interpretations of TE silencing and genomic gigantism in the literature, as well as the absence of a correlation between TE silencing pathway gene expression and genome size, can be reconciled by considering whether the TE community or the host is currently "on the attack" in the arms race dynamic.

## 1 Introduction

Transposable elements (TEs) are DNA sequences that can mobilize throughout the genomes of their hosts, typically replicating as part of the transposition life cycle (Doolittle and Sapienza, 1980;Orgel and Crick, 1980;Wicker et al., 2007). TEs are an ancient and diverse class of sequences, encompassing a range of replication mechanisms that rely on both TE- and host-encoded enzymatic machinery (Bourque et al., 2018). Eukaryotic genomes contain a substantial yet variable number of TEs; they make up well over half of the human genome, up to 85% of the maize genome (Haberer et al., 2005), yet only ~0.1% of the yeast *Pseudozyma antarctica* genome (de Koning et al., 2011;Castanera et al., 2017;Jiao et al., 2017). TE abundance is one of the major determinants of overall genome size, which ranges from ~0.002 Gb to ~150 Gb across eukaryotes and from ~0.4 Gb to ~130 Gb across vertebrates (Rodriguez and Arkhipova, 2018;Gregory, 2022). The mechanistic and evolutionary forces shaping TE abundance and, thus, genome size remain incompletely understood.

Individual TE insertions have a range of effects on host fitness; the majority are effectively neutral or slightly deleterious, while smaller proportions are either harmful (or lethal) on the one hand or adaptive on the other (Arkhipova, 2018;Almeida et al., 2022). For example, at least 120 human diseases have been attributed to the effects of de novo TE insertions, but so have classic adaptive traits including industrial melanism and the mammalian placenta (Hancks and Kazazian, 2016;Hof et al., 2016;Senft and Macfarlan, 2021). The likelihood of a novel TE insertion having an extreme effect on the host phenotype depends on properties of the host genome including gene density, which affects the probability of a new insertion disrupting a functional protein-coding or regulatory sequence (Medstrand et al., 2002).

In response to TEs' mutagenic properties, eukaryotes have evolved multiple mechanisms to silence their activity, particularly in the germline and early embryo where TE effects on host fitness are the most pronounced (Almeida et al., 2022). Some mechanisms act by transcriptionally silencing TE loci through targeted deposition of chromatin modifications (e.g. methylation of cytosines on DNA, or H3K9 methylation of histone proteins) (Deniz et al., 2019). Other mechanisms act post-transcriptionally, targeting TE transcripts for destruction in the cytoplasm before they can complete the replicative life cycle and generate a novel genomic TE insertion (Czech and Hannon, 2016).

In multicellular animals, TE silencing in the germline and during early embryogenesis is carried out by the piRNA pathway, a small RNA pathway that relies on RNA-induced silencing complexes (RISC) composed of PIWI proteins and associated guide piRNAs that identify TE transcripts by base complementarity (Aravin et al., 2006;Ozata et al., 2019;Iwakawa and Tomari, 2022). In the nucleus, piRNA-PIWI complexes identify chromatin-associated nascent TE transcripts, inducing transcriptional silencing of the genomic TE locus through recruitment of DNA methyltransferases and histone methyltransferases to establish a repressive chromatin structure (Aravin et al., 2008;Czech et al., 2018). In the cytoplasm, piRNA-PIWI complexes identify mature TE transcripts and cleave them between nucleotide positions 10 and 11 of the guide piRNA (Reuter et al., 2011;Iwasaki et al., 2015). The cleaved fragments of TE mRNA induce the production of more TE-targeting piRNAs through a feed-forward loop called the ping-pong cycle, which amplifies the cell's post-transcriptional TE silencing response (Brennecke et al., 2007;Gunawardane et al., 2007;Castel and Martienssen, 2013). Although present in both males and females, there are sex-specific differences in activity of the piRNA pathway, which may be associated with sex-biased contributions to overall genomic TE load (Saint-Leandre et al., 2020).

In lobe-finned fishes (tetrapods, coelacanth, and lungfishes), TE silencing is also carried out by a large family of transcriptional modulators called the Krüppel-associated box domain-containing zinc-finger proteins (KRAB-ZFPs) (Imbeault et al., 2017). These proteins include an array of zinc fingers, each of which binds short DNA sequences such that, together, they confer specificity to individual TE families (Thomas and Schneider, 2011). These proteins also include the KRAB domain, which recruits KAP1/TRIM28 and, in turn, a silencing complex of proteins that establish a repressive chromatin structure at TE loci (Ecco et al., 2017).

Although these TE silencing pathways are broadly conserved phylogenetically and functionally critical for maintaining genome integrity, they nonetheless evolve (Parhad and Theurkauf, 2019). Our work is motivated by the hypothesis that their evolution contributes to variation in TE content, and therefore overall genome size, across the tree of life (Mueller, 2017). Species that are extreme genome size outliers provide a powerful test of this hypothesis, as they are predicted to harbor strong signatures of divergent TE silencing compared with genomes of more typical size.

Among vertebrates, extreme genome expansion through TE accumulation evolved independently in salamanders and in lungfishes, with large increases in both lineages occurring over 200 million years ago (Liedtke et al., 2018;Meyer et al., 2021). Salamanders are one of the three clades of living amphibians; there are 775 extant species, and haploid genome sizes range from 9 to 120 Gb, reflecting ongoing genome size evolution (AmphibiaWeb, 2022;Gregory, 2022). Lungfishes are the sister taxon to tetrapods; there are 6 extant species, and haploid genome size estimates range from 40 Gb to 130 Gb (Meyer et al., 2021). Amphibians also include some of the smallest vertebrate genomes; the ornate burrowing frog *Platyplectrum ornatum* and the New Mexico spadefoot toad *Spea multiplicata* have genome sizes of 1.06 and 1.09 Gb, respectively (Lamichhaney et al., 2021;Gregory, 2022).

To date, several studies have begun to explore the relationship between TE silencing and genome size among vertebrates. At the smaller extremes, studies of frogs and fish with tiny genomes (≤ 1 Gb) revealed at least one additional duplicate copy of a PIWI gene, suggesting increased activity of the piRNA pathway in silencing TEs in genomes that have undergone size reduction (Malmstrøm et al., 2018;Lamichhaney et al., 2021). At the larger extremes, the data reveal a more complex picture; the Australian and African lungfish genomes (*Neoceratodus forsteri* and *Protopterus annectens*, ≥ 40 Gb) show neither gains nor losses of PIWI or related genes (Biscotti et al., 2017;Meyer et al., 2021). However, the African lungfish genome includes far more KRAB domains than other vertebrate genomes, suggesting a copy-number-based increase in activity. In contrast, the genome of the Mexican axolotl salamander *Ambystoma mexicanum* (~34 Gb) (Gregory, 2022) contains a comparable number of KRAB domains to mammalian and non-avian reptile genomes, suggesting no similar increase in this TE silencing activity (Wang et al., 2021b).

Expression data reveal a similarly mixed picture: for some piRNA pathway genes, germline expression is higher in salamanders (represented by the fire-bellied newt *Cynops orientalis*, ~43 Gb) than in the African lungfish, whereas for other genes, the pattern is reversed; comparisons with genomes of more typical size (coelacanth *Latimeria menadoensis* and zebrafish *Danio rerio*) show patterns of both higher and lower germline expression of TE silencing genes in the species with gigantic genomes (Biscotti et al., 2017;Carducci et al., 2021). Small RNA sequence data from the gonads of the northern dusky salamander *Desmognathus fuscus* (~15 Gb) reveal lower percentages of TE-mapping piRNAs than are found in smaller genomes, suggesting a less comprehensive TE-targeting piRNA pool in the gigantic genome (Madison-Villar et al., 2016). Taken together, these inconsistent patterns reveal that the relationship between TE silencing pathway activity and genome size evolution

remains incompletely understood, and that integrating genomic, transcriptomic, and small RNA analysis is critical for a complete picture.

Here we present a detailed analysis of TEs and germline TE silencing activity in the central Asian salamander *Ranodon sibiricus* — a range-restricted species endemic to China and Kazakhstan — adding both phylogenetic (family Hynobiidae) and genome size (~21 Gb) diversity to the small but growing dataset on TE silencing in gigantic genomes (AmphibiaWeb, 2022;Gregory, 2022). We quantify the expression of TEs in the male and female gonads, and we complement this data with analyses of the genomic TE landscape and TE amplification histories to reveal what TE superfamilies are active in the *R. sibiricus* genome. We quantify small RNAs expressed in male and female gonads and test whether small RNAs targeting TEs for silencing are expressed and amplified in proportion to TE expression. We quantify the relative expression of genes encoding proteins from two TE silencing pathways — piRNA and KRAB-ZFP. Finally, we extend these latter analyses to other vertebrates with a range of genome sizes to test for changes in TE silencing accompanying extreme increases in genome size.

## 2 Results

### 2.1 The genome of *R. sibiricus* contains diverse known, active TE superfamilies

We estimated the haploid genome size of *R. sibiricus* to be 17 Gb; averaging our result with published estimates (22.3 or 24.8 Gb) yields 21.3 Gb (Gregory, 2022). We used the PiRATE pipeline (Berthelier et al., 2018), which was designed to mine and classify repeats from low-coverage genomic shotgun data in taxa that lack genomic resources. The pipeline yielded 109,909 repeat contigs (Table1). RepeatMasker mined the most repeats (75,381 out of 109,909; 68.6%), followed by dnaPipeTE (21.9%), RepeatScout (3.3%), RepeatModeler (2.9%), and TE-HMMER (2.8%). TEdenovo, LTRharvest, HelSearch, SINE-Finder, and MITE-Hunter found few repeats, and MGEScan-non-LTR found none.

Repeat contigs were annotated as TEs to the levels of order and superfamily in Wicker's hierarchical classification system (Wicker et al., 2007), modified to include several recently discovered TE superfamilies, using PASTEC (Hoede et al., 2014). Of the 109,909 identified repeat contigs, 1,088 were filtered out as potential chimeras, 275 were classified as potential multiple-copy host genes, and 54,221 (49.33%) were classified as known TEs (Table 2), representing 23 superfamilies in eight orders as well as retrotransposon and transposon derivatives.

To calculate the proportion of different repeats in the genome, shotgun reads were masked with RepeatMasker using two *R. sibiricus*-derived repeat libraries: excluding or including unknown repeats. This comparison provided a rough approximation of the quantity of unknown repeats that were TE-derived, but divergent, fragmented, or otherwise unidentifiable by our pipeline.

Class I TEs (retrotransposons) make up 28.90%–48.43% (unknown repeats included or excluded in the repeat library, respectively) of the *R. sibiricus* genome; they are over 4 times more abundant than Class II TEs (DNA transposons; 5.48%–12.11%). *LINE/Jockey* is the most abundant superfamily (9.69%–12.12% of the genome), followed by *LINE/L1* (5.04%–6.62%), *LTR/Gypsy* (3.85–6.50%), and TRIM (3.80%–9.76%); all are retrotransposons or retrotransposon derivatives (Table 2). *TIR/PIF-harbinger* (2.98%–4.22%), *TIR/hAT* (1.12%–1.15%), and MITE (0.56%–4.07%) are the most abundant superfamilies of DNA transposons/transposon derivatives (Table 2).

Diversity of the overall genomic TE community was measured using both Simpson's and Shannon diversity indices, considering TE superfamilies as "species" and the total

number of base pairs for each annotated superfamily as individuals per "species." The Gini-Simpson Index (1-D) is 0.83, and the Shannon Index H is 1.92, similar to estimates of genomic diversity from other salamander species (Wang et al., 2021a;Haley and Mueller, 2022).

Seventeen superfamilies and three retrotransposon or transposon derivatives (each covering more than 0.05% of the genome) were selected for summaries of overall amplification history, generated by plotting the genetic distances between individual reads (representing TE loci) and the corresponding ancestral TE sequences as a histogram, with bins of 1%. All of the resulting distributions showed characteristics of ongoing or recent activity (*i.e.,* presence of TE sequences < 1% diverged from the ancestral sequence) (Figure 1).

