Dynamic co-evolution of transposable elements and the piRNA pathway in African cichlid fishes

Miguel Vasconcelos Almeida[1,2,*], Moritz Blumer[3,13], Chengwei Ulrika Yuan[1,2,3,13], Pío Sierra[3], Jonathan L. Price[1,2], Fu Xiang Quah[1,3], Aleksandr Friman[4,5], Alexandra Dallaire[1,2,6], Grégoire Vernaz[2,3,7], Audrey L. K. Putman[1,2,3], Alan M. Smith[8], Domino A. Joyce[8], Falk Butter[9,10], Astrid D. Haase[4], Richard Durbin[3,11], M. Emília Santos[12], Eric A. Miska[1,2,11,*]

[1]Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, CB2 1GA, UK
[2]Wellcome/CRUK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QN, UK
[3]Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK
[4]National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA
[5]Biophysics Graduate Program, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA
[6]Comparative Fungal Biology, Royal Botanic Gardens Kew, Jodrell Laboratory, Richmond TW9 3DS, UK
[7]Present address: Zoological Institute, Department of Environmental Sciences, University of Basel, Vesalgasse 1, Basel, 4051, Switzerland
[8]School of Natural Sciences, University of Hull, Hull, HU6 7RX, UK
[9]Institute of Molecular Biology (IMB), Quantitative Proteomics, Ackermannweg 4, Mainz, 55128, Germany
[10]Institute of Molecular Virology and Cell Biology, Friedrich-Loeffler-Institute, Südufer, Greifswald, 17493, Germany
[11]Wellcome Sanger Institute, Tree of Life, Wellcome Genome Campus, Hinxton, CB10 1SA, UK
[12]Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK
[13]These authors contributed equally.

*Correspondence: mdd34@cam.ac.uk, eam29@cam.ac.uk

Abstract

East African cichlid fishes have diversified in an explosive fashion, but the (epi)genetic basis of the phenotypic diversity of these fishes remains largely unknown. Although transposable elements (TEs) have been associated with phenotypic variation in cichlids, little is known about their transcriptional activity and epigenetic silencing. Here, we describe dynamic patterns of TE expression in African cichlid gonads and during early development. Orthology inference revealed an expansion of *piwil1* genes in Lake Malawi cichlids, likely driven by PiggyBac TEs. The expanded *piwil1* copies have signatures of positive selection and retain amino acid residues essential for catalytic activity. Furthermore, the gonads of African cichlids express a Piwi-interacting RNA (piRNA) pathway that target TEs. We define the genomic sites of piRNA production in African cichlids and find divergence in closely related species, in line with fast evolution of piRNA-producing loci. Our findings suggest dynamic co-evolution of TEs and host silencing pathways in the African cichlid radiations. We propose that this co-evolution has contributed to cichlid genomic diversity.

Introduction

The East African Great Lakes are home to prolific cichlid radiations, the most species-rich and phenotypically diverse adaptive radiations in vertebrates[1,2]. In the last 10 million years, more than 1,700 species of cichlid fishes (Cichlidae family) have evolved in virtually every lacustrine and riverine ecological niche in Lakes Victoria, Tanganyika, Malawi and surrounding bodies of water. The explosive diversification of East African cichlids is particularly striking in the haplochromine tribe and has resulted in astonishing variation in morphologies, colouration, diets, and behaviours[1,2]. The genetic and epigenetic basis for such phenotypic variability is of great interest and remains, by and large, unknown.

Initial genomic studies suggested very low genetic variability amongst East African cichlids[3]. In Lake Malawi cichlids, for example, the reported average single nucleotide polymorphism divergence between species pairs was 0.1-0.25%[3,4]. These low estimates were derived from approaches aligning short-read sequence data to a linear reference genome and generally ignore the contribution of structural variation. We have recently complemented these estimations using a pangenomic approach and long-read genome assemblies of representative Lake Malawi species[5]. With this approach, we estimated that 4.73-9.86% of Lake Malawi cichlid genomes can be attributed to interspecific structural variation[5]. Importantly, transposable elements (TEs) account for up to 74.65% of structural variant sequence. Thus, TEs comprise an underestimated source of genetic variability in East African cichlids.

TEs are diverse mobile genetic elements that inhabit nearly all eukaryotic genomes sequenced to date[6]. While most extant TEs and novel TE mobilisation events are selectively neutral or slightly deleterious to their hosts[7], several examples of TEs providing adaptive benefits to their hosts have been reported[8–10]. The TE landscapes of teleost fish genomes are highly dynamic[11–17], and cichlid genomes are no exception, as they contain varied TE populations with signs of recent transpositional activity[16,18]. TEs may be an important source of (epi)genetic variability that has fuelled the cichlid radiations. Consistent with this notion, presence/absence variation of TEs is associated with pigmentation traits[19,20], sex determination[21], and modulation of endogenous gene expression[18,22]. It has recently been shown that differentially methylated regions enriched in young TEs are associated with transcriptional changes[23], further supporting a role for TEs in modulating gene expression in cichlids. The same study found widespread DNA methylation at TEs, but besides this, little is known about the silencing pathways that direct TE silencing in cichlids and lead to the deposition of DNA methylation.

Several pathways have evolved in animals to silence TEs, particularly in the germline and early development to protect the next generations from deleterious effects of TE activity[8,24–30]. Here, we focus on the Piwi-interacting RNA (piRNA) pathway, a class of non-coding small RNAs (sRNAs) 21-35 ribonucleotides long, which drive silencing of TEs in the animal germline, including in fishes[27,31–33]. piRNAs bind to Piwi Argonaute proteins and guide them to target RNAs with base complementarity, leading to post-transcriptional and/or transcriptional silencing of their targets[26,27]. The latter can be achieved by piRNA-directed DNA methylation of targets. piRNA biogenesis is complex, requires a variety of co-factors, and can be conceptualised as two collaborating pathways that create sequence diverse piRNA populations in the animal germline: the ping-pong and phased biogenesis pathways[26,27,32,34–37]. These pathways depend mainly on the slicer activity of Piwi proteins, and endonucleolytic activity of Zucchini/PLD6 acting on long piRNA precursor transcripts.

The co-evolution of TE silencing factors and TEs is often thought to occur in the form of an arms race. TE silencing factors, including those of the piRNA pathway, often have signatures of fast, adaptive evolution that are interpreted as a consequence of such an arms race[24,25,38–41]. These signatures include positive selection and lability in terms of copy number variation, with recurrent gene duplications and turnover. Little is known about the co-evolution of TE silencing pathways and TEs in East African cichlids and whether these arms races could help fuel cichlid radiations.

Here, we describe dynamic TE expression in the gonads and early development of African cichlids. We identify cichlid orthologs of known factors required for TE silencing in vertebrates and discover an expanded repertoire of *piwil1* genes in Lake Malawi cichlids, which may have been driven by PiggyBac TEs. The additional *piwil1* paralogs retain amino acid residues required for the catalytic activity of the PIWI domain and have signatures of adaptive evolution, suggesting acquisition of novel regulatory functions. TE silencing factors are expressed in cichlid gonads, alongside an abundant piRNA population with signatures consistent with active piRNA-driven TE silencing. Lastly, we observe divergence in the genomic origins of piRNA production in closely related Lake Malawi cichlids.

Results

TE transcriptional activity in cichlid gonads and early development

To profile TE expression in African cichlids, we sequenced mRNAs of representative species of haplochromine cichlids from each of the major East African Great Lakes (**Figure 1A**). We chose *Pundamilia nyererei* (PN) as a representative for Lake Victoria, *Astatotilapia burtoni* (AB) for Lake Tanganyika, and *Astatotilapia calliptera* (AC) for Lake Malawi. To compare closely related species within the same Lake, we included two species from Lake Malawi, alongside AC: *Maylandia zebra* (MZ), and *Tropheops* sp. 'mauve' (TM). In addition, we included *Oreochromis niloticus* (ON, commonly known as Nile tilapia) as an outgroup. ON is a representative of the tilapine tribe that has a broad geographical distribution in Africa and is not as phenotypically diverse as haplochromines[42]. We profiled TE expression in cichlid gonads, as these contain the germline, where the arms race between TEs and their silencing factors is most apparent in other animals[8,27]. For a comprehensive analysis of younger TE populations in Lake Malawi, we created an additional curated TE annotation for AC, which we used throughout this work alongside the uncurated annotation (**Figure S1A**, see **Methods**).

We found that 515-746 (86-93%) cichlid TE families are expressed in gonads (**Figure S1B**). Two trends are recognisable when considering the expression of TE families grouped by class. First, long terminal repeat (LTR) families have the highest median expression (**Figure S1C**). This trend is reversed when TE expression is quantified based on the curated TE annotation of AC, which has more annotated LTR families (**Figure S1B**) and where LTR annotations were improved, including both the long-terminal repeats and intervening genes. This suggests that uncurated LTR annotation may lead to an overestimation of LTR expression. Second, TE families of the same class tend to be more highly expressed in testes rather than ovary, revealing differences in TE expression between sexes (**Figure S1C**). Higher median expression of annotated protein-coding genes was also observed in cichlid testes (**Figure S1D**), suggesting the sex-specific differences in TE expression may follow general sex-specific differences in transcriptional output.

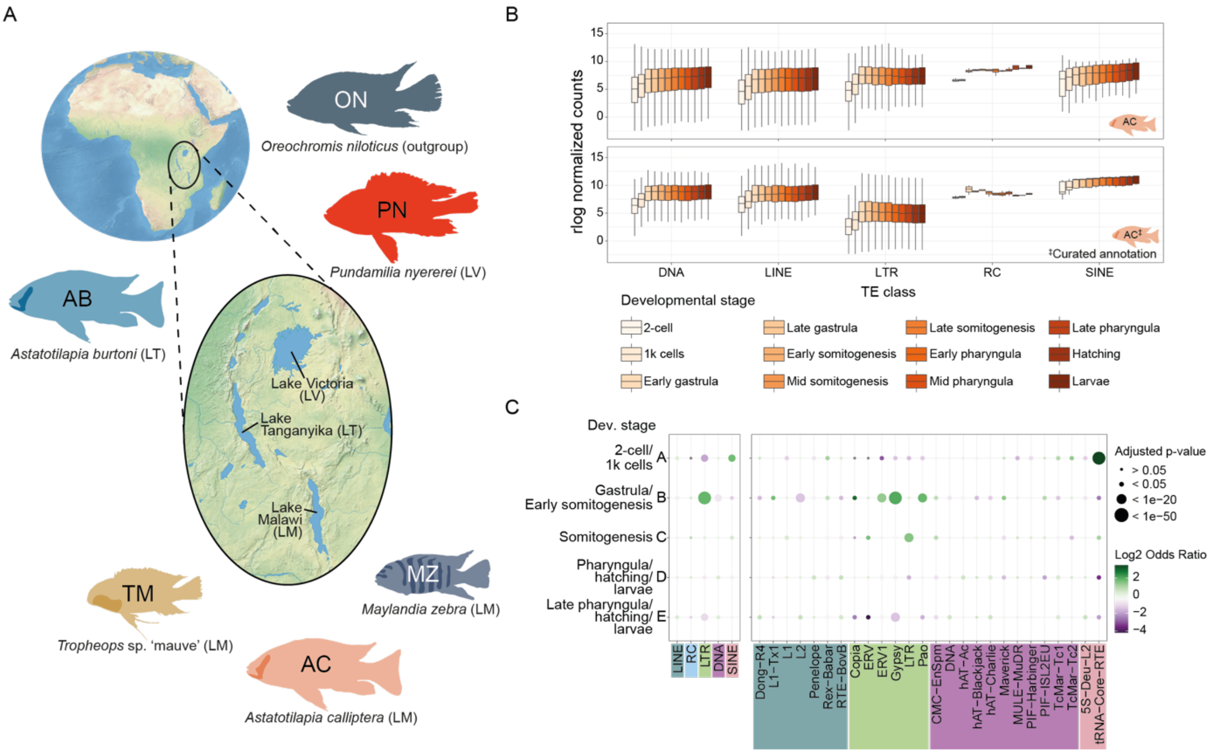


Figure 1. Dynamic patterns of TE expression during cichlid early development. (**A**) The East African Great Lakes and surrounding bodies of water, along with the species used in this study, each representative of a major lake. *Oreochromis niloticus* (Nile tilapia) is used as an outgroup to the radiations of the Great Lakes. For Lake Malawi, we use three species to address within-lake dynamics of TE expression and epigenetic silencing. *Astatotilapia calliptera* is a generalist omnivore, which inhabits shallow water environments in the lake and surrounding rivers and streams[4,42], while *Maylandia zebra* and *Tropheops* sp. 'mauve' are Mbuna rock-dwelling cichlids specialised in eating algae[4,42].; (**B**) Expression of TE families belonging to major TE classes throughout early development of *A. calliptera*, displayed as rlog normalised counts. TE Expression was calculated using the default (panel above) and curated annotations (panel below). (**C**) Enrichment of TE classes and superfamilies in particular developmental stages, according to clusters A-E of differentially expressed TEs as defined in Figure S1G. Only TE superfamilies significantly enriched/depleted in at least one developmental stage are depicted. Grey dots represent lack of significant enrichment. Analysis done as in Chang et al., 2022[13], using the curated TE annotation of AC. AB, *Astatotilapia burtoni*; AC, *Astatotilapia calliptera*; LM, Lake Malawi; LT, Lake Tanganyika; LV, Lake Victoria; MZ, *Maylandia zebra*; ON, *Oreochromis niloticus*; PN, *Pundamilia nyererei*; rlog, regularised log; TM, *Tropheops* sp. 'mauve'.

Embryogenesis and early development are periods known to display signs of TE transcriptional activity[8,10,13]. We therefore conducted bulk mRNA sequencing in early developmental stages of Lake Malawi cichlids and found that 91-94% of cichlid TE families are expressed during early development (**Figure S1E**). Expression in these developmental stages is overall identical between AC and TM (**Figure 1B** and **S1F**). The temporal expression pattern of all TE classes is similar: lower expression before gastrulation rising to peak or near peak expression at early gastrula followed by relatively constant levels of transcriptional activity. Analysing TE expression at the locus level reveals the overall expression pattern at the family level is not universal, as several individual TEs have expression patterns specific to distinct developmental stages (**Figure S1G**). Interestingly, we find a major enrichment of ERV1, Gypsy and Pao LTRs in gastrula stage and early somitogenesis (**Figure 1C**, cluster B), and SINE (short interspersed nuclear element) enrichment at the earliest stages (**Figure 1C**, cluster A). Overall, these results support substantial transcriptional TE activity in gonads and during early development of African cichlids.

An expanded repertoire of *piwil1* genes in Lake Malawi cichlids

Given the dynamic TE expression patterns observed, we reasoned that active silencing pathways must be in place in cichlids to counteract TE activity. First, we identified orthologs of sRNA-based TE silencing factors in cichlids (**Supplemental Table 1**)[8]. With three exceptions, all genes are present in cichlid genomes (**Figure S2A** and **Supplemental Table 1**).

Then, we addressed whether these factors are expressed in the germline by performing quantitative proteomics on gonads of representative cichlid species. TE silencing factors are detected most prominently in testes (**Figure S2B**). Abundant yolk proteins, from the substantial yolk fraction of cichlid eggs[43], precluded protein detection in ovary samples at a depth similar to other organs (**Figure S2C**). Despite the influence of the yolk, Piwil1, a core piRNA pathway factor was detected in the ovaries of all species (**Figure S2B**). Somatic roles for the piRNA pathway have been increasingly recognized in animals, including in brain and nervous system[44]. We also profiled the proteome of brain tissues of the representative cichlid species, but obtained no consistent evidence supporting expression of core piRNA factors in the brain of all cichlid species (**Figure S2B**). These results point to strong conservation of germline-expressed TE silencing factors in African cichlids.

While inspecting TE silencing factor orthologs, we detected multiple copies of *piwil1* genes, homologs of zebrafish *ziwi*[33], in cichlids representative of Lake Malawi, but not in representatives of Lakes Tanganyika and Victoria (**Figure 2A-B**, **Supplemental Table 1**. While fishes generally have one *piwil1* copy, AC has four *piwil1* copies, which we named *piwil1.1-1.4*. Two of these are full-length copies, whereas the other two are truncations containing only the PIWI domain (**Figure 2A-B**). *piwil1.1* of AC is located in the conserved syntenic context of vertebrate *piwil1* genes (**Figure S3A**), indicating that *piwil1.1* is the ancestral cichlid *piwil1* gene. By aligning all additional *piwil1* copies of AC to the coding sequence of *piwil1.1* and projecting the coding sequence to the aligned paralogs, we observe that the full-length paralog *piwil1.2* likely contains stop codons that are bypassed in existing gene annotations produced by automated annotation pipelines (**Figure 2A**). Also, *piwil1.2* is expressed at negligible levels in cichlid gonads and brain (**Figure S3B**) and is therefore likely a pseudogene.

piwil1.2, *piwil1.3*, and *piwil1.4* reside in genomic regions rich in TEs (**Figure 2A**). The 3' regions of *piwil1.2, piwil1.3* and *piwil1.4* share a PiggyBac TE insertion (**Figures 2A** and **S3C**). PiggyBac is a DNA TE family known to be very proficient at carrying large DNA segments upon transposition, a quality that has promoted its use in genome engineering[45,46]. Autonomous PiggyBac TEs consist of two terminal inverted repeats (TIRs) flanking a transposase gene[47]. Like other DNA TEs with TIRs, PiggyBacs mobilise when two transposase proteins each bind to one of the TIRs[6]. The *piwil1*-associated PiggyBacs have mutations that preclude production of a functional transposase (**Figure S3C**). These *piwil1*-associated PiggyBac belong to the same TE family (PiggyBac-1), of which we identified 377 high quality copies in the AC reference genome. Considering the genome size of >880 Megabase, one PiggyBac-1 element is expected, on average, every 2.3 Megabase. A phylogeny of all high-confidence PiggyBac-1 TE fragments in the AC genome shows that the three *piwil1*-associated PiggyBac TEs are closely related, particularly the PiggyBacs associated with *piwil1.3* and *piwil1.4* (**Figure S3D**). Finding all three *piwil1* paralogs on different chromosomes with closely related flanking PiggyBac-1 insertions either directly 3' adjacent (*piwil1.2* and *piwil1.4*) or 7 kb downstream (*piwil1.3*) is therefore highly unlikely to be coincidental.