Ten of these showed right-skewed, essentially monotonically decreasing distributions with a maximum or near-maximum at < 1% diverged from the ancestral sequence: *LTR/ERV*, *LTR/Copia*, *LTR/Bel-Pao*, *LINE/Jockey*, *LINE/L1*, *SINE/5S*, LARD, *TIR/hAT*, MITE, and *Helitron/Helitron*, suggesting TE superfamilies or derivatives that continue to be replicating today at their highest-ever rates of accumulation. In contrast, ten TE superfamilies or derivatives showed right-skewed, uni- or multimodal distributions: *LTR/Gypsy*, *DIRS/DIRS*, *PLE/Penelope*, *LINE/I*, *LINE/RTE*, TRIM, *TIR/PIF-Harbinger*, *TIR/Tc1-Mariner*, *TIR/PiggyBac*, and *Maverick/Maverick*. These 10 distributions suggest TE superfamilies that continue to be active today, but whose accumulation peaked at some point in the past.

**2.2 Germline relative TE expression is higher in males than females, but is correlated with genomic abundance in both sexes**

Our *de novo* gonad transcriptome assembly yielded 510,439 contigs (N50 = 1,250 bp; min and max contig lengths = 201 bp and 28,590 bp; total assembly length = 362,097,394 bp). The BUSCO pipeline revealed the presence of 95.3% of core vertebrate genes and 89.8% of core tetrapod genes. 47,182 contigs were annotated as TEs (representing 28 superfamilies), 64,409 as endogenous genes (representing 28,283 different genes), and 1,257 as having both a TE and an endogenous gene; the majority of contigs (72%) remained unannotated.

Endogenous genes account for the majority of expression in the gonads of both sexes (68% and 51% of summed TPM in females and males, respectively), followed by unannotated contigs (29% and 42%). Relative expression of TEs is an order of magnitude lower than endogenous gene expression in both sexes (2.4% and 5.6%) (Table 3, Supplementary File S3). Nine superfamilies (*LTR/Retrovirus*, *LINE/R2*, *SINE/7SL*, *TIR/MuDR*, *TIR/CACTA*, *TIR/ISL2EU*, *TIR/Ginger*, *TIR/Academ*, *TIR/P*) were detected at low expression levels in the transcriptome but were not initially detected in the genomic data (Table 2); mapping the genomic reads to these transcriptome contigs with Bowtie2 identified an average of 4 reads per superfamily, indicating their extremely low frequency in the genome. In contrast, only one superfamily (*LTR/Bel-Pao*) was detected in the genomic data but not in the transcriptome data. Overall, 19 superfamilies were identified both in the genomic contigs and transcriptome contigs.

In ovaries, autonomous TEs account for 8.9% of the total transcriptome contigs and 1.8% of the overall transcripts (summed TPM = 17,890) (Table 3). Non-autonomous TEs account for only 0.6% of the total transcriptome contigs, but still represent 0.6% of the overall transcripts (summed TPM = 5,538). In testes, relative TE expression is more than double that seen in females, with 4.4% and 1.2% of the overall gonad transcriptome accounted for by autonomous and non-autonomous TEs, respectively.

Differential expression analysis identified 780 contigs of 18 TE superfamilies as differently expressed between testes and ovaries. 678 TE transcripts were more highly expressed in testes, while only 102 TE transcripts were more highly expressed in ovaries

(Figure 2A). Of the nine superfamilies with more than ten differentially expressed transcripts between males and females, eight of them showed significantly higher relative expression in testes than ovaries, and one showed a non-significant trend towards higher testis expression (Figure 2B). Across the 19 TE superfamilies detected in both the genomic and transcriptomic datasets, genomic abundance is positively correlated with overall relative expression both in ovaries ($R = 0.786$, $P < 0.001$) and testes ($R=0.837$, $P < 0.001$), with male relative expression higher overall (Figure 2C).

**2.3 Expression of TE-mapping piRNAs correlates with TE expression in the gonads**

In both testes and ovaries, the length distribution of small RNAs includes a peak at 29 nucleotides (Figure 3A), and sequences up to 30 nt show a strong 5'-U bias at the first nucleotide position, consistent with expectations for the piRNA pool (Figure 3B). In ovaries, there is a second peak at 22 nucleotides. The relative expression of putative piRNAs (25–30 nt) is lower in females than in males. In contrast, the relative expression of ~22 nt RNAs is higher in females than in males; 41%–59% of 22 nt RNAs correspond to known miRNAs in females, versus 37%–41% in males.

A higher percentage of total putative piRNAs map to TEs in females than in males; on average, 22.7% map in the antisense direction and 23.5% map in the sense direction in females, and 11.0% map in the antisense direction and 10.1% map in the sense direction in males. Considering unique putative piRNA sequences, 16.9% map in the antisense direction and 17.1% map in the sense direction in females, and 15.1% map in the antisense direction and 14.7% map in the sense direction in males. Overall, more total putative piRNAs map to TEs in males than females, although the ranges overlap (1,264,088–3,045,727 in males vs. 897,586–2,503,814 in females) (Figure 3B).

In both sexes, we identify a peak overlap length between TE-mapping sense and anti-sense piRNAs of 10 base pairs, consistent with ping-pong amplification of piRNAs in response to TE transcription (Figure 3C, Supplementary S5, S6). The strength of the ping-pong signal, indicated by the Z-scores of the 10-nt overlap, is greater in females.

piRNA expression is correlated with TE expression, measured at the TE superfamily level, in both females and males (Figure 4, Supplementary File S4). At higher levels of TE expression, females show a trend of having more piRNAs relative to TE expression level than males.

**2.4 Germline expression of piRNA pathway genes is higher in males in *R. sibiricus*, whereas KRAB-ZFP silencing and miRNA pathway genes are higher in females**

The expression of piRNA pathway genes in *R. sibiricus* is higher in testes than in ovaries, measured both relative to miRNA pathway gene expression and as TPM (Figure 5; Supplementary Files 6,7). In contrast, the expression of genes establishing a repressive transcriptional environment (NuRD complex + related proteins) is comparable between the sexes relative to miRNA pathway gene expression and slightly higher in females measured as TPM. The expression of TRIM28 — which links KRAB-ZFP proteins to the NuRD complex + related proteins — is higher in females than males relative to miRNA expression, yet miRNA pathway expression levels (TPM) are slightly higher in females (consistent with higher miRNA expression, Figure 3; Supplementary Files 6,7). Taken together, these results suggest that males may rely more heavily on piRNA machinery to recruit repressive transcriptional machinery, whereas females may rely more heavily on KRAB-ZFP proteins to recruit repressive transcriptional machinery.

**2.5 Relative expression of TE silencing pathways between males and females varies across species**

Across species, higher piRNA pathway expression relative to miRNA pathway expression in males is seen in the majority of taxa across the range of genome sizes; the exceptions are *Gallus gallus* and *Protopterus annectens* (Figure 5, Supplementary File S6). Similarly, higher miRNA pathway expression in females (TPM) is seen in all taxa except *G. gallus* and *D. rerio*, although the difference is not always as pronounced as in *R. sibiricus* (Supplementary File S7). *Platyplectrum ornatus*, *Anolis carolinensis*, and *C. orientalis* show the same pattern of TE silencing expression differences between males and females as in *R. sibiricus*, with higher reliance on piRNA machinery in males and higher reliance on KRAB-ZFP in females. However, this pattern does not hold in other taxa (Figure 5).

### 2.6 Germline expression of TE silencing machinery does not correlate with genome size in either sex

Across the range of genome sizes from ~1 Gb to ~130 Gb, we find no correlations between genome size and 1) the expression of piRNA processing genes, 2) NuRD complex and associated genes establishing a repressive transcriptional environment, or 3) TRIM28. Interestingly, five of the six highest piRNA pathway expression levels are found in amphibians, the clade with the most variation in genome size (Figure 5, Supplementary File S9).

## 3 Discussion

### 3.1 Permissive TE environment despite comprehensive piRNA-mediated silencing

The transposable element community in *Ranodon sibiricus* is comparable in diversity to other gigantic amphibian genomes and shares many of the same abundant TE superfamilies (e.g., LTR/Gypsy, DIRS and LINE/L1) (Sun et al., 2012;Sun and Mueller, 2014;Wang et al., 2021a). However, *R. sibiricus* differs from other salamanders in having high levels of LINE/Jockey, a superfamily shown to be abundant in the caecilian *Ichthyophis bannanicus*, but rare in other salamanders (Wang et al., 2021a). Both genomic and transcriptomic data suggest that all TE superfamilies are potentially experiencing ongoing transposition in *R. sibiricus*. Thus, like other gigantic amphibian genomes, *R. sibiricus* appears to have attained its large size because of the expansion of multiple types of TEs, supporting the notion of amphibian genomes as permissive TE environments. At the same time, all of these putatively active TE superfamilies are targeted by piRNAs, and piRNA levels are correlated with TE expression at the superfamily level, consistent with patterns in *Drosophila* and suggesting that the scope, if not efficacy, of TE silencing in *Ranodon sibiricus* is comparable to other species (Kelleher and Barbash, 2013;Saint-Leandre et al., 2020).

### 3.2 Sex-biased TE expression and silencing

Transposable element expression is higher in *R. sibiricus* testes than ovaries, a pattern also reported in *Drosophila* (Wei et al., 2022) and *Oryzias latipes* (the medaka fish) (Saint-Leandre et al., 2020;Dechaud et al., 2021), but opposite the pattern reported in *Corvus corone* (carrion crow) (Warmuth et al., 2022) and different from the non-sex-biased TE expression in the fire-bellied newt *Cynops orientalis* (Carducci et al., 2021). In *Drosophila*, testes expression levels of TE-mapping piRNAs and piRNA pathway genes, as well as the ping-pong signature, are lower than ovary expression levels, suggesting that lower male piRNA-mediated silencing contributes to higher male TE expression (Saint-Leandre et al., 2020;Chen et al., 2021). In *O. latipes*, on the other hand, testes expression levels of TE-mapping piRNAs are higher than ovary expression levels — a pattern also observed in zebrafish — suggesting that higher male piRNA-mediated silencing may actually be correlated with high male TE expression in these two fish species (Houwing et al., 2007;Kneitz et al., 2016). In *R. sibiricus*, TE-mapping piRNA counts are similar between the sexes, despite higher relative expression

of putative piRNAs in males than females (Figure 3a,b). However, the ping-pong amplification signature is higher in females, as is the number of piRNAs per expressed TE transcript, suggestive of a more robust piRNA-directed silencing response in females (Figure 3c, 4). On the other hand, the piRNA pathway protein expression levels are higher in males, suggesting the opposite case of a more robust response in males (Figure 5).

Sex-specific differences in gonadal TE expression have also been explained by factors other than variation in TE silencing; for example, in systems with heteromorphic sex chromosomes, sex-biased TE expression has been attributed to different TE dynamics on the sex-limited chromosome (Y in XY systems, or W in ZW systems). TE abundance is higher on sex-limited chromosomes because of lower effective population size, lack of recombination, and lower gene density. In addition, TE expression per locus has been shown to be higher on the sex-limited chromosome itself — as well as genome-wide — in the heterogametic sex in *Drosophila* (Y, males) and in crows (*Corvus corone*; W, females) (Wei et al., 2020;Warmuth et al., 2022). The mechanism for higher TE expression in the heterogametic sex in these cases remains incompletely understood, but may involve TEs affecting their own genome-wide regulation *in trans* or the heightened conflict between creating a repressive chromatin state to silence TEs while maintaining open chromatin to allow genic transcription on a degenerating chromosome (Wei et al., 2020;Warmuth et al., 2022). Neither *Ranodon sibiricus* nor *Oryzias latipes* has heteromorphic sex chromosomes (Hillis and Green, 1990;Matsuda et al., 2002;Evans et al., 2012;Perkins et al., 2019), yet both show sex-biased TE expression, and only *Oryzias* shows unambiguous sex-biased TE-targeting piRNA expression; taken together, these data reveal that the difference in TE expression between sexes reflects different underlying causes across species. The relationships among sex-biased TE expression, sex determination, and TE silencing are an important target for future research, but irrespective of the underlying mechanisms, the sex with higher TE activity contributes more to the species' genomic TE load. In *R. sibiricus*, that sex is the male, provided that TE expression is a reasonable proxy for transposition.