Given the presence of related PiggyBac-1 TEs associated with all three *piwil1* paralogs, we reasoned that the initial expansion of *piwil1* genes in Lake Malawi cichlids was likely driven by transposition of PiggyBac-1, either at a time when its transposase was active, or in a non-autonomous fashion using the transposase of other PiggyBacs. This could have happened if a piggyBac transposase used one of its own TIRs together with an alternative TIR-like sequence from the *piwil1* locus. To address this, we searched for sequence signatures of PiggyBac mobilisation: the preferred insertion

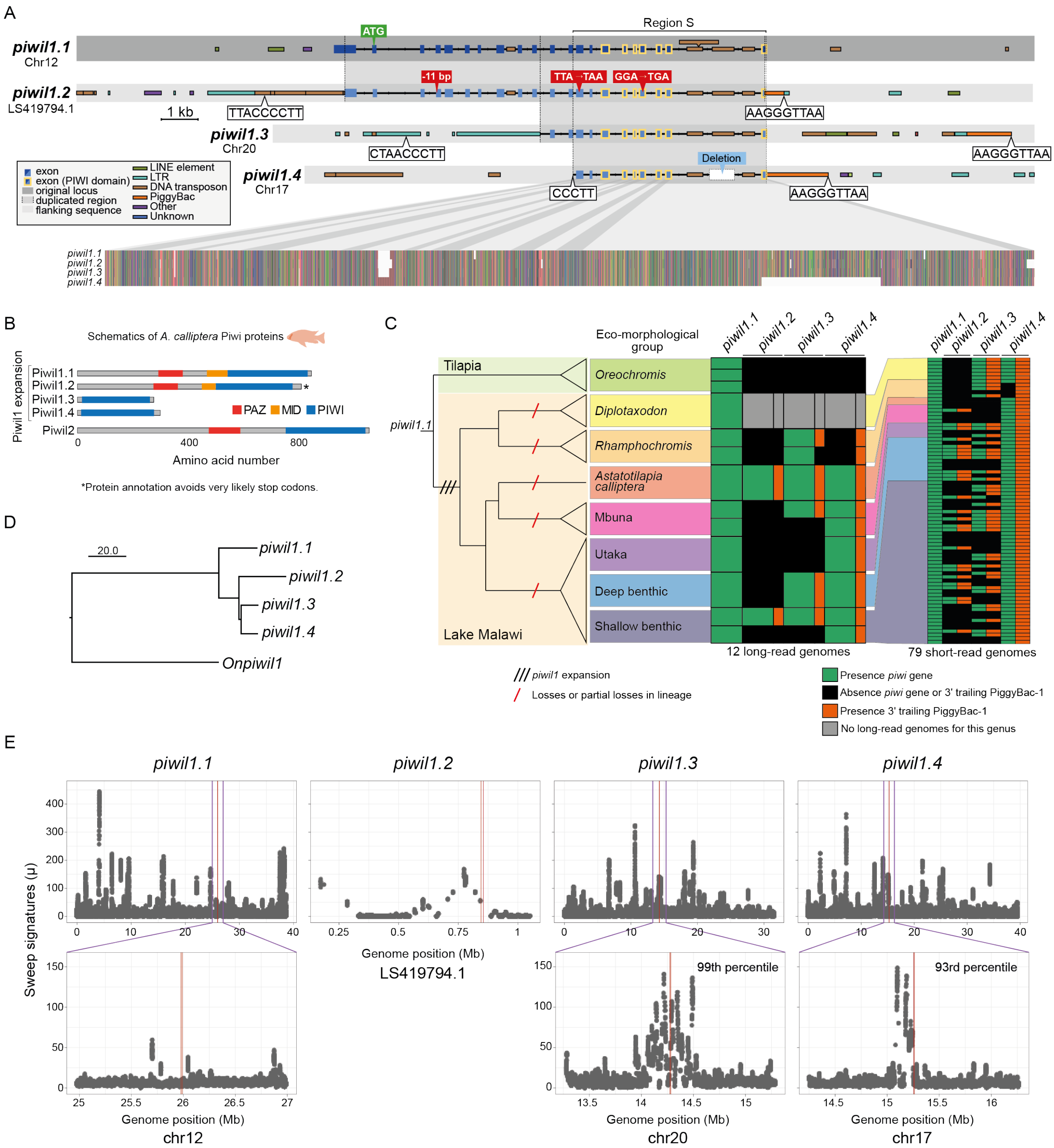


Figure 2. An expansion of *piwil1* paralogs in Lake Malawi cichlids likely mediated by PiggyBac TEs. (**A**) Detailed schematics of the four *piwil1* loci in the *A. calliptera* reference genome. Exons and TEs are shown, along with other relevant sequence features, such as start and stop codons, deletions, etc. The sequences of the putative PiggyBac TIRs (terminal inverted repeats) and preferred insertion sites are shown in white boxes, from 5' to 3'. Of note, the putative TIR and insertion site sequences distal to the PiggyBac are the reverse complement of 5'-CCCTT-3' and 5'-TTAA-3', respectively. The dotted lines represent the borders of duplicated regions, according to multiple sequence alignment. Region S marks the genomic region shared by all *piwil1* genes. The image in the lower portion of the panel is a zoomed-out image of the multiple sequence alignment, color-coded by nucleotide. The putative stop codons were identified manually from an alignment of the genomic regions of all *piwil1* copies with the coding sequence of *piwil1.1*, the *piwil1* gene most conserved in vertebrates. No putative stop codons were found in *piwil1.3* and *piwil1.4*. (**B**) Schematics of the domain structure of the five Piwi proteins annotated in the *A. calliptera* genome, including the expanded Piwil1 protein repertoire. Due to the putative stop codons found in the *piwil1.2* locus, it is likely that the protein is misannotated and that the full-length protein will not be produced. (**C**) Presence (green)/absence (black) of each *piwil1* gene in genomes of Lake Malawi and Tilapia cichlids. Presence of *piwi*-associated PiggyBac TEs is indicated in orange. Presence/absence of *piwil1* genes and PiggyBac TEs was ascertained from long-read sequencing of 12 individuals and short-read sequencing of 79 individuals spanning all the major eco-morphological clades in Lake Malawi. The cladogram of the Malawi radiation reflects the current understanding of the radiation based on genomic studies[4]. The proposed model for *piwil1* gene evolution involves gene expansion early in the Lake Malawi radiation, followed by losses in particular lineages. (**D**) Neighbour-joining tree representing the Hamming distance between the non-coding regions of the *piwil1* genomic sequences of *A. calliptera* along with the genomic sequence of *piwil1* of *O. niloticus* (*Onpiwil1*) as an outgroup. The multiple sequence alignment used to build this tree included the introns shared by all *piwil1* genes (Region S in **Figure 2A**). (**E**) The plots show genome-wide results of Raised Accuracy in Sweep Detection (RAiSD)[84]. μ is a metric incorporating three selective sweep signatures, with higher μ values indicative of a stronger signature of selection. Upper panels show μ across the entire chromosome, or entire scaffold in case of *piwil1.2*. Lower panels are insets of the *piwil1* gene regions +/- 1 Megabase (Mb). As the entire scaffold where *piwil1.2* resides is less than 2 Mb, no inset is shown. We calculated a per-gene μ for all genes (see **Methods**), and with this approach *piwil1.3* and *piwil1.4* are in the 99th and 93rd percentile, respectively, of per-gene μ.

sequence (5'-TTAA-3'), directly preceding the predicted PiggyBac TIR sequence (5'-CCCTT-3')[47,48]. We found potential TIRs adjacent to the PiggyBac-1 elements, and close to the border of the *piwi* duplications distal to the PiggyBac (**Figure 2A**). Putative PiggyBac insertion signatures distal to the PiggyBac-1 element of *piwil1.2* and *piwil1.3* were harder to identify because of additional transposition in that area that could have pushed the PiggyBac sequence signature further upstream from *piwil1* (**Figure 2A**). We could find a 5'-CCCTT-3' sequence upstream of *piwil1.4*, but the downstream 5'-TTAA-3' insertion sequence may have eroded. The consistent association of closely related PiggyBac TEs to *piwil1* paralogs, and the presence of putative TIR sequences flanking the genes are compatible with a model whereby PiggyBac-1 transposition mediated the expansion of *piwil1* genes in Lake Malawi cichlids.

Evolution and functional potential of *piwil1* genes in Lake Malawi cichlids

Next, we assessed the prevalence of each *piwil1* paralog in the major eco-morphological clades of Lake Malawi cichlids. We mapped genomic reads to the AC reference genome (which contains all four *piwil1* copies) and manually assessed the presence or absence of each *piwil1* gene from mapped reads. We used 12 sets of long reads and 79 sets of short reads of Lake Malawi cichlids, corresponding to 80 species (**Supplemental Table 1**). We did not find any of the three extra *piwil1* paralogs in tilapias, which form an outgroup to the haplochromine radiations (**Figure 2C**). However, we find each additional *piwil1* paralog in all major eco-morphological clades within the Lake Malawi radiation (**Figure 2C**). *piwil1.1* and *piwil1.4* are most widespread, with *piwil1.1* identified in all individuals and *piwil1.4* found in 82/88 individuals (exceptions are 6/7 individuals of the *Rhamphochromis* genus, **Figure 2C** and **Supplemental Table 1**). Conversely, *piwil1.2* and *piwil1.3* have a patchier distribution (27/88 and 46/88 individuals). We found support for a 3' trailing PiggyBac-1 TE in the vast majority of *piwil1.2*, *piwil1.3*, and *piwil1.4* copies (153/155, **Figure 2C** and **Supplemental Table 1**). In 8 individuals we found support for a 3' trailing PiggyBac-1 TE in their expected location 3' of *piwil1.3* and *piwil1.4*, but found no support for the *piwil1* gene itself (**Figure 2C** and **Supplemental Table 1**). This observation may reflect rare events of *piwil1* gene elimination by recombination processes. Alternatively, these individuals could be heterozygous for the presence/absence of *piwil1.3* or *piwil1.4*, leading to fewer supporting genomic reads.

Inspection of alignments of all AC *piwil1* paralogs revealed that *piwil1.2*, *piwil1.3* and *piwil1.4* all share variation that is not shared with *piwil1.1* (**Figure 2A**). Moreover, *piwil1.3* and *piwil1.4* share the most variation. This, together with the relatedness of the *piwil1*-associated PiggyBac-1 elements (**Figure S3D**), suggests that *piwil1.2* was the first paralog to duplicate via transposition and that *piwil1.3* and *piwil1.4* originated from *piwil1.2*. A tree representing the distance between the non-coding regions shared by all four *piwil1* genes of AC (within region S in **Figure 2A**) and *piwil1* of ON (as an outgroup) support this hypothesis (**Figure 2D**). A similar tree created from the exons shared by these same *piwil1* genes (within region S in **Figure 2A**) did not produce a tree topology congruent with the non-coding tree (compare **Figures 2D** with **S3E**). We suggest that this discrepancy could reflect selective processes acting on the coding sequences of *piwil1* genes.

Following gene duplication, paralogs can undergo a number of evolutionary routes, including towards sub- or neofunctionalisation[49], with distinct signatures of selection. To learn about the selective pressures at play, we tested for the presence of signatures of selective sweeps in 79 Lake Malawi cichlid genomes (**Supplemental Table 1**). While the genomic region of *piwil1.1* does not display a clear signature of selective sweep (**Figure 2E**, left panels), *piwil1.3* and *piwil1.4* are in the 99th and 93rd

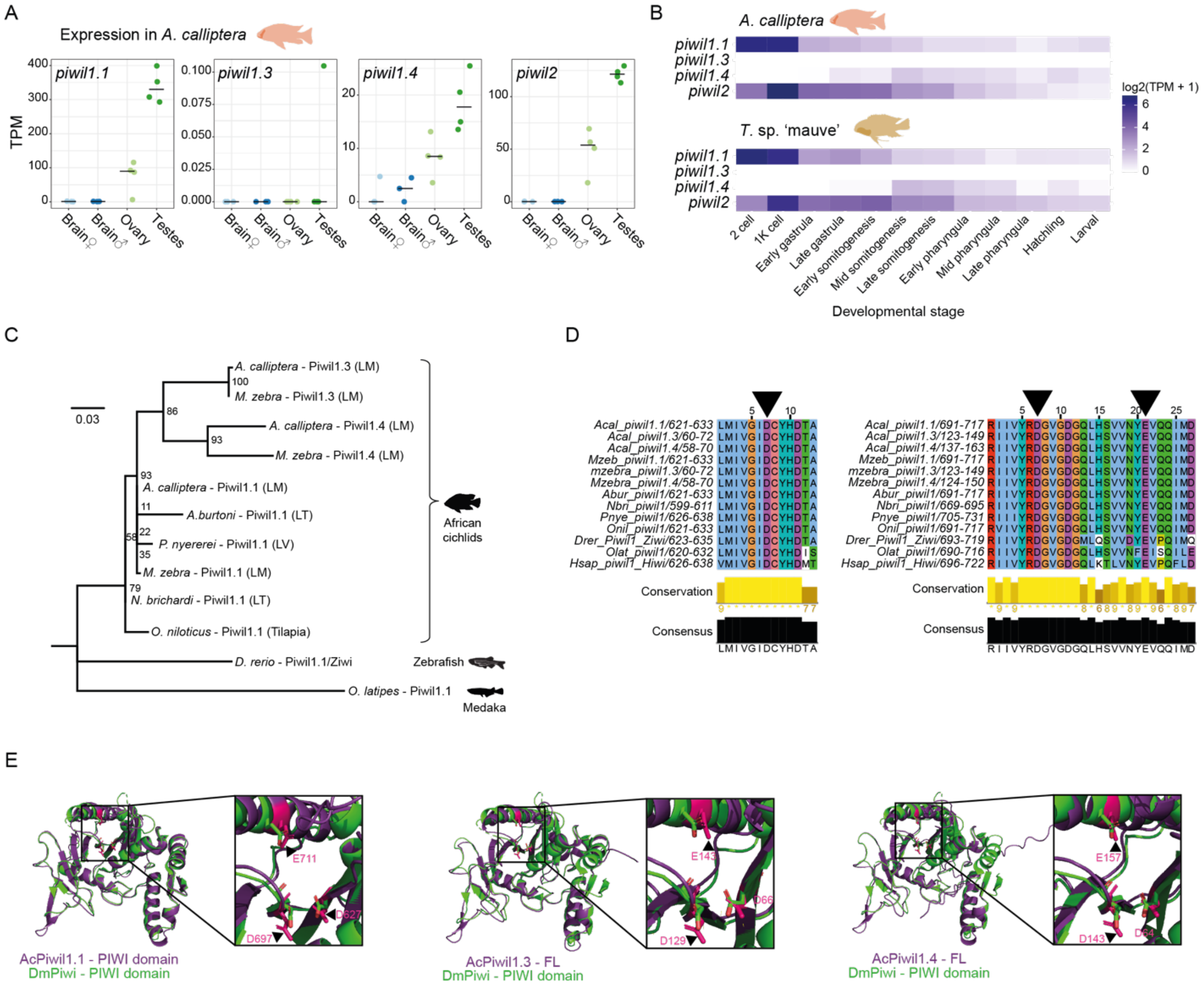


Figure 3. Expression and functional potential of Piwil1 proteins in Lake Malawi. (**A**) Expression, in Transcripts per Million (TPM), of *piwil1* paralogs and *piwil2* in gonads and brain of *A. calliptera*. (**B**) Expression of the three *piwil1* genes and *piwil2* throughout early development of *A. calliptera* and *Tropheops* sp. 'mauve', another Lake Malawi cichlid. (**C**) Phylogenetic tree constructed from an alignment of the PIWI domain of Piwil1 proteins of African cichlids, using zebrafish and medaka as outgroups. Branch support numbers are shown at the tree nodes and were calculated with 10,000 ultrafast bootstrap replicates. (**D**) Specific regions of the multiple sequence alignment of several PIWI domains, surrounding the integral residues of the catalytic triad, indicated with black arrowheads, the catalytic residues within the PIWI domain known to be important for Piwi-mediated cleavage. These residues are conserved in Piwil1 proteins of African cichlids, including in *piwil1.3* and *piwil1.4* in Lake Malawi. (**E**) Structural alignments of the PIWI domain of *Drosophila melanogaster* (Dm) Piwi protein and AlphaFold predictions of Piwil1.1 (using only PIWI domain, left), Piwil1.3 (full-length, centre), and Piwil1.4 (full-length, right) of *A. calliptera*. Regions of the structural alignment encompassing the catalytic triad are augmented in the insets and the triad residues are highlighted with black or white arrowheads.

percentiles, respectively, of genes with highest values of integrative sweep signatures, supporting positive selection at these loci (**Figure 2E** and **Supplemental Table 1**). Moreover, we found evidence of positive selection in cichlid Piwil1 proteins beyond Lake Malawi, particularly in amino acid residues in the PIWI domain or immediately C-terminally adjacent to the annotated domain (**Figure S4A**). The results above are in line with positive selection acting on cichlid Piwi proteins, most notably in the expanded Piwi repertoire of Lake Malawi cichlids. Overall, the data suggests a scenario consistent with *piwil1* expansion early in the radiation, followed by positive selection and gene losses.

Next, we sought to determine whether the expanded copies of *piwil1* genes in Lake Malawi are expressed. We excluded *piwil1.2* from further analysis, because both the premature stop codons in conserved exons (**Figure 2A**) and low expression (**Figure S3B**), suggest that it is a pseudogene. First, we interrogated *piwil1* gene expression at the mRNA level. We also probed the expression of *piwil2*, the *piwi* gene homolog of zebrafish *zili*[33], which did not undergo gene duplication. *piwil1.1* and *piwil2* are strongly expressed in gonads but not in brain (**Figures 3A** and **S4B**), in line with known TE silencing roles in the germline of other organisms[8,27,33]. *piwil1.4* was expressed in gonads, and lowly expressed in brain. During early development of Lake Malawi cichlids, we detected strong maternal deposition of *piwil1.1* and *piwil2* transcripts (**Figure 3B**). In contrast, *piwil1.4* seems to be expressed mainly after gastrulation, likely after the onset of zygotic expression. No expression of *piwil1.3* was detected in these organs and in early development (**Figure 3A-B).**

To gain further insights into the potential function of these Piwil1 proteins, we analysed their protein sequence and structure. As Piwil1.3 and Piwil1.4 have only the PIWI domain (**Figure 2B**), we focused on the portion of Piwil1 proteins encompassing this domain. We found low overall variation in African cichlid Piwil1 proteins, but the Lake Malawi truncations showed higher divergence than their full-length orthologs (**Figure 3C**). This divergence is not expected to disrupt protein structure, as the predicted structures of full-length Piwil1.3 and Piwil1.4 proteins align well with the known structures of Piwi proteins of *Drosophila melanogaster* and *Bombyx mori*[50,51], and the predicted PIWI domain of Piwil1.1 (**Figure S4C**). The PIWI domain is a ribonuclease H-like domain, the catalytic centre of Argonaute proteins responsible for their slicer activity. Within the PIWI domain, a DDE motif of amino acid residues is required for Argonaute cleavage[52,53]. Despite the higher divergence of Piwil1.3 and Piwil1.4, they retain a conserved DDE motif, as Piwil1.1 (**Figure 3D**). Furthermore, the PIWI domain structures of Lake Malawi Piwil1 proteins are predicted to be identical to those of *D. melanogaster* and *B. mori* Piwi proteins, including the relative position of the DDE motif residues (**Figures 3E** and **S4D**). These data indicate the genomes of Lake Malawi cichlids encode three Piwil1 proteins with potentially catalytically active PIWI domains.

Cichlids express TE-targeting piRNAs with signatures of active silencing

To characterise the piRNA cofactors of cichlid Piwi proteins, we sequenced sRNAs from gonads of the selected cichlid species (**Figure 1A**). The sRNA length distribution profiles in gonads have prominent peaks at lengths of 21-22 nucleotides, likely corresponding to microRNAs (**Figure S5A**). Contrary to microRNAs, piRNAs have high sequence diversity[27]. When sRNA reads are collapsed into unique sequences, we observed prominent sRNA populations between 24-31 nucleotides long, consistent with the length distribution of piRNAs (**Figure 4**, left panels). In testes, sRNA populations peaked at lengths of 26-27 nucleotides, whereas in ovaries the peak was shifted to 28-29 nucleotides long (**Figures 4** and **S5B**).