### 3.3 The TE-silencing arms race dynamic across genome sizes

Transposable elements and the silencing machinery of their hosts are engaged in an arms race in which a novel TE family initially proliferates, the host evolves silencing based on TE sequence identification, and the TE subsequently diverges to evade silencing — or that particular TE remains permanently silenced, but a novel TE invades the host genome and begins the cycle anew (Luo et al., 2020;Zhang et al., 2020;Said et al., 2022;Wei et al., 2022). If balanced by deletion of TE sequences, this arms-race dynamic can be associated with fairly stable genome size over evolutionary timescales, despite turnover in TE content (Kapusta et al., 2017). To yield an overall evolutionary trend in genome size, the long-term balance between TE insertion and deletion has to become skewed in favor of one or the other, with deletion bias leading to genome contraction and insertion bias leading to genome expansion (Nam and Ellegren, 2012).

It has been suggested in the literature that large genomes may be manifestations of an arms race between TEs and the silencing of their hosts, but that this arms race involves the TE community as a whole rather than individual TE families — and the species with the gigantic genome has been interpreted as both the current "attacker" and "defender" in the arms race. For example, Meyer et al (2021) suggested that TE silencing machinery did not adapt to reduce TE expansion in the Australian lungfish, based on high genomic TE abundance and ongoing TE expression. In the strawberry poison dart frog, which has a moderately expanded genome size of 6.76 Gb, a widespread failure to silence TEs was suggested based on high genomic TE abundance, germline TE expression of diverse TEs, and the presence of an identified *piwi2* and *piwi3* transcript in only one individual sampled each

(Rogers et al., 2018). In the salamander *Desmognathus fuscus*, less comprehensive piRNA pathway-mediated TE silencing was suggested based on a relatively low percentage of TE-mapping piRNAs (Madison-Villar et al., 2016). In all three of these examples, the TEs are suggested to be currently on the attack. On the other hand, in the African lungfish, Wang et al. (2021) suggested that the KRAB-ZFP TE transcriptional silencing machinery has expanded in scope in response to the high genomic TE load. Similarly, in the fire-bellied newt *Cynops orientalis*, it was suggested that TE silencing is now enhanced in response to high TE load, which accumulated in the past during a period of increased TE mobilization (Carducci et al., 2021). In both of these examples, the TEs are suggested to be currently on the defensive. Thus, two opposite predictions for TE silencing in gigantic genomes — either increased or decreased — have been proposed to be met in different extant organisms, albeit with datasets that are not necessarily comparable. This apparent conflict can be resolved by considering that the attack/defense status of TEs and their silencing reflect where the lineage currently exists in the dynamic cycle between TE and host dominance. Our results revealing no consistent pattern in TE silencing pathway expression levels and genome size (Figure 5) are consistent with this interpretation. At large sizes, it is not the size of the genome itself that likely predicts the efficacy of the TE silencing machinery, but more likely the directional trend in genome size evolution; genomes that are contracting are more likely to have effective TE silencing, whereas genomes undergoing expansion are more likely to have reduced TE silencing. In the absence of comparable data (including small RNA, TE expression and amplification, and silencing pathway expression) for other species with known trends in genome size evolution, we opt for a conservative position and do not infer *Ranodon sibiricus'* current position in the TE/host dynamic arms race cycle.

The mechanisms by which global TE silencing mechanisms can be subverted by a community of TEs, and then evolve to regain stricter control, are not yet well-understood. A few studies in invertebrates have begun to reveal differences in TE and silencing dynamics in genomes of different sizes. In a pairwise comparison of grasshopper species with different genome sizes, Liu et al. revealed that the species with the larger genome had higher TE expression, lower piRNA abundance, and lower expression of the piRNA biogenesis gene *HENMT*, which suggested that lower piRNA-mediated TE silencing was permissive to higher TE activity and genome expansion (Liu et al., 2022). A comparison between *Drosophila melanogaster* and the mosquito *Aedes aegypti*, which has a larger genome (1.38 Gb versus 180 Mb), revealed that the mosquito has a higher TE load and a smaller percentage of TE-mapping piRNAs (Arensburger et al., 2011). Future studies that leverage the large range of genome sizes present in vertebrates, emphasizing comprehensive across-species data on TE activity and TE silencing in a phylogenetic context to allow ancestral genome size reconstruction, will continue to shed light on how TEs and their hosts coevolve to achieve gigantic genomes.

## 4 Materials and methods

### 4.1 Specimen information

We collected three male adult *Ranodon sibiricus* from the wild of Wenquan County, Xinjiang Uygur Autonomous Region of China, and egg-hatched and raised one male and four females in an aquarium of Xinjiang Normal University from eggs originally collected from the same field site as the wild males. All these individuals were collected during the breeding season of August, 2017, and all adults had a snout-tail length of 16–21 cm and a body mass of 12–35 g prior to euthanasia (Supplementary File S1). Wild-caught adults were euthanized upon return to the laboratory and were not kept alive in captivity. Collection, hatching, and euthanasia were performed following Animal Care & Use Protocols approved by Chengdu Institute of Biology, Chinese Academy of Sciences.

### 4.2 Genome size estimation

Blood smears were prepared from a formalin-fixed specimen of *Ranodon sibiricus* and nuclear area was measured from Fuelgen-stained red blood cell nuclei using the ImagePro® image analysis program (Itgen et al., 2022). Blood smears of the reference standards *Ambystoma mexicanum* (34 Gb; Gregory, 2022) and the Iberian ribbed newt *Pleurodeles waltl* (20 Gb; Gregory 2022) were prepared and analyzed at the same time under the same conditions.

### 4.3 Genomic shotgun library creation, sequencing, and assembly

Total DNA was extracted from muscle tissue using the modified low-salt CTAB extraction of high-quality DNA procedure (Arseneau et al., 2017). DNA quality and concentration were assessed using agarose gel electrophoresis and a NanoDrop Spectrophotometer (ThermoFisher Scientific, Waltham, MA), and a PCR-free library was prepared using the NEBNext Ultra DNA Library Prep Kit for Illumina. Sequencing was performed on one lane of a Hiseq 2500 platform (PE250). Library preparation and sequencing were performed by the Beijing Novogene Bioinformatics Technology Co. Ltd. Raw reads were quality-filtered and adaptor-trimmed using Trimmomatic-0.39 (Bolger et al., 2014) with default parameters. In total, the genomic shotgun dataset included 11,960,858 reads. After filtering and trimming, 11,168,678 reads covering a total length of 2,314,096,923 bp remained. Thus, the sequencing coverage is about 10.1% (0.1X coverage). Filtered, trimmed reads were assembled into contigs using dipSPAdes 3.12.0 (Bankevich et al., 2012) with default parameters, yielding 478,991 contigs with an N50 of 447 bp and a total length of 249,425,929 bp.

### 4.4 Mining and classification of repeat elements

The PiRATE pipeline was used as in the original publication (Berthelier et al., 2018), including the following steps: **1)** Contigs representing repetitive sequences were identified from the assembled contigs using similarity-based, structure-based, and repetitiveness-based approaches. The similarity-based detection programs included RepeatMasker v-4.1.0 (http://repeatmasker.org/RepeatMasker/, using Repbase20.05_REPET.embl.tar.gz as the library instead) and TE-HMMER (Eddy, 2011). The structural-based detection programs included LTRharvest (Ellinghaus et al., 2008), MGEScan non-LTR (Rho and Tang, 2009), HelSearch (Yang et al., 2009), MITE-Hunter (Han and Wessler, 2010), and SINE-finder (Wenke et al., 2011). The repetitiveness-based detection programs included TEdenovo (Flutre et al., 2011) and RepeatScout (Price et al., 2005). **2)** Repeat consensus sequences (*e.g.*, representing multiple subfamilies within a TE family) were also identified from the cleaned, filtered, and unassembled reads with dnaPipeTE (Goubert et al., 2015) and RepeatModeler (http://www.repeatmasker.org/RepeatModeler/). **3)** Contigs identified by each individual program in steps 1 and 2, above, were filtered to remove those < 100 bp in length and clustered with CD-HIT-est (Li and Godzik, 2006) to reduce redundancy (100% sequence identity cutoff). This yielded a total of 155,999 contigs. **4)** All 155,999 contigs were then clustered together with CD-HIT-est (100% sequence identity cutoff), retaining the longest contig and recording the program that classified it. 46,090 contigs were filtered out at this step. **5**) The remaining 109,909 repeat contigs were annotated as TEs to the levels of order and superfamily in Wicker's hierarchical classification system (Wicker et al., 2007), modified to include several recently discovered TE superfamilies using PASTEC (Hoede et al., 2014), and checked manually to filter chimeric contigs and those annotated with conflicting evidence (Supplementary File S2). **6**) All classified repeats ("known TEs" hereafter), along with the unclassified repeats ("unknown repeats" hereafter) and putative multi-copy host genes, were combined to produce a *Ranodon*-derived repeat library. **7**) For

each superfamily, we collapsed the contigs to 95% and 80% sequence identity using CD-HIT-est to provide an overall view of within-superfamily diversity; 80% is the sequence identity threshold used to define TE families (Wicker et al., 2007).

### 4.5 Characterization of the overall repeat element landscape

Overlapping paired-end shotgun reads were merged using PEAR v.0.9.11 (Zhang et al., 2014) with the following parameter values based on our library insert size and trimming parameters: min-assemble-length 36, max-assemble-length 490, min-overlap size 10. After merging, 7,385,166 reads remained (including both merged and singletons), with an N50 of 388 bp and total length of 1,997,175,501 bp. To calculate the percentage of the *R. sibiricus* genome composed of different TEs, these shotgun reads were masked with RepeatMasker v-4.1.0 using two versions of our *Ranodon*-derived repeat library: one that included the unknown repeats and the other that excluded them. In both cases, simple repeats were identified using the Tandem Repeat Finder module implemented in RepeatMasker. The overall results were summarized at the levels of TE class, order, and superfamily.

### 4.6 Measuring diversity of the genomic TE community

Unknown repeats were excluded from the analysis, as were TEs that could only be annotated down to the level of Class. Simpson's diversity index is expressed as the variable D, calculated by: $D = \frac{\sum n(n-1)}{N(N-1)}$ (Simpson, 1949). D is the probability that two individuals at random pulled from a community will be from the same species. We report 1 – D, or the Gini-Simpson's index, which is more intuitive. The Shannon's diversity index H is calculated by: $H = -\sum_{i=1}^{s} p_i \, ln \, p_i$ (Shannon, 1948). The higher the value of H, the greater the diversity.

### 4.7 Amplification history of TE superfamilies

To summarize the overall amplification history of TE superfamilies and test for ongoing activity, the perl script parseRM.pl (Kapusta et al., 2017) was used to parse the raw output files from RepeatMasker (.align) and report the sequence divergence between each read and its respective consensus sequence (parameter values = -l 50,1 and -a 5). The repeat library used to mask the reads comprised the 55,327 TE contigs classified by the PiRATE pipeline and clustered at 100% sequence identity. Each TE superfamily is therefore represented by multiple consensus sequences corresponding to the family and subfamily TE taxonomic levels (*i.e.*, not the distant common ancestor of the entire superfamily). For each superfamily, histograms were plotted to summarize the percent divergence of all reads from their closest (*i.e.,* least divergent) consensus sequence. These histograms do not allow the delineation between different amplification dynamics scenarios (*i.e.,* a single family with continuous activity *versus* multiple families with successive bursts of activity). Rather, these global overviews were examined for overall shapes consistent with ongoing activity (*i.e.,* the presence of TE loci < 1% diverged from the ancestral sequence and a unimodal, right-skewed, J-shaped, or monotonically decreasing distribution).

### 4.8 Transcriptome library creation, sequencing, assembly, and TE annotation

Total RNA was extracted separately from testes ($n$ = 4) and ovary ($n$ = 4) tissues using TRIzol (Invitrogen). For each sample, RNA quality and concentration were assessed using agarose gel electrophoresis, a NanoPhotometer spectrophotometer (Implen, CA), a Qubit 2.0 Fluorometer (ThermoFisher Scientific), and an Agilent BioAnalyzer 2100 system (Agilent Technologies, CA), requiring an RNA integrity number (RIN) of 8.5 or higher; one ovary sample failed to meet these quality standards and was excluded from downstream analyses. Sequencing libraries were generated using the NEBNext Ultra RNA Library Prep Kit for

Illumina following the manufacturer's protocol. After cluster generation of the index-coded samples, the library was sequenced on one lane of an Illumina Hiseq 4000 platform (PE 150). Transcriptome sequences were filtered using Trimmomatic-0.39 with default parameters (Bolger et al., 2014). 30,848,170 to 39,695,323 reads were retained for each testis or ovary sample, and in total, 290,925,984 reads remained, with a total length of 42,385,060,050 bp. Remaining reads of all testes and ovary samples were combined and assembled using Trinity 2.12.0 (Haas et al., 2013), yielding 573,144 contigs (i.e., putative assembled transcripts). Contigs were clustered using CD-hit-est (95% identity). Completeness of this final *de novo* transcriptome assembly were assessed using the BUSCO pipeline (Simao et al., 2015).