We selected sRNAs between 24-35 nucleotides long for subsequent analysis and searched for the two typical sequence signatures of piRNAs: a bias for uridine at position 1 (1U), and a bias for an adenine at position 10 (10A)[27,31,32,34,35,54]. Unique sRNA sequences between 24-35 nucleotides long clearly show both the 1U and 10A biases in cichlid gonads, as well as additional signatures consistent with active piRNA ping-pong and phased biogenesis pathways (**Figures 4** and **S5B-G**). Of note, while phased piRNA biogenesis is pervasive in cichlid testes, ovary sRNAs display no clear signatures of phased biogenesis, except in AC ovaries (**Figures 4** and **S5F-G**). As a control, we sequenced sRNAs from muscle, as a representative somatic tissue of AC and found no prominent population of sRNAs in the piRNA length range with piRNA signatures (**Figures 4A** and **S5E-F**). Thus, cichlid gonads express sRNA populations consistent in length and sequence signatures with an active piRNA pathway.

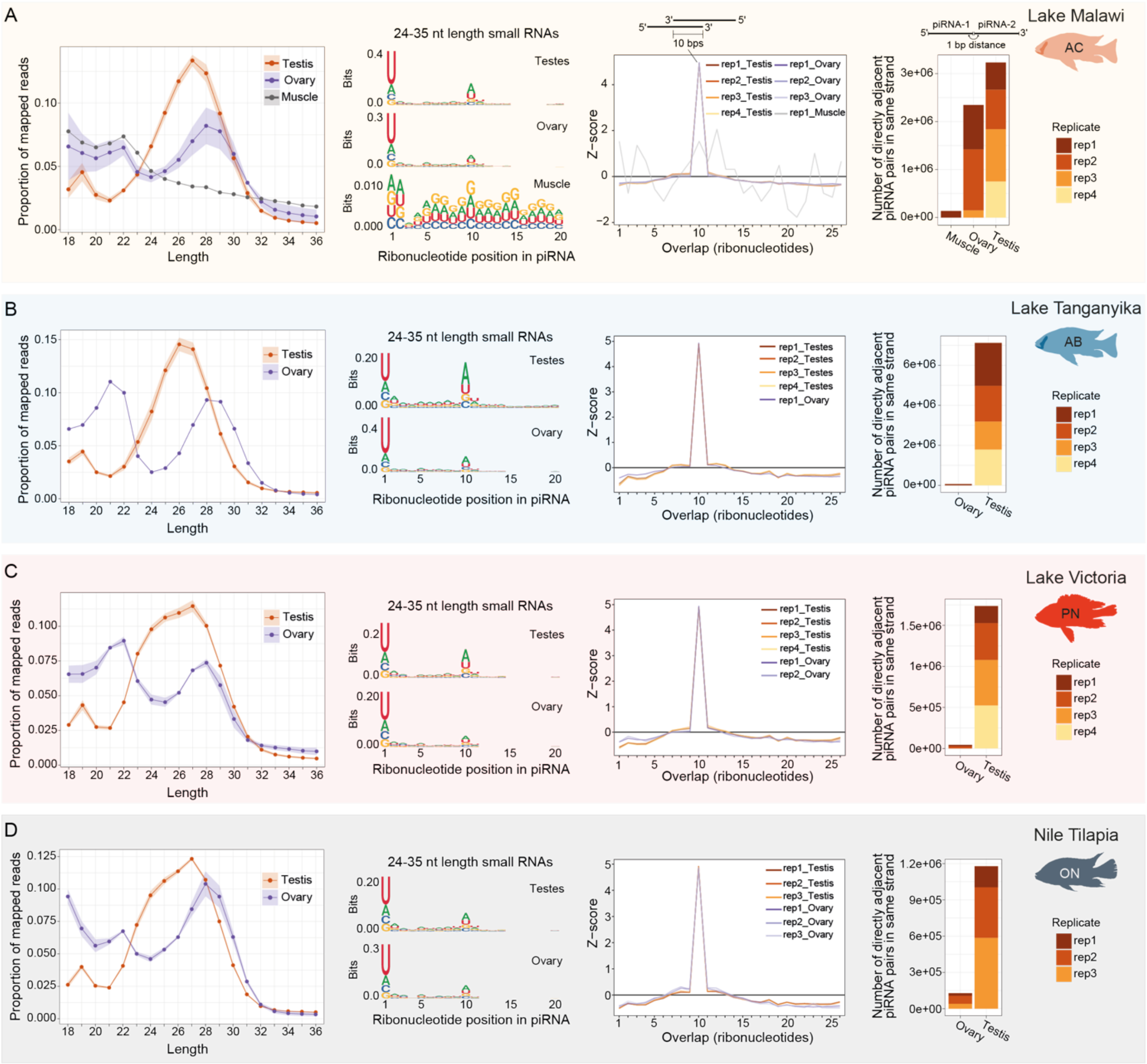


Figure 4. African cichlids express piRNAs in gonads. (**A-D**) sRNA length distribution profiles and piRNA sequence signatures in sRNAs 24-35 nucleotides long. sRNA length profiles shown here (left-most panels) comprise only reads of unique sequence. The shading in the sRNA length distribution profiles indicates standard deviation of replicates. Sequence logos (second set of panels from left) denote the 1U bias typical of piRNAs, and the 10A signature of ping-pong amplification in gonad sRNAs but not in muscle tissues of *A. calliptera* (A). Third set of panels from left show ping-pong signature with a robust overlap of 10 ribonucleotides in piRNA pairs. Right hand-side panels show number of piRNA pairs in same orientation that are directly adjacent, indicative of phased piRNA biogenesis. Signature is observable in the testes of all species, but in ovaries it is detectable only in AC. AB, *Astatotilapia burtoni*; AC, *Astatotilapia calliptera*; CPM, Counts Per Million; ON, *Oreochromis niloticus*; PN, *Pundamilia nyererei*.

piRNAs are often created from discrete genomic regions termed piRNA clusters[26,27,31,32]. To finely map piRNA clusters, we used a novel computational approach that identifies piRNA clusters by incorporating information from uniquely- and multi-mapping reads in a stepwise manner (see **Methods**). We restricted the analysis to sRNA sequencing data of Lake Malawi (AC, TM, and MZ, all mapped to the AC genome) and ON, because these chromosomal level assemblies allow us to define the piRNA clusters within genomic coordinate systems and to understand their biological context. We identified thousands of genomic sources of piRNAs in Lake Malawi (**Figure 5A-B** and **Supplemental table 2**, between 3,091-3,251 in ovaries and 3,494-4,252 in testes) and ON (**Figure 5C-D** and **Supplemental table 2**, between 3,194-7,352 in ovaries and 4,053-4,781 in testes). The clusters explain 65-80% of piRNA reads in the library (**Figure S6A**). Although the total number of clusters are comparable in testes of distinct Lake Malawi species (**Figure 5B**, compare testis of AC, MZ, and TM), the number of clusters shared between all three species are a fraction of the total (**Figure 5B**, 1,377 shared clusters), revealing variation in piRNA production in closely related species of Lake Malawi cichlids. Moreover, the even lower