Expression levels of contigs in each sample were measured with Salmon (Patro et al., 2017), and contigs with no raw counts were removed. To annotate the remaining contigs containing autonomous TEs, BLASTp and BLASTx were used against the Repbase Database (downloaded on January 5, 2022) with an E-value cutoff of 1E-5 and 1E-10, respectively. The aligned length coverage was set to exceed 80% of the queried transcriptome contigs. To annotate contigs containing non-autonomous TEs, RepeatMasker was used with our *Ranodon*-derived genomic repeat library of non-autonomous TEs (LARD-, TRIM-, MITE-, and SINE-annotated contigs) and the requirement that the transcriptome/genomic contig overlap was > 80 bp long, > 80% identical in sequence, and covered > 80% of the length of the genomic contig. Contigs annotated as conflicting autonomous and non-autonomous TEs were filtered out.

To identify contigs that contained endogenous *R. sibiricus* genes, the Trinotate annotation suite (Bryant et al., 2017) was used with an E-value cutoff of 1E-5 for both BLASTx and BLASTp against the Uniport database, and 1E-5 for HMMER against the Pfam database (Wheeler and Eddy, 2013). To identify contigs that contained both a TE and an endogenous gene (i.e., putative cases where a TE and a gene were co-transcribed on a single transcript), all contigs that were annotated both by Repbase and Trinotate were examined, and the ones annotated by Trinotate to contain a TE-encoded protein (*i.e.*, the contigs where Repbase and Trinotate annotations were in agreement) were not further considered. The remaining contigs annotated by Trinotate to contain a non-TE gene (*i.e.*, an endogenous *Ranodon* gene) and also annotated either by Repbase to include a TE-encoded protein or by RepeatMasker to include a non-autonomous TE were identified for further examination and expression-based analysis.

### 4.9 Germline TE expression quantification in males and females

Expression levels of the individual TE superfamilies were calculated by averaging the TPM values among replicates of each sex and then summing the average TPM of all contigs annotated to each superfamily. For TE superfamilies detected in both the genomic and transcriptomic datasets, we tested for a relationship between genomic abundance and expression levels in each sex using linear regression on log-transformed data.

To identify differentially expressed contigs between males and females, DESeq2 (Love et al., 2014) was used with an adjusted P-value cut off of 0.05. Among the 15,011 total differentially expressed transcripts between sexes (including TEs, endogenous genes, and unannotated contigs), 869 were TEs, representing 18 superfamilies and other unknown TEs. Superfamilies with fewer than 10 differentially expressed transcripts between sexes were removed, leaving 9 superfamilies; for each, we tested for a difference in expression between males and females using a t-test.

### 4.10 Identification of putative piRNAs from small RNA-seq data

Small RNA libraries were prepared for each sample using the NEBNext® Multiplex Small RNA Library Prep Set for Illumina® (NEB, USA) following the manufacturer's

recommendations, and index codes were added to attribute sequences to each sample. Briefly, the NEB 3' SR Adaptor (5'-AGATCGGAAGAGCACACGTCT-3') was ligated to the 3' end of small RNA molecules. After the 3' ligation reaction, the SR RT Primer was hybridized to the excess 3' SR Adaptor (that remained free after the 3' ligation reaction), transforming the single-stranded DNA adaptor into a double-stranded DNA molecule (dsDNAs). This step was important for preventing adaptor-dimer formation, and because dsDNAs are not substrates for ligation mediated by T4 RNA Ligase 1, they therefore would not ligate to the 5' SR Adaptor (5'-GTTCAGAGTTCTACAGTCCGACGATC-3') in the subsequent ligation step. The 5' end adapter was then ligated to the 5' ends of small RNA molecules. First strand cDNA was synthesized using M-MuLV Reverse Transcriptase (Rnase H). PCR amplification was performed using LongAmp Taq 2X Master Mix, SR Primer for Illumina, and index primers. PCR products were purified on an 8% polyacrylamide gel (100V, 80 min). DNA fragments corresponding to ~140-160 bp (the length of a small noncoding RNA plus the 3' and 5' adaptors) were recovered and dissolved in 8 µL elution buffer. Library quality was assessed on the Agilent Bioanalyzer 2100 using DNA High Sensitivity Chips. Clustering of index-coded samples was performed on a cBot Cluster Generation System using the TruSeq SR Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, libraries were sequenced on an Illumina Hiseq 2500 platform (SE50).

We filtered low-quality sequences using the fastq_quality_filter (-q 20, -p 90) in the FASTX-Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/). We removed the adapter sequence with a minimum overlap of 10 bp from the 3'-end, discarded untrimmed reads, and selected those with a minimum length of 18 bp and a maximum length of 40 bp (after cutting adapters) and no Ns using cutadapt v2.8 (Martin, 2011). Reads mapping to the mitochondrial genome (NCBI code: AJ419960) and riboRNAs (NCBI codes: DQ283664, AJ279506, MH806872) were identified and filtered out using Bowtie v1.1.0 (Langmead et al., 2009). Overall, more reads were filtered out based on length and rRNA identity in females than males (Supplementary File S5). miRNAs 21-24 nt in length were annotated using Bowtie v1.1.0 to identify hits to miRbase 22.1 (Kozomara et al., 2019) in each male and female.

To test for the predicted bias towards U at the first 5' nucleotide position of piRNAs, we calculated the proportion of small RNAs with each nucleotide in the first position. Based on this result, and the overall length distribution of RNAs between 18 and 40 nt, we conservatively defined putative piRNAs as those ranging from 25-30 nt, and we selected these using the seqkit software (Shen et al., 2016). We mapped these putative piRNAs to our transcriptome assembly using Bowtie and identified piRNAs that map to autonomous TEs (i.e. those that include transposition-related ORFs) in the sense and antisense orientations.

### 4.11 Ping-pong signature analysis

Secondary piRNA biogenesis associated with piRNA-targeted post-transcriptional TE silencing produces a distinctive "ping-pong signature" in the piRNA pool, which consists of a 10 bp overlap between the 5' ends of antisense and sense piRNAs. The ping-pong signature for each individual was analyzed using the following approach: First, TE transcripts that were not mapped by both sense-oriented and antisense-oriented piRNAs were filtered out using Bowtie, allowing 0 mismatches for sense mapping under the assumption that piRNAs derived directly from an RNA target should have the identical sequence (Teefy et al., 2020) and 3 mismatches for antisense mapping because cleavage of RNA targets can occur with imperfect base-pairing (Zhang et al., 2015). Second, the fractions of overlapping pairs of sense/antisense piRNAs corresponding to specific lengths, as well as the Z-score measuring the significance of each ping-pong signature, were generated using the script of 1_piRNA_and_Degradome_Counts.RMD (Teefy et al., 2020).

### 4.12 Putative piRNAs targeting TE superfamilies

To estimate levels of piRNAs targeting each TE superfamily, putative piRNAs were mapped to the TE transcripts using the 'align_and_estimate_abundance.pl' script of Trinity (Haas et al., 2013). Reads per million (RPM) values were calculated for each TE contig and then averaged across individuals of each sex. For each sex, overall putative piRNA levels targeting each TE superfamily were calculated by summing across all contigs annotated to the same TE superfamily. We tested for a correlation between TE superfamily expression level and targeting piRNA abundance using linear regression on the log-transformed variables.

### 4.13 Germline TE silencing pathway expression across genome sizes

To test for an association between overall piRNA or KRAB-ZFP pathway activity and genome size, we first compiled male and female gonad RNA-Seq datasets for vertebrates of diverse genome sizes, including *Platyplectrum ornatum* (ornate burrowing frog), *Gallus gallus* (chicken), *Danio rerio* (zebrafish), *Xenopus tropicalis* (Western clawed frog), *Anolis carolinensis* (green anole), *Mus musculus* (mouse), *Geotrypetes seraphini* (Gaboon caecilian), *Rhinatrema bivittatum* (two-lined caecilian), and *Caecilia tentaculata* (bearded caecilian) spanning genomes sizes from 1.0 – 5.5 Gb, and *Pleurodeles waltl* (the Iberian ribbed newt), *Ambystoma mexicanum* (the Mexican axolotl), *Cynops orientalis* (the fire-bellied newt), *Protopterus annectens,* and *P. aethiopicus* (African and marbled lungfishes) spanning genome sizes from 20 – ~130 Gb (Supplementary File S8). We performed de novo assemblies using the same pipeline as for *R. sibiricus* on all obtained datasets.

We identified transcripts of 21 genes receiving a direct annotation of piRNA processing in vertebrates in the Gene Ontology knowledgebase that were present in the majority of our target species: ASZ1, BTBD18 (BTBDI), DDX4, EXD1, FKBP6, GPAT2, HENMT1 (HENMT), MAEL, MOV10l1 (M10L1), PIWIL1, PIWIL2, PIWIL4, PLD6, TDRD1, TDRD5, TDRD6, TDRD7, TDRD9, TDRD12 (TDR12), TDRD15 (TDR15), and TDRKH. In addition, we identified transcripts of 14 genes encoding proteins that create a transcriptionally repressive chromatin environment in response to recruitment by PIWI proteins or KRAB-ZFP proteins, 12 of which received a direct annotation of NuRD complex in the Gene Ontology knowledgebase and 2 of which were taken from the literature: CBX5, CHD3, CHD4, CSNK2A1 (CSK21), DNMT1, GATAD2A (P66A), MBD3, MTA1, MTA2, RBBP4, RBBP7, SALL1, SETDB1 (SETB1), and ZBTB7A (ZBT7A) (cite Ecco et al 2017) (Wang et al., 2022). Finally, we identified TRIM28, which bridges this repressive complex to TE-bound KRAB-ZFP proteins in lobe-finned fishes (Ecco et al., 2017). For comparison, we identified transcripts of 14 protein-coding genes receiving a direct annotation of miRNA processing in vertebrates in the Gene Ontology knowledgebase, which we did not predict to differ in expression based on genome size: ADAR (DSRAD), AGO1, AGO2, AGO3, AGO4, DICER1, NUP155 (NU155), PUM1, PUM2, SNIP1, SPOUT1 (CI114), TARBP2 (TRBP2), TRIM71 (LIN41), and ZC3H7B. Expression levels for each transcript in each individual were measured with Salmon (Patro et al., 2017) (Supplementary File S9).

As a proxy for overall piRNA silencing activity, for each individual, we calculated the ratio of total piRNA pathway expression (summed TPM of 21 genes) to total miRNA pathway expression (summed TPM of 14 genes). As a proxy for transcriptional repression driven by both the piRNA pathway and KRAB-ZFP binding activity, we calculated the ratio of total transcriptional repression machinery expression (summed TPM of 14 genes) to total miRNA pathway expression. Finally, we calculated the ratio of TRIM28 expression to total miRNA pathway expression for each individual. We also calculated these ratios with a more conservative dataset allowing for no missing genes; this yielded 15 piRNA pathway genes, 9

KRAB-ZFP genes, and 13 miRNA genes. We plotted these ratios to reveal any relationship between TE silencing pathway expression and genome size.

## 5 Data availability statement

Genomic shotgun and transcriptome sequences have been deposited in the Genome Sequence Archive at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation (GSA: CRA008892, CRA008899, CRA008900), and are publicly accessible at http://bigd.big.ac.cn/gsa.

## 6 Author contributions

**Jie Wang**: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition. **Liang Yuan**: Specimen collection. **Jiaxing Tang Jiongyu Liu, Cheng Sun, Michael W. Itgen, & Guiying Chen**: Methodology, Software. **Stanley K. Sessions**: Investigation, Resources. **Guangpu Zhang**: Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Rachel Lockridge Mueller**: Conceptualization, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition. All authors read and approved the final manuscript.