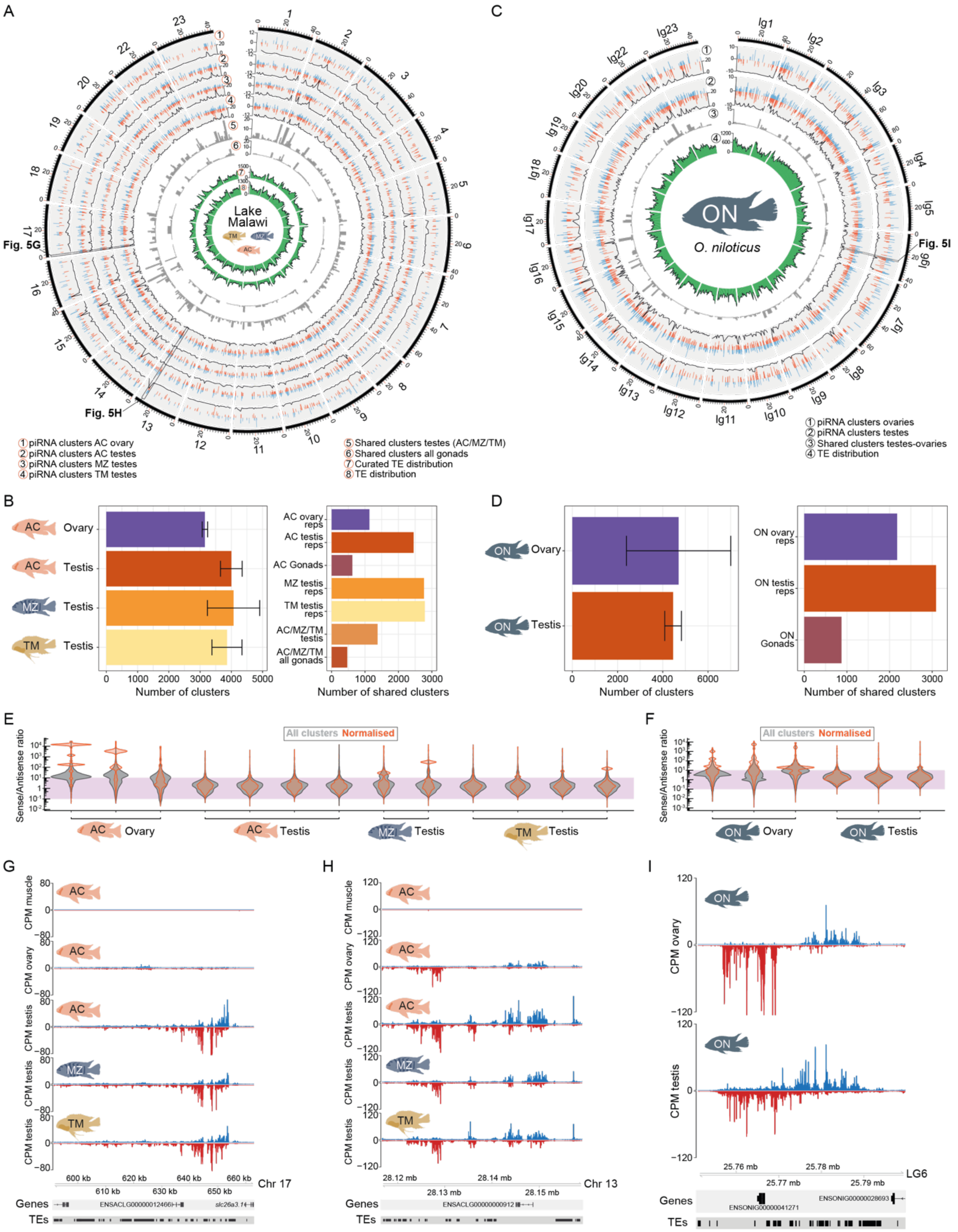


Figure 5. Fluid genomic origins of cichlid piRNAs. (**A**) Circos plot showing the chromosomal locations of piRNA clusters in Lake Malawi cichlid gonads (tracks 1-4), clusters shared between all replicates of each organ (tracks 5-6), and TE distributions (tracks 7-8) from curated (track 7) and non-curated annotations (track 8). In tracks 1-4, blue and red represent the log2 mean Reads Per Kilobase Million (RPKM) of piRNA clusters in the plus and minus strands, respectively. In the bottom of tracks 1-4 is a line plot with the density of clusters. (**B**) Left panel shows the mean number of clusters identified in Lake Malawi cichlid gonads. Error bars represent standard deviation. Right panel depicts the number of clusters shared between the replicates of the organs indicated. (**C**) Circos plot showing the chromosomal locations of piRNA clusters in ovaries (track 1) and testes (track 2), the shared clusters between these two organs (track 3), and the TE distribution (track 4). In tracks 1-2, blue and red represent the log2 mean RPKM of piRNA clusters in the plus and minus strands, respectively. In the bottom of tracks 1-2 is a line plot with the density of clusters. (**D**) Left panel represents the mean number of clusters identified in *O. niloticus* gonads. Error bars represent standard deviation. Right panel shows the shared clusters between the replicates indicated. (**E-F**) Strand biases in piRNA production, shown as the ratio of sense over antisense piRNAs intersecting each piRNA cluster. The grey violin plot represents all piRNA clusters identified, while the orange violin plot represents the sense/antisense ratio normalised according to cluster productivity. The purple region highlights piRNA clusters with piRNA production less than 100-fold different between the sense and anti-sense strands. Thus, values that fall within this range likely account for piRNA clusters producing piRNAs from both strands. (**G-I**) Genome tracks with examples of clusters identified in Lake Malawi cichlids (G-H) and in *O. niloticus* (I). Blue and red tracks represent 24-35 nucleotide long piRNAs, in Counts per Million (CPM), mapping to the plus and minus strands, respectively. (G) shows a testes-specific piRNA cluster in Lake Malawi. (H-I) are examples of clusters shared by ovary and testis of Lake Malawi cichlids (H) and *O. niloticus* (I). AC, *Astatotilapia calliptera*; CPM, Counts Per Million; MZ, *Maylandia zebra*; ON, *Oreochromis niloticus*; TM, *Tropheops* sp. ‘mauve’.

380

number of clusters shared between testes and ovaries illustrates sex differences in piRNA production (**Figure 5B, D**, 622 shared clusters in AC gonads, 469 clusters shared across all Lake Malawi testes and ovaries, and 872 clusters shared in ON gonads). Overall, these results suggest considerable fluidity in the sources of piRNA production in cichlids, including in cichlids inhabiting the same Lake.

Next, we explored additional features of the cichlid piRNA clusters identified. Most piRNA clusters are shorter than 50 kb (**Figure S6B-C**). In testes, clusters tended to be larger than in ovaries (**Figure S6B-C**, median length of 12.3 kb in AC testes versus 2.89 kb in AC ovaries, and median length of 13.70 in ON testes versus 4.62 kb in ON ovaries). Within Lake Malawi, median cluster lengths in testes were consistent in AC, MZ, and TM (**Figure S6B**). piRNA clusters are spread throughout the entire genome and do not tend to be in close proximity (**Figures 5A, C** and **S6D**). piRNA clusters tend to produce piRNAs from both strands, although in ovaries there is a bias for sense piRNAs (**Figure 5E-F**). In terms of productivity, we found that a fraction of clusters generate the majority of the piRNAs in the library (**Figure S6A**). We found no relationship between the productivity and length of piRNA clusters (**Figure S6E**). Examples of large, highly productive piRNA clusters are shown in **Figure 5G-I**.

The majority of piRNA clusters in AC and ON overlap with intergenic regions and TEs, and the observed overlap with TEs is higher than expected by chance (**Figures 6A-B** and **S7A**). Of all TEs, LTRs are significantly enriched in piRNA clusters (**Figures 6A-B** and **S7A**). The piRNA clusters not detected in the testes of the three Lake Malawi species tend to follow similar enrichment trends (**Figure S7B**). Furthermore, these species-variable piRNA clusters are enriched in genomic regions of Lake Malawi cichlids associated with structural variation (**Figure S7C**), defined in recent work[5]. This suggests the existence of piRNA-producing sequences that are structural variants in Lake Malawi cichlids. We overlapped all 24-35 nucleotide long piRNAs of AB and PN with genome features and TE classes and observed enrichments similar to those of AC and ON piRNA clusters (**Figure S7D-E**). Next, we further explored sequence signatures of TE-mapping piRNAs. We found 1U bias and 10A bias in piRNAs mapping sense and antisense to TEs, consistent with active targeting of TEs (**Figures 6C-D** and **S7F-G**). Sense piRNAs have higher 10A bias and lower 1U bias than piRNAs antisense to TEs (**Figure 6C-D** and **S7F-G**). These signatures are absent from muscle (**Figure 6C**). In terms of piRNAs mapping to distinct TE classes, we did not find consistent differences between testes and ovaries across all species (**Figure S7H**), contrasting with the sex differences in TE expression (**Figure S1C**).

We identified TE families likely to be transpositionally active in Lake Malawi cichlids (P. Sierra & R. Durbin, unpublished results) and compared expression in gonads of likely mobile TE families with transpositionally inactive families of the same superfamily. We found higher expression of mobile versus immobile families in five TE superfamilies in at least one organ (**Figure 6E**, upper panel). Conversely, in three TE superfamilies we observed lower expression of the mobile versus immobile TE families, in line with ongoing silencing (**Figure 6E**, lower panel). Two of these TE superfamilies (CMC-EnSpm and Maverick) showed the opposite correlation in terms of piRNA levels: mobile TE families were targeted by higher piRNA levels (**Figure 6F**, lower panels), further supporting ongoing piRNA-driven TE silencing. No clear trend was observed in relation to piRNA levels for mobile TE families more highly expressed than their related immobile families (**Figure 6F**, upper panels). Additional data agree with robust targeting of CMC-EnSpm and Maverick families by piRNAs (**Figures 6G** and **S7I**). Furthermore, these families showed dynamic expression patterns in early development: mobile families of CMC-EnSpm, which are DNA TEs, are more highly expressed throughout early development than their immobile relatives (**Figure 6H**,