## 7 Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 32170435, 31570391 to WJ) and the National Science Foundation of USA (Grant No. 1911585 to RLM).

## 8 Acknowledgements

We gratefully acknowledge Ye Xu and Yuzhou Gong for target tissue dissection; Xiuling Wang for help in field work; Jianping Jiang, Ava Louise Haley, and Chaochao Yan for their expertise in facilitating data analysis.

## 9 Literature cited

Almeida, M.V., Vernaz, G., Putman, A.L.K., and Miska, E.A. (2022). Taming transposable elements in vertebrates: from epigenetic silencing to domestication. *Trends in Genetics* 38**,** 529-553.

Amphibiaweb (2022). AmphibiaWeb: Information on amphibian biology and conservation. Berkeley, California.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., Chien, M., Russo, J.J., Ju, J., Sheridan, R., Sander, C., Zavolan, M., and Tuschl, T. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442**,** 203-207.

Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T., and Hannon, G.J. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular Cell* 31**,** 785-799.

Arensburger, P., Hice, R.H., Wright, J.A., Craig, N.L., and Atkinson, P.W. (2011). The mosquito *Aedes aegypti* has a large genome size and high transposable element load

but contains a low proportion of transposon-specific piRNAs. *BMC Genomics* 12**,** 606.

Arkhipova, I.R. (2018). Neutral theory, transposable elements, and eukaryotic genome evolution. *Mol Biol Evol* 35**,** 1332-1337.

Arseneau, J.R., Steeves, R., and Laflamme, M. (2017). Modified low-salt CTAB extraction of high-quality DNA from contaminant-rich tissues. *Mol Ecol Resour* 17**,** 686-693.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., and Prjibelski, A.D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19**,** 455-477.

Berthelier, J., Casse, N., Daccord, N., Jamilloux, V., Saint-Jean, B., and Carrier, G. (2018). A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga Tisochrysis lutea. *BMC Genomics* 19**,** 378.

Biscotti, M.A., Canapa, A., Forconi, M., Gerdol, M., Pallavicini, A., Schartl, M., and Barucca, M. (2017). The small noncoding RNA processing machinery of two living fossil species, lungfish and coelacanth, gives new insights into the evolution of the Argonaute protein family. *Genome Biology and Evolution* 9**,** 438-453.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30**,** 2114-2120.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., and Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biol* 19**,** 199.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* 128**,** 1089-1103.

Bryant, D.M., Johnson, K., Ditommaso, T., Tickle, T., Couger, M.B., Payzin-Dogru, D., Lee, T.J., Leigh, N.D., Kuo, T.-H., and Davis, F.G. (2017). A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Reports* 18**,** 762-776.

Carducci, F., Carotti, E., Gerdol, M., Greco, S., Canapa, A., Barucca, M., and Biscotti, M.A. (2021). Investigation of the activity of transposable elements and genes involved in their silencing in the newt *Cynops orientalis*, a species with a giant genome. *Scientific Reports* 11**,** 1-11.

Castanera, R., Borgognone, A., Pisabarro, A.G., and Ramírez, L. (2017). Biology, dynamics, and applications of transposable elements in basidiomycete fungi. *App Microbiol Biotech* 101**,** 1337-1350.

Castel, S.E., and Martienssen, R.A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature Reviews Genetics* 14**,** 100-112.

774 Chen, P., Kotov, A.A., Godneeva, B.K., Bazylev, S.S., Olenina, L.V., and Aravin, A.A.
775 (2021). piRNA-mediated gene regulation and adaptation to sex-specific transposon
776 expression in *D. melanogaster* male germline. *Genes & Development* 35**,** 914-935.

777 Czech, B., and Hannon, G.J. (2016). One loop to rule them all: The ping-pong cycle and
778 piRNA-guided silencing. *Trends in Biochemical Sciences* 41**,** 324-337.

779 Czech, B., Munafò, M., Ciabrelli, F., Eastwood, E.L., Fabry, M.H., Kneuss, E., and Hannon,
780 G.J. (2018). piRNA-guided genome defense: from biogenesis to silencing. *Annual*
781 *Review of Genetics* 52**,** 131-157.

782 De Koning, A.J., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive
783 elements may comprise over two-thirds of the human genome. *PLoS Genetics* 7**,**
784 e1002384.

785 Dechaud, C., Miyake, S., Martinez-Bengochea, A., Schartl, M., Volff, J.-N., and Naville, M.
786 (2021). Clustering of sex-biased genes and transposable elements in the genome of
787 the medaka fish *Oryzias latipes*. *Genome Biology and Evolution* 13**,** evab230.

788 Deniz, Ö., Frost, J.M., and Branco, M.R. (2019). Regulation of transposable elements by
789 DNA modifications. *Nature Reviews Genetics* 20**,** 417-431.

790 Doolittle, W.F., and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome
791 evolution. *Nature* 284**,** 601-603.

792 Ecco, G., Imbeault, M., and Trono, D. (2017). KRAB zinc finger proteins. *Development* 144**,**
793 2719-2729.

794 Eddy, S.R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol* 7**,** e1002195.

795 Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible
796 software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9**,** 18.

797 Evans, B.J., Alexander Pyron, R., and Wiens, J.J. (2012). "Polyploidization and sex
798 chromosome evolution in amphibians," in *Polyploidy and Genome Evolution*.
799 Springer), 385-410.

800 Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable
801 element diversification in de novo annotation approaches. *PLoS One* 6**,** e16526.

802 Goubert, C., Modolo, L., Vieira, C., Valientemoro, C., Mavingui, P., and Boulesteix, M.
803 (2015). De novo assembly and annotation of the Asian tiger mosquito (*Aedes*
804 *albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative
805 analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biology and*
806 *Evolution* 7**,** 1192-1205.

807 Gregory, T.R. (2022). "Animal Genome Size Database (http://www.genomesize.com)".).

808 Gunawardane, L.S., Saito, K., Nishika, K.M., Kawamura, Y., Nagami, T., Siomi, H., and
809 Siomi, M.C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end
810 formation in Drosophila. *Science* 315**,** 1587-1590.

811 Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger,
812 M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N.,
813 Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R.,
814 Leduc, R.D., Friedman, N., and Regev, A. (2013). De novo transcript sequence
815 reconstruction from RNA-seq using the Trinity platform for reference generation and
816 analysis. *Nature Protocols* 8**,** 1494-1512.

817 Haberer, G., Young, S., Bharti, A.K., Gundlach, H., Raymond, C., Fuks, G., Butler, E., Wing,
818 R.A., Rounsley, S., Birren, B., Nusbaum, C., Mayer, K.F.X., and Messing, J. (2005).
819 Structure and architecture of the maize genome. *Plant Physiology* 139**,** 1612-1624.

820 Haley, A.L., and Mueller, R.L. (2022). Transposable Element Diversity Remains High in
821 Gigantic Genomes. *J Mol Evol* 90**,** 332-341.

822 Han, Y.J., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature
823 inverted-repeat transposable elements from genomic sequences. *Nucleic Acids*
824 *Research* 38.

825 Hancks, D.C., and Kazazian, H.H. (2016). Roles for retrotransposon insertions in human
826 disease. *Mobile DNA* 7**,** 9.

827 Hillis, D.M., and Green, D.M. (1990). Evolutionary changes of heterogametic sex in the
828 phylogenetic history of amphibians. *Journal of Evolutionary Biology* 3**,** 49-64.

829 Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville,
830 H. (2014). PASTEC: an automatic transposable element classification tool. *PloS One*
831 9**,** e91929.

832 Hof, A.E., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby,
833 A.C., and Saccheri, I.J. (2016). The industrial melanism mutation in British peppered
834 moths is a transposable element. *Nature* 534**,** 102-105.

835 Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., Van Den Elst, H.,
836 Filippov, D.V., Blaser, H., Raz, E., Moens, C.B., Plasterk, R.H.A., Hannon, G.J.,
837 Draper, B.W., and Ketting, R.F. (2007). A role for Piwi and piRNAs in germ gell
838 maintenance and transposon silencing in zebrafish. *Cell* 129**,** 69-82.

839 Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute
840 to the evolution of gene regulatory networks. *Nature* 543**,** 550-554.

841 Itgen, M.W., Siegel, D.S., Sessions, S.K., and Mueller, R.L. (2022). Genome size drives
842 morphological evolution in organ-specific ways. *Evolution* in press.

843 Iwakawa, H.-O., and Tomari, Y. (2022). Life of RISC: Formation, action, and degradation of
844 RNA-induced silencing complex. *Molecular Cell* 82**,** 30-43.

845 Iwasaki, Y.W., Siomi, M.C., and Siomi, H. (2015). PIWI-Interacting RNA: its biogenesis and
846 functions. *Annual Review of Biochemistry* 84**,** 405-433.

847 Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C.,
848 Wei, X., Chin, C.-S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J.,
849 Schneider, K.L., Wolfgruber, T.K., May, M.R., Springer, N.M., Antoniou, E.,

Mccombie, W.R., Presting, G.G., Mcmullen, M., Ross-Ibarra, J., Dawe, R.K., Hastie, A., Rank, D.R., and Ware, D. (2017). Improved maize reference genome with single-molecule technologies. *Nature* 546**,** 524-527.

Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences* 114**,** E1460-E1469.

Kelleher, E.S., and Barbash, D.A. (2013). Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. *Molecular Biology and Evolution* 30**,** 1816-1829.

Kneitz, S., Mishra, R.R., Chalopin, D., Postlethwait, J., Warren, W.C., Walter, R.B., and Schartl, M. (2016). Germ cell and tumor associated piRNAs in the medaka and *Xiphophorus* melanoma models. *BMC Genomics* 17**,** 1-13.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res* 47**,** D155-D162.

Lamichhaney, S., Catullo, R., Keogh, J.S., Clulow, S., Edwards, S.V., and Ezaz, T. (2021). A bird-like genome from a frog: Mechanisms of genome size reduction in the ornate burrowing frog, *Platyplectrum ornatum*. *Proc Nat Acad Sci* 118.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10**,** R25.

Li, W.Z., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22**,** 1658-1659.

Liedtke, H.C., Gower, D.J., Wilkinson, M., and Gomez-Mestre, I. (2018). Macroevolutionary shift in the size of amphibian genomes and the role of life history and climate. *Nat Ecol Evol* 2**,** 1792-1799.

Liu, X., Majid, M., Yuan, H., Chang, H., Zhao, L., Nie, Y., He, L., Liu, X., He, X., and Huang, Y. (2022). Transposable element expansion and low-level piRNA silencing in grasshoppers may cause genome gigantism. *BMC **B**iology* 20**,** 1-16.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15**,** 550.

Luo, S., Zhang, H., Duan, Y., Yao, X., Clark, A.G., and Lu, J. (2020). The evolutionary arms race between transposable elements and piRNAs in *Drosophila melanogaster*. *BMC Evol Biol* 20**,** 14.

Madison-Villar, M.J., Sun, C., Lau, N., Settles, M., and Mueller, R.L. (2016). Small RNAs from a big genome: the piRNA pathway and transposable elements in the salamander species *Desmognathus fuscus*. *J Mol Evol* 83**,** 126-136.

Malmstrøm, M., Britz, R., Matschiner, M., Tørresen, O.K., Hadiaty, R.K., Yaakob, N., Tan, H.H., Jakobsen, K.S., Salzburger, W., and Rüber, L. (2018). The most developmentally truncated fishes show extensive Hox gene loss and miniaturized genomes. *Genome Biology and Evolution* 10**,** 1088-1103.

889 Matsuda, M., Nagahama, Y., Shinomiya, A., Sato, T., Matsuda, C., Kobayashi, T., Morrey,
890 C.E., Shibata, N., Asakawa, S., and Shimizu, N. (2002). DMY is a Y-specific DM-
891 domain gene required for male development in the medaka fish. *Nature* 417**,** 559-563.

892 Medstrand, P., Van De Lagemaat, L.N., and Mager, D.L. (2002). Retroelement distributions
893 in the human genome: variations associated with age and proximity to genes. *Genome*
894 *Research* 12**,** 1483-1495.