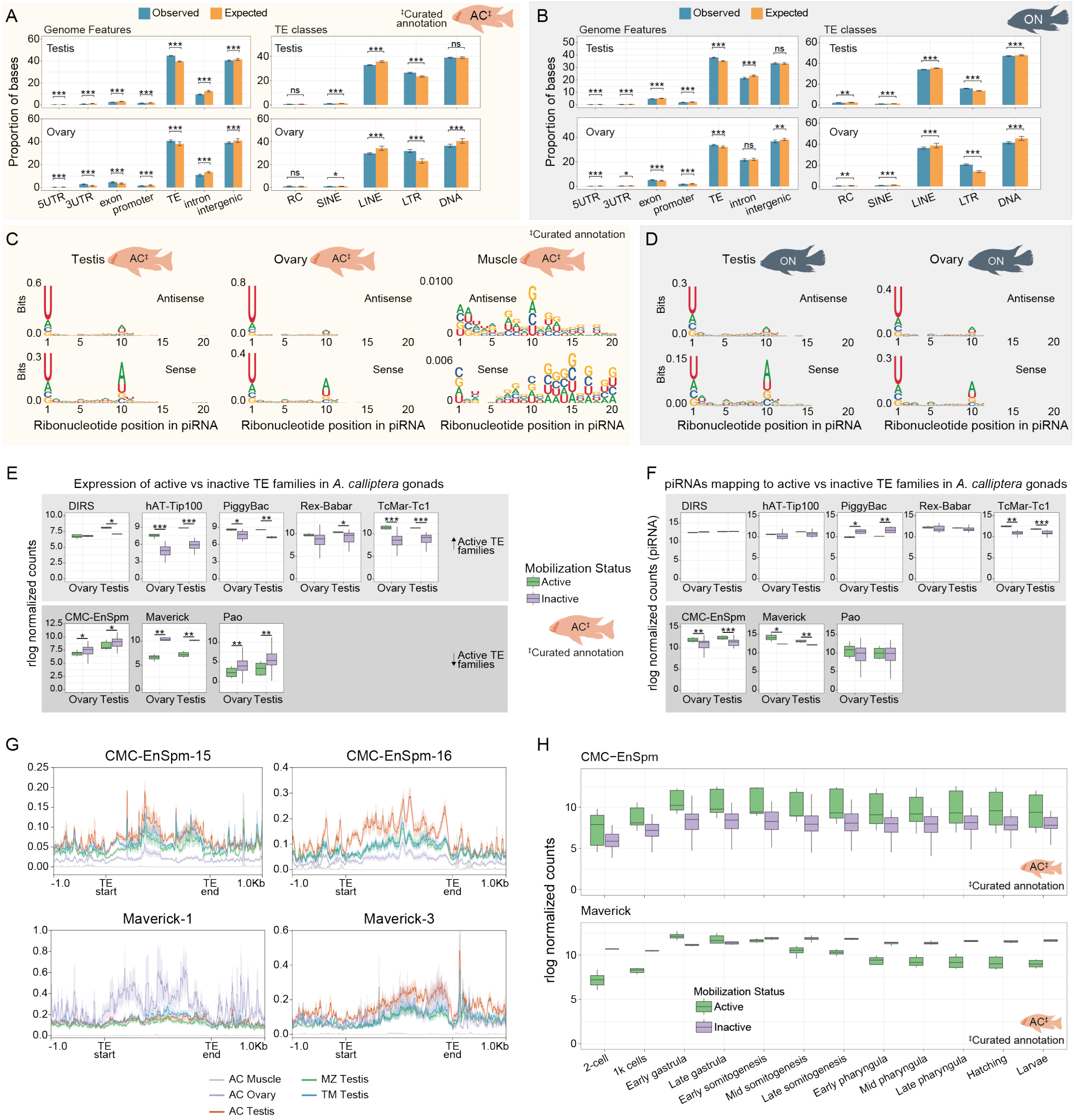


Figure 6. Cichlid piRNAs target TEs. (**A-B**) Observed and expected values at genomic features and TE classes that piRNA clusters overlap with in *A. calliptera* (A) and *O. niloticus* (B). ns, not statistically significant. In (A) we overlapped the piRNA clusters with TE features from the TE annotation produced with a curated library. (**C-D**) Sequence logos of 24-35 nucleotide long piRNAs mapping sense or antisense in regard to TE orientation in *A. calliptera* (C) and *O. niloticus* (D). The 1U and 10A signatures are observable in gonads but not in muscle. For *A. calliptera*, TE features were extracted from the curated TE annotation. (**E**) The mRNA expression in cichlid gonads of likely transpositionally active vs inactive families of a given TE superfamily. Panels above: superfamilies where the active families are more highly expressed than the inactive families. Panels below: TE superfamilies with higher expression of inactive families. P-values were calculated with Wilcoxon rank-sum tests (using Benjamini & Hochberg correction) comparing TE families with distinct mobilisation status in each gonad. The expression data was quantified using a curated annotation of Lake Malawi TEs. (**F**) piRNAs mapping to transpositionally active vs inactive TE families of the same TE superfamily. P-values were calculated with Wilcoxon rank-sum tests (using Benjamini & Hochberg correction) comparing TE families with distinct mobilisation status in each gonad. piRNAs were mapped to a curated TE annotation in Lake Malawi cichlids. Panels above and below, represent TE superfamilies with opposing relative expression of active versus inactive TE families, according to **Figure 6E**. (**G**) Metagene plots depicting mean piRNA levels mapping to all TEs of likely transpositionally active families. The shading represents standard error of replicates. TE start and TE end indicate start and end coordinates, respectively, of TE in the annotation. (**H**) The expression of transpositionally active versus inactive CMC-EnSpm (panel above) and Maverick (panel below) families throughout early development of *A. calliptera*. A curated annotation of Lake Malawi TEs was used to calculate expression data. AB, *Astatotilapia burtoni*; AC, *Astatotilapia calliptera*; ON, *Oreochromis niloticus*; PN, *Pundamilia nyererei*; sRNA, small RNA.

upper panel), a reverse pattern to that observed in gonads (**Figure 6E**, lower panel). In turn, mobile TE families of the Maverick superfamily, also DNA TEs, show an expression pattern in accordance with early silencing due to maternal silencing factors, followed by weakening of silencing at gastrulation until zygotic silencing can be re-established (**Figure 6H**, lower panel). Altogether, our data shows piRNAs target TEs and are likely to be engaged in ongoing silencing of transpositionally active TE families.

Discussion

In this work, we describe three main findings that altogether suggest dynamic TE co-evolution with host control mechanisms in East African cichlids: 1) dynamic TE expression; 2) an expanded repertoire of *piwil1* genes; and 3) fast evolution of piRNA clusters. We will elaborate on these points below.

First, hundreds of TEs families are dynamically expressed in gonads and early development of cichlids (**Figures 1** and **S1**). Given the extensive shared polymorphism in African cichlids due to hybridisation[3,4,18,55–58], we adopted a more conservative approach by initially quantifying TE expression at the family level. We found sex-biased expression patterns of TEs and protein-coding genes (**Figure S1C**), with higher median expression in testes versus in ovaries. This asymmetry is likely to be the result of overall higher transcriptional output in testes. Interestingly, this sex asymmetry does not consistently extend to piRNAs mapping to TEs (**Figure S7H**), suggesting that piRNA precursor transcription may not follow general transcription trends. Testes contain the male germline and are a relevant organ in the context of genetic conflict between TEs and host silencing factors[8,27]. Higher gene and TE expression in testes is consistent with previous studies describing more widespread expression and increased transcriptome complexity in mammalian testes[59,60]. TEs have been shown to contribute to transcriptome complexity in the mammalian germline[61] and in fish testes[62]. It may be worth exploring in depth whether cichlid testes, much like mammalian testes, have increased transcriptome complexity and diversity, and if this has contributed to the cichlid radiations. Indeed, gonad transcriptomes are evolving faster than transcriptomes of other organs in Lake Tanganyika cichlids[63]. In early development, we found that the majority of TE families are most highly expressed during gastrulation, a period that may coincide with the maternal-to-zygotic transition in cichlids. Zygotic transcription of TE silencing factors may initiate concomitant to the onset of general zygotic transcription, leading to zygotic TE silencing. Expression of transpositionally active Maverick TEs in early development could illustrate just that (**Figure 6H**, lower panel). Expression analysis of individual TE loci revealed TEs with expression in discrete developmental times (**Figures 1C** and **S1G**). As expected given their evolutionary distance, the TE classes enriched in particular developmental stages in cichlids differ substantially from those enriched in the same developmental stages in zebrafish[13]. However, a striking similarity to zebrafish is enrichment of TEs belonging to ERV1, Gypsy, and Pao LTR superfamilies in gastrula stages (**Figure S1F-G**). It will be relevant to investigate how the maternal-to-zygotic transition and/or epigenetic reprogramming affect LTR transcription and transposition during early fish development.

Second, we find an expanded repertoire of *piwil1* genes in Lake Malawi cichlids and signatures of positive selection on the novel copies (**Figure 2**). Lability in copy number and positive selection on TE silencing factors are two signatures associated with arms races between TEs and their animal hosts[24,25,38–41]. These findings also add to the notion that piRNA pathway factors, including *piwil1* genes, evolve fast in teleosts[64]. Interestingly, TEs, the targets of Piwi proteins, likely have mediated, at least partially, the expansion of *piwil1* genes in Lake Malawi cichlids. We found closely related

PiggyBac elements associated with the three novel *piwil1* genes, but not with the *piwil1* copy sharing synteny with other vertebrate *piwil1* genes, presumably the original copy (**Figures 2** and **S3** and **Supplemental Table 1**). We also found TIRs flanking the PiggyBac and putative TIRs distal to the PiggyBac and 5' to the *piwil1* copies. The non-coding differences of the four *piwil1* genes suggest the succession of events underlying the expansion: first a duplication of *piwil1.1* creating *piwil1.2*, followed by creation of one of the truncated copies from *piwil1.2*, and its subsequent duplication (**Figure 2D**). Given the PiggyBac TIR signatures, it is likely that at least the first duplication was mediated by transposition, but we cannot exclude that subsequent duplications were driven by a recombination-based mechanism. By leveraging available genomic resources we determined that *piwil1.1* and *piwil1.4* seem to be fixed or nearly fixed in Lake Malawi, whereas *piwil1.2* and *piwil1.3* are less widespread (**Figure 2C**). *piwil1.3* seems to have negligible expression in the germline and early development (**Figures 3A-B**). It is possible that *piwil1.3* is expressed and functional in other organs beyond the gonads and brain, or in juvenile developmental stages between larval stage and sexual maturity. An alternative is that *piwil1.3* is a pseudogene, similar to *piwil1.2*.

The exact function of *piwil1.3* and *piwil1.4* remains to be determined. Knock-outs of *piwil1.1* and *piwil1.4* will be key to inform on their function. The annotated Piwil1.3 and Piwil1.4 proteins are predicted to encode a catalytically competent PIWI domain (**Figures 2A-B** and **3D-E**), the catalytic centre of Argonaute proteins responsible for slicer activity[52,53]. The Argonaute domains lacking in Piwil1.3 and Piwil1.4, the MID and PAZ domains (**Figure 2B**), are predicted to serve as binding pockets for the 5' and 3' ends of the piRNA, respectively[65–67]. Without these domains, Piwil1.3 and Piwil1.4 are most likely not able to bind to piRNAs or other sRNAs, and will probably function independently of piRNAs. Thus, these truncated Piwi proteins were likely repurposed for a piRNA-independent gene regulatory role, related, or not, to TE silencing.