895 Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J.M., Irisarri, I., Wong, W.Y.,
896 Nowoshilow, S., Kneitz, S., and Kawaguchi, A. (2021). Giant lungfish genome
897 elucidates the conquest of land by vertebrates. *Nature* 590**,** 284-289.

898 Mueller, R.L. (2017). piRNAs and evolutionary trajectories in genome size and content.
899 *Journal of Molecular Evolution* 85**,** 169-171.

900 Nam, K., and Ellegren, H. (2012). Recombination drives vertebrate genome contraction.
901 *PLoS Genetics* 8**,** e1002680.

902 Orgel, L.E., and Crick, F.H. (1980). Selfish DNA: the ultimate parasite. *Nature* 284**,** 604.

903 Ozata, D.M., Gainetdinov, I., Zoch, A., O'carroll, D., and Zamore, P.D. (2019). PIWI-
904 interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics* 20**,** 89-
905 108.

906 Parhad, S.S., and Theurkauf, W.E. (2019). Rapid evolution and conserved function of the
907 piRNA pathway. *Roy Soc Open Biol* 9**,** 180181.

908 Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides
909 fast and bias-aware quantification of transcript expression. *Nat Methods* 14**,** 417-419.

910 Perkins, R.D., Gamboa, J.R., Jonika, M.M., Lo, J., Shum, A., Adams, R.H., and Blackmon,
911 H. (2019). A database of amphibian karyotypes. *Chromosome Research* 27**,** 313-319.

912 Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families
913 in large genomes. *Bioinformatics* 21 Suppl 1**,** i351-358.

914 Reuter, M., Berninger, P., Chuma, S., Shah, H., Hosokawa, M., Funaya, C., Antony, C.,
915 Sachidanandam, R., and Pillai, R.S. (2011). Miwi catalysis is required for piRNA
916 amplification-independent LINE1 transposon silencing. *Nature* 480**,** 264-267.

917 Rho, M., and Tang, H. (2009). MGEScan-non-LTR: computational identification and
918 classification of autonomous non-LTR retrotransposons in eukaryotic genomes.
919 *Nucleic Acids Res* 37**,** e143.

920 Rodriguez, F., and Arkhipova, I.R. (2018). Transposable elements and polyploid evolution in
921 animals. *Current Opinion in Genetics & Development* 49**,** 115-123.

922 Rogers, R.L., Zhou, L., Chu, C., Márquez, R., Corl, A., Linderoth, T., Freeborn, L.,
923 Macmanes, M.D., Xiong, Z., and Zheng, J. (2018). Genomic takeover by transposable
924 elements in the strawberry poison frog. *Molecular Biology and Evolution* 35**,** 2913-
925 2927.

926 Said, I., Mcgurk, M.P., Clark, A.G., and Barbash, D.A. (2022). Patterns of piRNA regulation
927 in *Drosophila* revealed through transposable element clade inference. *Molecular*
928 *Biology and Evolution* 39**,** msab336.

929 Saint-Leandre, B., Capy, P., Hua-Van, A., and Filée, J. (2020). piRNA and transposon
930 dynamics in *Drosophila*: A female story. *Genome Biology and Evolution* 12**,** 931-947.

931 Senft, A.D., and Macfarlan, T.S. (2021). Transposable elements shape the evolution of
932 mammalian development. *Nature Reviews Genetics* 22**,** 691-711.

933 Shannon, C.E. (1948). A mathematical theory of communication. *Bell Syst Tech J* 27**,** 379-
934 423.

935 Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for
936 FASTA/Q file manipulation. *PloS One* 11**,** e0163962.

937 Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015).
938 BUSCO: assessing genome assembly and annotation completeness with single-copy
939 orthologs. *Bioinformatics* 31**,** 3210-3212.

940 Simpson, E.H. (1949). Measurement of Diversity. *Nature* 163**,** 688-688.

941 Sun, C., and Mueller, R.L. (2014). Hellbender genome sequences shed light on genomic
942 expansion at the base of crown salamanders. *Genome Biol Evol* 6**,** 1818-1829.

943 Sun, C., Shepard, D.B., Chong, R.A., Lopez Arriaza, J., Hall, K., Castoe, T.A., Feschotte, C.,
944 Pollock, D.D., and Mueller, R.L. (2012). LTR retrotransposons contribute to genomic
945 gigantism in plethodontid salamanders. *Genome Biol Evol* 4**,** 168-183.

946 Teefy, B.B., Siebert, S., Cazet, J.F., Lin, H., and Juliano, C.E. (2020). PIWI-piRNA pathway-
947 mediated transposable element repression in Hydra somatic stem cells. *RNA* 26**,** 550-
948 563.

949 Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger
950 genes. *Genome Research* 21**,** 1800-1812.

951 Wang, J., Itgen, M.W., Wang, H., Gong, Y., Jiang, J., Li, J., Sun, C., Sessions, S.K., and
952 Mueller, R.L. (2021a). Gigantic genomes provide empirical tests of transposable
953 element dynamics models. *Genom Proteom Bioinform* 19**,** 123-139.

954 Wang, K., Wang, J., Zhu, C., Yang, L., Ren, Y., Ruan, J., Fan, G., Hu, J., Xu, W., and Bi, X.
955 (2021b). African lungfish genome sheds light on the vertebrate water-to-land
956 transition. *Cell* 184**,** 1362-1376. e1318.

957 Wang, X., Ramat, A., Simonelig, M., and Liu, M.F. (2022). Emerging roles and functional
958 mechanisms of PIWI-interacting RNAs. *Nat Rev Mol Cell Biol*.

959 Warmuth, V.M., Weissensteiner, M.H., and Wolf, J.B. (2022). Accumulation and ineffective
960 silencing of transposable elements on an avian W Chromosome. *Genome Research*
961 32**,** 671-681.

Wei, K.H.-C., Gibilisco, L., and Bachtrog, D. (2020). Epigenetic conflict on a degenerating Y chromosome increases mutational burden in Drosophila males. *Nature Communications* 11**,** 1-9.

Wei, K.H.-C., Mai, D., Chatla, K., and Bachtrog, D. (2022). Dynamics and impacts of transposable element proliferation in the *Drosophila nasuta* species group radiation. *Molecular Biology and Evolution* 39**,** msac080.

Wenke, T., Döbel, T., Sörensen, T.R., Junghans, H., Weisshaar, B., and Schmidt, T. (2011). Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *The Plant Cell* 23**,** 3117-3128.

Wheeler, T.J., and Eddy, S.R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29**,** 2487-2489.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., Sanmiguel, P., and Schulman, A.H. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8**,** 973-982.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N., and Wessler, S.R. (2009). Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science* 325**,** 1391-1394.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30**,** 614-620.

Zhang, P., Kang, J.Y., Gou, L.T., Wang, J., Xue, Y., Skogerboe, G., Dai, P., Huang, D.W., Chen, R., Fu, X.D., Liu, M.F., and He, S. (2015). MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. *Cell Res* 25**,** 193-207.

Zhang, S., Pointer, B., and Kelleher, E.S. (2020). Rapid evolution of piRNA-mediated silencing of an invading transposable element was driven by abundant de novo mutations. *Genome Research* 30**,** 566-575.

## 10 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## 11 Figure Legends

**Figure 1** Amplification plots for TE superfamilies and derivatives. All of the amplification plots suggest current activity. Note that the y-axes differ in scale.

**Figure 2** (A) TE transcripts that are differentially expressed between testes and ovaries. Positive fold-change value indicates higher expression in testes. The majority of transcripts show male-biased expression. (B) Expression levels of TE superfamilies represented by >10 TE transcripts in ovaries vs. testes. Expression is higher in males. Each point represents the average expression of a TE transcript across same-sex individuals. * and NS indicate $P < 0.05$ and not significant, respectively. (C) Genomic abundance and gonadal expression level of TE superfamilies are positively correlated in both sexes.

**Figure 3** (A) The length distribution of small RNA molecules. (B) Composition of the first nucleotide (green means U) of small RNAs that could be mapped to TE transcripts in different genders. (C) The 5' to 5' overlapping of putative piRNA (25–30 bp) in gonads of males and females, a typical signature of ping-pong biogenesis. The number above the peak at the $10^{th}$ overlapping is the Z-score.

**Figure 4** The relationship between TE expression and putative piRNAs (25–30 bp) mapped to TEs in both sexes. Gray lines connect male and female datapoints for the same TE superfamily.

**Figure 5** (A) The ratio of the summed expression of piRNA pathway genes, (B) NuRD and associated repressive complex genes, and (C) TRIM28 to the summed expression of miRNA pathway genes in species with diverse genome sizes. The species and their genome sizes (Gb) from left to right are: *Platyplectrum ornatum* (1), *Gallus gallus* (1.3), *Danio rerio* (1.4), *Xenopus tropicalis* (1.7), *Anolis carolinensis* (2.2), *Mus musculus* (2.5), *Geotrypetes seraphini* (3.8), *Rhinatrema bivittatum* (5.3), *Caecilia tentaculate* (5.5), *Pleurodeles waltl* (20), *Ranodon sibiricus* (21), *Ambystoma mexicanum* (32), *Protopterus annectens* (43), *Cynops orientalis* (44), and *Protopterus aethiiopicus* (~130).

**Table 1** Repeat contigs (≥ 100 bp) identified by different methods/software in the PiRATE pipeline (Berthelier et al., 2018).

| TE-mining method | Software | Repeats clustered at 100% identify |
|---|---|---|
| Similarity-based | RepeatMasker | 75,381 (68.6%) |
| | TE-HMMER | 3,108 (2.8%) |
| Structure-based | HelSearch | 1 (0.0%) |
| | LTR harvest | 84 (0.1%) |
| | MGEScan-non-LTR | 0 |
| | MITE hunter | 7 (0.0%) |
| | SINE finder | 48 (0.0%) |
| Repetitiveness-based | TEdenovo | 306 (0.3%) |
| | RepeatScout | 3,671 (3.3%) |
| Repeat-building-based | dnaPipeTE | 24,090 (21.9%) |
| | RepeatModeler | 3,213 (2.9%) |
| In total | | 109,909 (100%) |

1030

1031

1032 **Table 2** Classification of repeat contigs (modified from Wicker 2007) and summary of
1033 repeats detected in the genome.

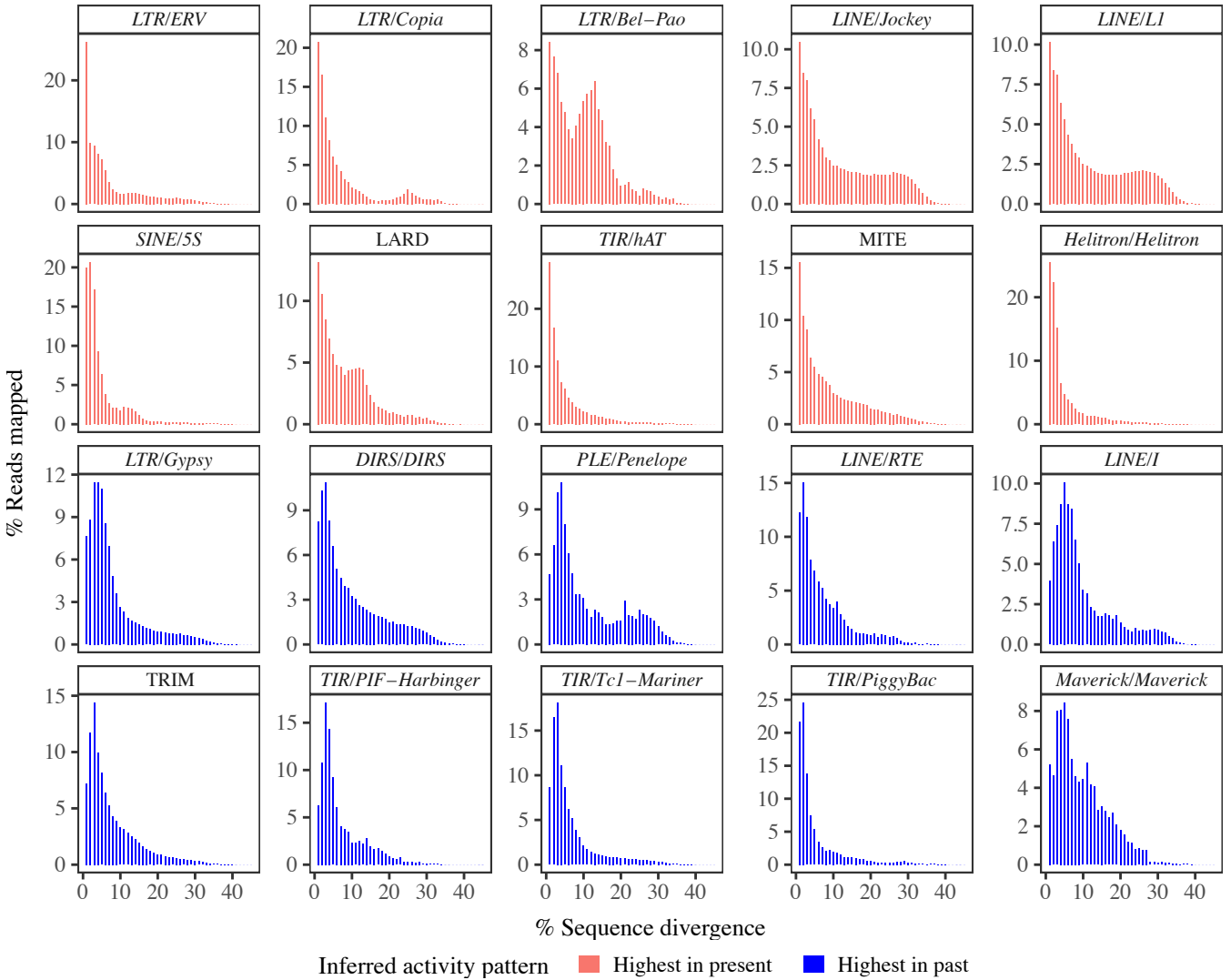
| Order | Superfamily | Percent of Genome[a] | Genomic Contigs (100% Identical) | Genomic Contigs (95% Identical) | Genomic Contigs (80% Identical) | Average Genomic Contig Length (100% identical) (bp) | Longest Genomic Contig (bp) | Transcriptome contigs (80% Identical) | Expression Level in females (TPM) | Expression Level in males (TPM) |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Class I - Retrotransposons - Autonomous** | | | | | | | | | |
| LTR | *Gypsy* | 3.85-6.50 | 5,985 | 5,464 | 3,835 | 548 | 7,587 | 2,057 | 827 | 3,024 |
| | *ERV* | 0.41-0.43 | 413 | 394 | 264 | 593 | 11,567 | 321 | 392 | 694 |
| | *Copia* | 0.10-0.25 | 23 | 21 | 19 | 592 | 1,353 | 27 | 31 | 26 |
| | *Bel-Pao* | 0.05 | 12 | 7 | 2 | 575 | 766 | - | - | - |
| | *Retrovirus* | - | 3 | 2 | 2 | 1,375 | 1,719 | 10 | 1 | 10 |
| | *THE1* | - | 4 | 4 | 3 | 491 | 674 | - | - | - |
| | *Unknown LTR* | - | 4 | 4 | 4 | 469 | 955 | - | - | - |
| DIRS | *DIRS* | 4.44-5.95 | 8,087 | 7,821 | 3,844 | 379 | 4,266 | 4,844 | 5,427 | 11.962 |
| PLE | *Penelope* | 0.09-0.12 | 482 | 462 | 376 | 407 | 3,208 | 460 | 123 | 352 |
| LINE | *Jockey* | 9.69-12.12 | 25,276 | 23,361 | 13,287 | 426 | 3,625 | 11,622 | 6,206 | 15,535 |
| | *L1* | 5.04-6.62 | 10,189 | 8,997 | 5,941 | 672 | 6,546 | 8,241 | 2,954 | 7,610 |
| | *RTE* | 0.12-0.24 | 227 | 201 | 137 | 676 | 4,540 | 399 | 130 | 360 |
| | *I* | 0.09-0.17 | 32 | 25 | 10 | 905 | 3,858 | 48 | 42 | 125 |
| | *R2* | - | | | | | | 1 | - | 1 |
| | *Unknown LINE* | 0.39-1.89 | 90 | 77 | 72 | 1,393 | 5,881 | 1 | - | - |
| | **Class I - Retrotransposons - Non-autonomous** | | | | | | | | | |
| SINE | *5S* | 0.23-0.18 | 57 | 41 | 15 | 773 | 1,783 | 4 | 5 | 1 |
| | *7SL* | - | 43 | 41 | 15 | 243 | 450 | 2 | 1 | - |
| | *tRNA* | - | 22 | 22 | 22 | 273 | 403 | - | - | - |
| | *Unknown SINE* | 0.45-3.42 | 365 | 348 | 338 | 377 | 695 | 219 | 2,901 | 2,471 |
| Retrotransposon Derivatives | TRIM | 3.80-9.76 | 748 | 631 | 465 | 588 | 2,619 | 492 | 1,651 | 5,847 |
| | LARD | 0.15-0.73 | 45 | 37 | 35 | 2,834 | 8,148 | 153 | 515 | 2,643 |
| | **Class II - DNA Transposons - Subclass 1** | | | | | | | | | |
| TIR | *PIF-Harbinger* | 2.98-4.22 | 1,164 | 1,014 | 434 | 411 | 5,169 | 416 | 575 | 1,359 |
| | *hAT* | 1.15-1.12 | 177 | 135 | 55 | 869 | 5,706 | 35 | 83 | 26 |
| | *Tc1-Mariner* | 0.18-0.63 | 73 | 40 | 24 | 1025 | 3,781 | 20 | 58 | 87 |
| | *PiggyBac* | 0.05-0.08 | 9 | 8 | 6 | 1,087 | 1,956 | 6 | 21 | 15 |
| | *MuDR* | - | 3 | 3 | 3 | 266 | 417 | 13 | 4 | 10 |
| | *CACTA* | - | 2 | 2 | 2 | 645 | 846 | 3 | 2 | 1 |
| | *ISL2EU* | - | | | | | | 1 | 18 | 16 |
| | *Ginger* | - | | | | | | 4 | 10 | 6 |
| | *Academ* | - | | | | | | 11 | 5 | 8 |
| | *P* | - | | | | | | 1 | 1 | 1 |

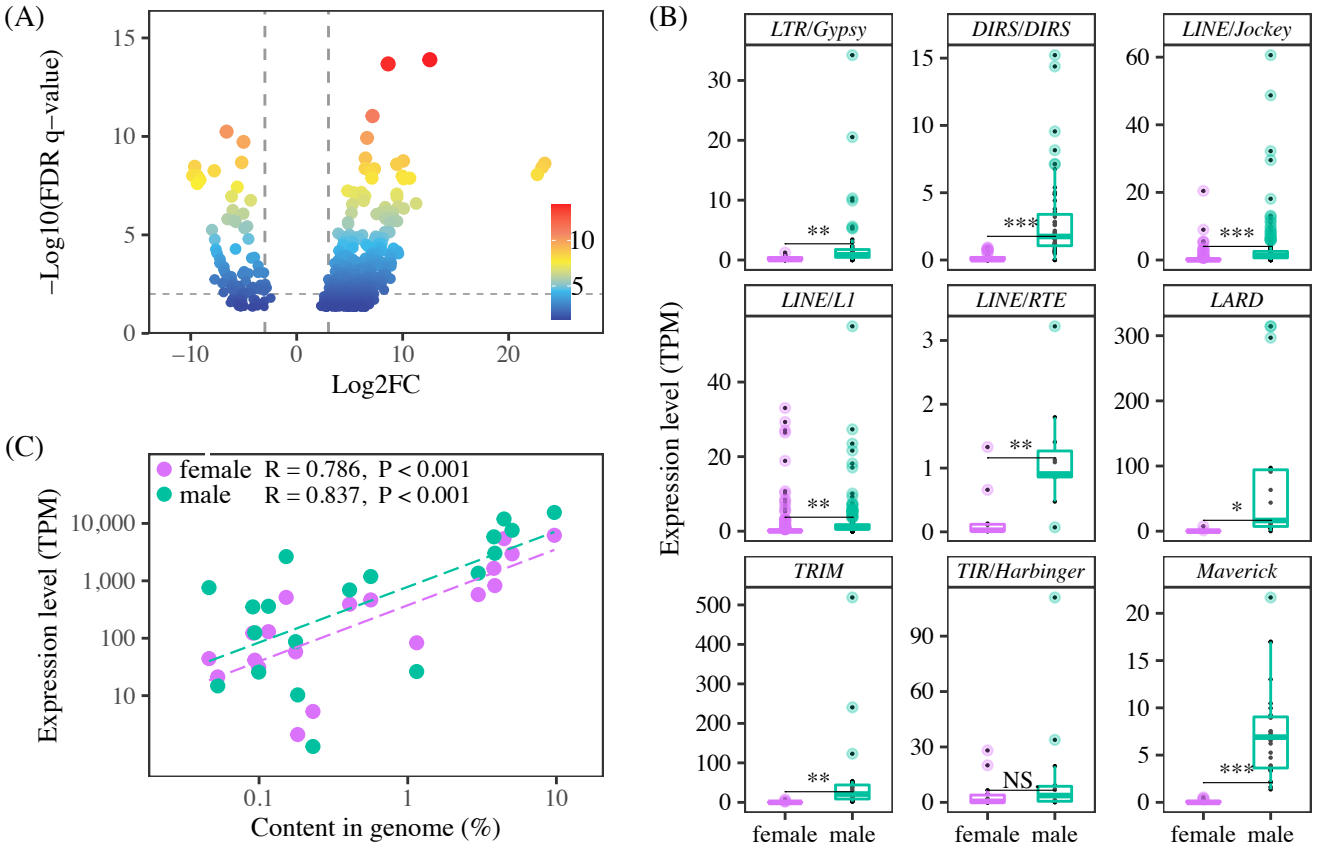
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Unknown TIR* | 0.33-1.46 | 22 | 20 | 19 | 754 | 2,174 | - | - | - |
| Transposon Derivatives | MITE | 0.56-4.07 | 357 | 346 | 321 | 258 | 1,942 | 137 | 465 | 1,188 |
| | **Class II - DNA Transposons - Subclass 2** | | | | | | | | | |
| Maverick | *Maverick* | 0.05 | 258 | 228 | 65 | 583 | 6,090 | 123 | 44 | 762 |
| Helitron | *Helitron* | 0.18-0.48 | 49 | 38 | 19 | 939 | 6,551 | 13 | 2 | 10 |
| **Total** | | 34.38-60.54 | 54,221 | 49,794 | 29,634 | 489 | 11,567 | 29,684 | 22,494 | 42,200 |

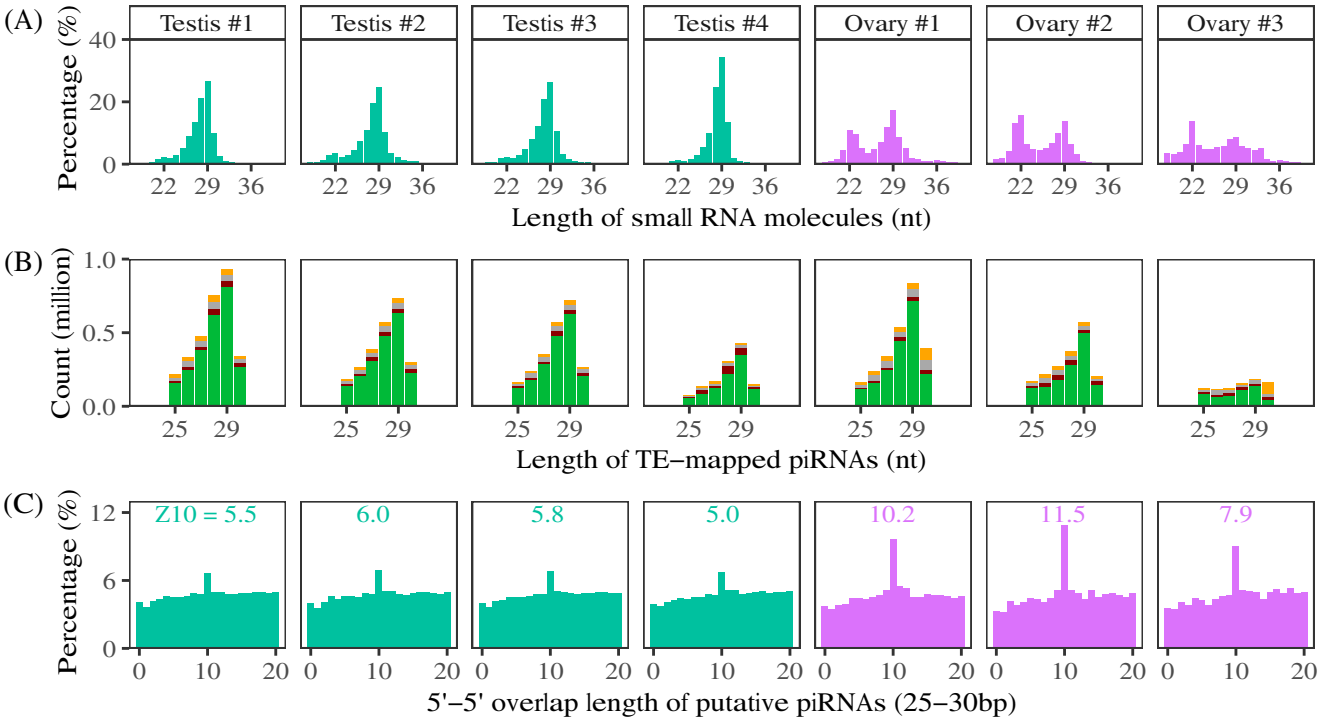
[a]The first and second numbers were estimated including and excluding unknown repeats, respectively, from the repeat library.

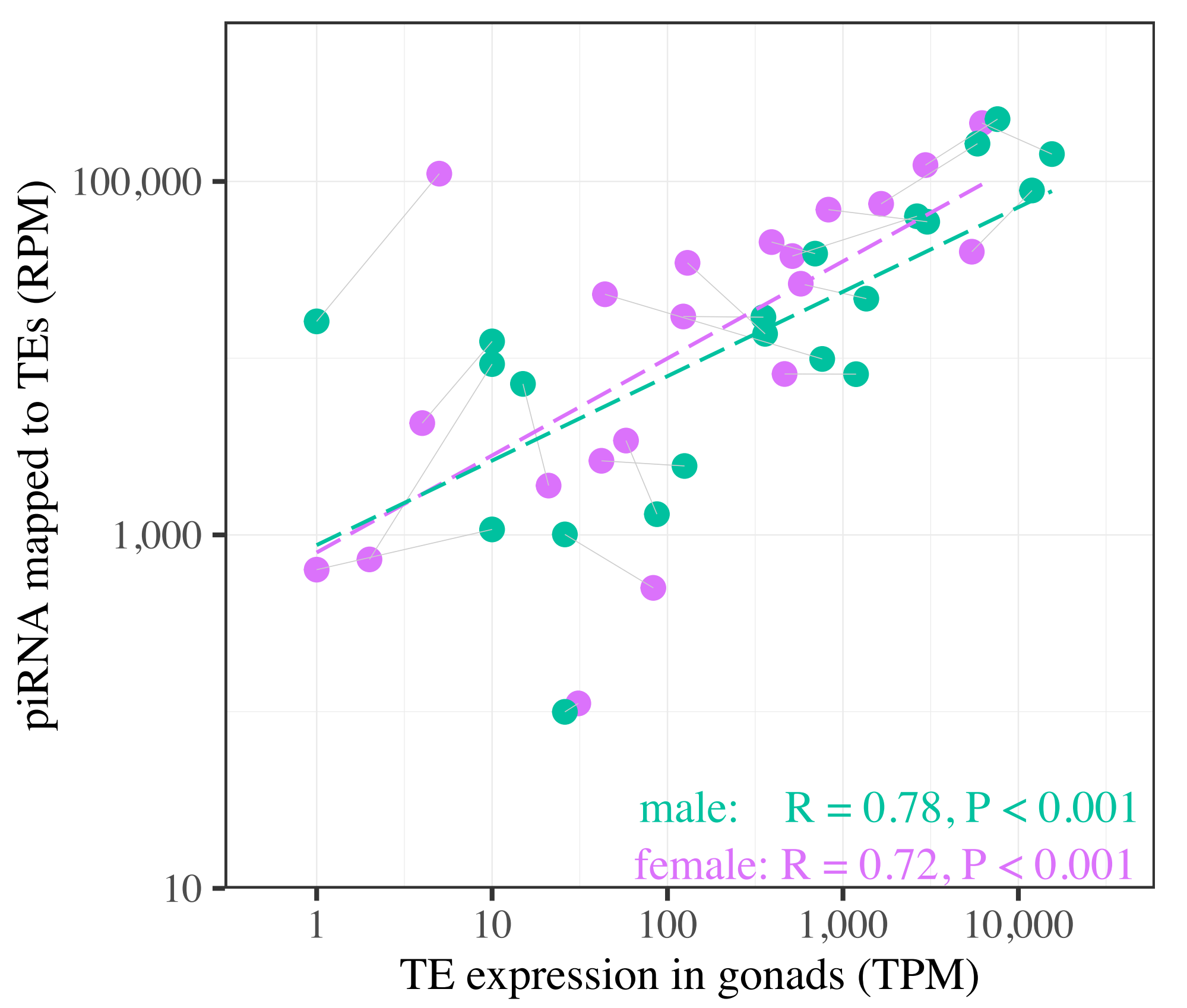
**Table 3** Overall summary of transcriptome annotation and expression in each sex.

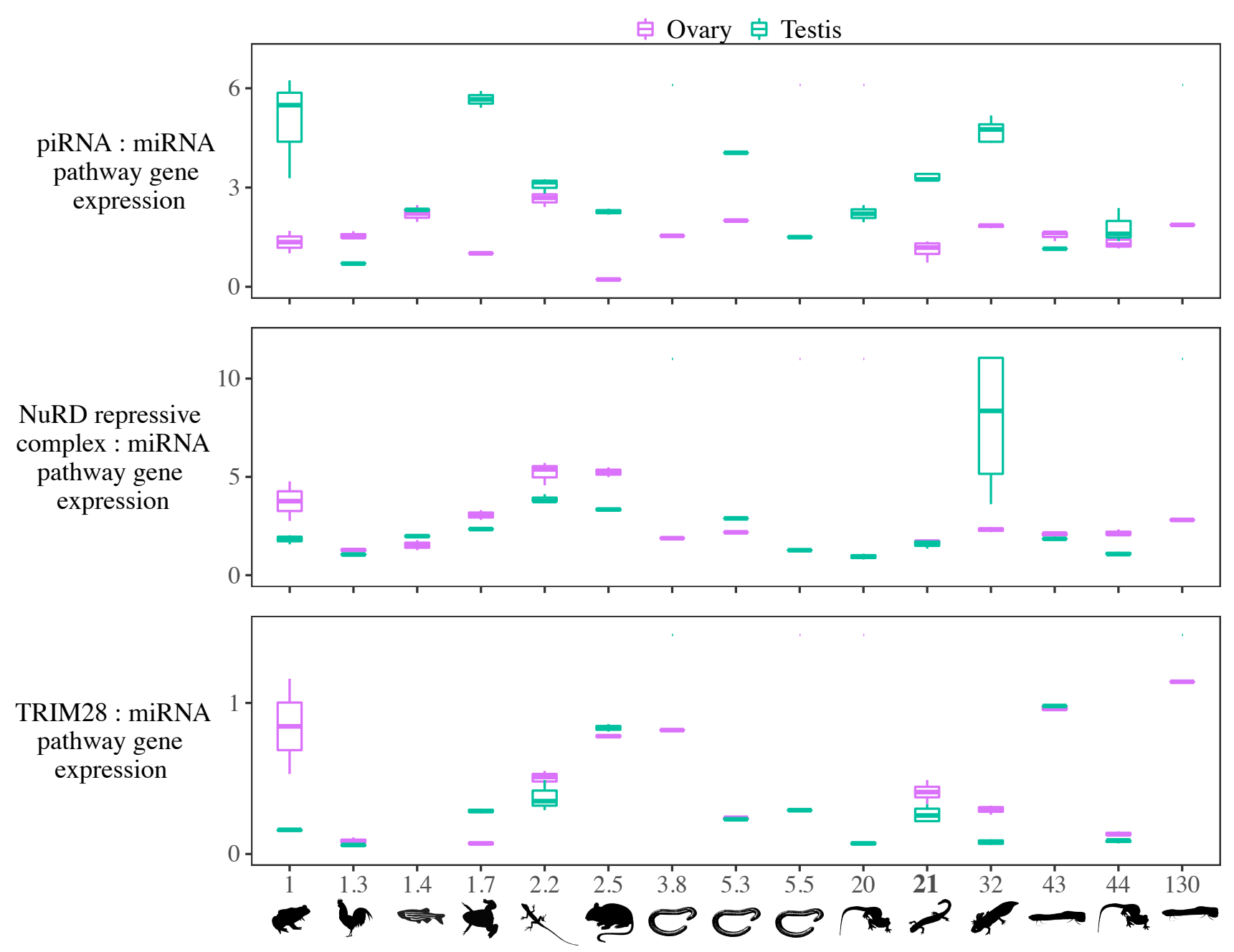
| | No. of contigs expressed in females (percentage of total contigs) | Summed TPM in females (percentage of total expression) | No. of contigs expressed in males (percentage of total contigs) | Summed TPM in males (percentage of total expression) |
|---|---|---|---|---|
| Endogenous gene | 51,647 (17.4%) | 678,366 (67.8%) | 58,041 (13.3%) | 513,783 (51.4%) |
| Autonomous TE | 26,358 (8.9%) | 17,890 (1.8%) | 41,421 (9.5%) | 43,694 (4.4%) |
| Non-autonomous TE | 1,700 (0.6%) | 5,538 (0.6%) | 2,348 (0.5%) | 12,151 (1.2%) |
| Gene/autonomous TE | 722 (0.2%) | 2,618 (0.3%) | 1,001 (0.2%) | 5,748 (0.6%) |
| Gene/nonautonomous TE | 168 (0.1%) | 2,779 (0.3%) | 189 (0.0%) | 1,073 (0.1%) |
| Unannotated | 215,508 (72.8%) | 292,828 (29.3%) | 334,533 (76.5%) | 423,550 (42.4%) |
| Total | 296,103 (100%) | 1,000,029 (100%) | 437,533 (100%) | 999,999 (100%) |


LTR/ERV
LTR/Copia
LTR/Bel−Pao
LINE/Jockey
LINE/L1
SINE/5S
LARD
TIR/hAT
MITE
Helitron/Helitron
LTR/Gypsy
DIRS/DIRS
PLE/Penelope
LINE/RTE
LINE/I
TRIM
TIR/PIF−Harbinger
TIR/Tc1−Mariner
TIR/PiggyBac
Maverick/Maverick
% Reads mapped
% Sequence divergence
Inferred activity pattern
Highest in present
Highest in past


(A)
−Log10(FDR q−value)
Log2FC
(B)
Expression level (TPM)
LTR/Gypsy
DIRS/DIRS
LINE/Jockey
LINE/L1
LINE/RTE
LARD
TRIM
TIR/Harbinger
Maverick
female
male
NS
(C)
female R = 0.786, P < 0.001
male R = 0.837, P < 0.001
Expression level (TPM)
Content in genome (%)


(A)
Testis #1
Testis #2
Testis #3
Testis #4
Ovary #1
Ovary #2
Ovary #3
Percentage (%)
40
20
0
22 29 36
Length of small RNA molecules (nt)
(B)
Count (million)
1.0
0.5
0.0
25 29
Length of TE−mapped piRNAs (nt)
(C)
Z10 = 5.5
6.0
5.8
5.0
10.2
11.5
7.9
Percentage (%)
12
6
0
0 10 20
5'−5' overlap length of putative piRNAs (25−30bp)


piRNA mapped to TEs (RPM)
100,000
1,000
10
1
10
100
1,000
10,000
TE expression in gonads (TPM)
male: R = 0.78, P < 0.001
female: R = 0.72, P < 0.001


Ovary
Testis
piRNA : miRNA pathway gene expression
NuRD repressive complex : miRNA pathway gene expression
TRIM28 : miRNA pathway gene expression
0
3
6
0
5
10
0
1
1
1.3
1.4
1.7
2.2
2.5
3.8
5.3
5.5
20
21
32
43
44
130