Third, we find fast evolution of piRNA clusters in cichlids. The majority of piRNAs were produced from intergenic regions and TEs (**Figures 5,6**) and 65-80% of these sequences can be grouped into discrete piRNA-producing clusters (**Figure S6A**). We identify piRNA clusters with sex-biased expression, and, interestingly, variation in piRNA clusters even in testes of closely related Lake Malawi cichlids (**Figure 5**). These observations indicate that piRNA clusters are fast-evolving modules in Lake Malawi. An in-depth population-wide analysis of piRNA populations and piRNA clusters in Lake Malawi will be useful to determine just how rapidly these units are evolving in cichlids. In terms of piRNA biogenesis, we find conserved differences in cichlid piRNA populations with peaks at 26-27 nucleotides long piRNAs in testes versus 28-29 nucleotide long piRNAs in ovaries (**Figure 4**). These piRNA size differences may be driven by Piwi Argonaute size preferences. The most striking difference in terms of piRNA biogenesis however, is the lack of consistent phasing signature in the ovaries of East African cichlids outside Lake Malawi (**Figures 4** and **S5**). It will be interesting to determine the factor(s) inhibiting phased biogenesis in cichlid ovaries.

Three sets of observations point towards TEs as key genetic elements contributing to cichlid diversification: 1) TEs represent a previously underestimated source of genetic diversity in African cichlids[5]; 2) TEs have been linked with pigmentation and vision traits, sex determination, and gene expression changes[18,20–23]; and 3) the ongoing dynamic TE-host co-evolution and arms races that our findings suggest. It remains unclear how the latter connects with cichlid phenotypic diversification. We expect it does not come down to the number of TE families or the proportion of the genome comprised by TEs. In this regard, zebrafish provides a much more striking example, with nearly 2,000 distinct TE families, occupying more than 50% of its genome[6,13], versus 557-828 TE families and 16-41% of the genome in cichlids (**Figure S1A-B**, **D**). However, the *Danio* genus of zebrafish did not diversify nearly as prolifically as East African cichlids despite its massive TE content.

What led to the unparalleled rates of phenotypic diversification observed in East African cichlids? Recent work on the cichlid radiation of Lake Victoria suggests that ecological versatility is the key[68,69]. Key features contributing to cichlid versatility include strong sexual selection, highly plastic jaw structures, and abundant interspecific hybridisation[1,68]. The regulatory consequences of hybridisation are one possible avenue to pursue to study the influence of TE-host co-evolution in cichlid radiations. Genomic studies have elucidated a complex evolutionary history of East African cichlids, marked by substantial amounts of gene flow occurring through hybridisation[4,18,55–58]. It will be important to determine how interspecific cichlid hybrids tolerate regulatory mismatches driven by genetic conflict between TEs and the piRNA pathway. It is conceivable that bouts of TE expansion following hybridisation generated the (epi)genetic potential for the radiations. This study provides a platform to investigate this hypothesis, an initial understanding of TEs and piRNAs as two co-evolving modules. Going forward, learning about the co-evolution of these modules in the context of recurring hybridisation has the potential to give valuable insights into the genetic and molecular basis of the cichlid radiations.

Methods

Animal sampling and housing conditions

Astatotilapia calliptera and *Tropheops* sp. 'mauve' animals were grown in 220 Litre tanks, with pH 8, at approximately 28°C, and with a 12 h dark/light cycle. Males and females of each species were housed only with conspecifics. Feeding, housing, and handling were conducted in strict adherence to local regulations and with the protocols listed in Home Office project license PP9587325. Fish were fed twice a day with cichlid flakes and pellets (Vitalis). Tank environment was enriched with plastic plants, plastic hiding tubes, and sand substrate. Aquaria grown animals were euthanised with approved Home Office schedule 1 protocols, namely using 1 g/L MS-222 (Ethyl 3-aminobenzoate methanesulfonate, Merck #E10521) and subsequent exsanguination by cutting the gill arches, in accordance with local regulations. Afterwards, gonads, brain and dorsal muscle tissue were carefully dissected, swiftly snap frozen in dry ice and stored at approximately -80°C.

Dominant adult male *Maylandia zebra* bred and raised in captivity were obtained from commercial supplier Kevs Rifts and culled in Cambridge animal facilities, following an ethically approved post–transport adjustment period. *M. zebra* animals were euthanised using approved Home Office schedule 1 protocols as above. *Pundamilia nyererei* animals were raised in stock tanks of dimensions 59 cm(L) x 45 cm(B) x 39 cm(H) and moved to larger tanks 177cm (L) x 45cm(B) x 39cm(H) once they reached approx. 7 cm long. Temperatures were kept at 26C°, with constant daily water change of about 10% and 12:12 light dark regime. Frozen tissue samples of *Astatotilapia burtoni* were provided by Hans Hofmann and Caitlin Friesen (University of Texas at Austin, Austin, TX, USA). *Oreochromis niloticus* frozen tissue samples were provided by David Penman, Alastair McPhee, and James F. Turnbull (Institute of Aquaculture, University of Stirling, Stirling, Scotland, UK).

Orthology analysis

To identify orthologs of conserved factors involved in TE silencing pathways, we used OrthoFinder[70,71] v2.3.12. We used Ensembl proteomes (downloaded on 02/06/2020) of Homo sapiens (GRCh38), *Mus musculus* (GRCm38), *Oryzias latipes* (ASM223467v1), *Danio rerio* (GRCz11), *Takifugu rubripes* (fTakRub1.2), *Gasterosteus aculeatus* (BROADS1), *Amphilophus citrinellus* (Midas_v5), *Oreochromis aureus* (ASM587006v1), *Oreochromis niloticus* (O_niloticus_UMD_NMBU), *Astatotilapia burtoni* (AstBur1.0), *Neolamprologus brichardi* (NeoBri1.0), *Pundamilia nyererei* (PunNye1.0), *Astatotilapia calliptera* (fAstCal1.2) and *Maylandia zebra* (M_zebra_UMD2a). OrthoFinder was run on proteomes containing the longest protein isoform, parsed using a script provided with OrthoFinder (https://github.com/davidemms/OrthoFinder/blob/master/tools/primary_transcript.py). Initially, we ran OrthoFinder with the fish genomes above as inputs (except *M. zebra*), using option -f. Afterwards, we added human, mouse, and an additional Lake Malawi cichlid species *M. zebra* to this analysis using options -b and -f. We subsequently pinpointed the orthogroups containing known human, mouse and zebrafish TE silencing factors and extracted the gene IDs of their cichlid orthologs.

Piwil1 evolutionary analysis

Piwil1 protein orthologs were identified with OrthoFinder (see **Orthology analysis** above). Schematic of domain structure of Piwil1 proteins was plotted in R[72], with packages drawProteins[73] and tidyverse[74]. Coordinates of the MID domain were manually added to Piwil1 proteins, as this information was not present in Uniprot, which drawProteins relies on. MID domain coordinates in *A. calliptera* Piwil1 proteins were inferred from the MID domain coordinates of zebrafish Ziwi in Uniprot, through a multiple sequence alignment of *A. calliptera* Piwil1 proteins and Ziwi.

To determine the presence and absence of *piwil1* copies and their 3' trailing PiggyBac-1 TEs across Lake Malawi cichlid eco-morphological groups and genera, we probed the reads of 74 previously published short-read genomes[4], 5 new short-read genomes, as well as 12 long-read genomes (**Supplemental table 1**). Short-read genomes were aligned to the *A. calliptera* reference genome (fAstCal1.2, GCA_900246225.3) using bwa mem v0.7.17-r1188 (arguments: -C -p) using default settings[75]. Using samtools v1.9[76], the resulting alignment files were then further processed with fixmate (arguments: -m), sort (arguments: -l0) and mardup. Long-read genomes were aligned to the same reference using minimap2 v2.17-r974-dirty[77] (arguments: -ax map-pb --MD) and then sorted and indexed using samtools v 1.16-9-g99f3988. We manually checked whether read alignments showed robust support in specific eco-morphological groups/genera for the presence of each *piwil1* paralog and 3' trailing piggyBac copy using IGV v2.9.4[78]. Next, we manually determined the exact features of these regions using the *piwil1* gene annotations of fAstCal1.2[79], our TE annotation created from a curated TE library (see section Transposable element annotations), and genomic alignments of the entire regions encompassing all *piwil1* paralogs. Initial alignments of the paralog loci were generated by aligning the fAstCal1.2 reference genome to itself using Winnowmap2[80] (options: -ax asm5 --MD). Potential stop codons in *piwil1* paralogs were assessed in a multiple sequence alignment between *piwil1.1*, *piwil1.2* (reverse complement), *piwil1.3*, and *piwil1.4* (reverse complement) genomic regions, which was created MUSCLE v3.8.31[81] using default settings and then curated manually in AliView v1.27[82]. The exons of ENSACLT00000021959, the canonical ENSEMBL isoform of *piwil1.1*, the best evolutionarily conserved *piwil1* gene, was projected to the aligned sequences of the paralogs. A second alignment was created analogously, which additionally included the homologous *piwil1* sequence from

Oreochromis niloticus. Based on the latter alignment, we calculated Hamming distances (github.com/ssciwr/hammingdist) separately for intronic and exonic regions and built neighbour joining trees (github.com/scikit-bio/scikit-bio). Alignment files can be found at https://github.com/migueldvalmeida/Cichlid_TEs_piRNAs2024. The *A. calliptera* TE annotation created from a curated TE library was used to identify the PiggyBac-1 TE and its terminal inverted repeats. The terminal inverted repeat sequence of PiggyBac TEs (5'-CCCTT-3') was extracted from https://dfam.org/classification/dna-termini[47].

For the selection analysis we restricted our existing callset of more than 2,000 whole-genome sequenced Lake Malawi cichlids (github.com/tplinderoth/cichlids/tree/master/callset), which are all aligned against the chromosome level fAstCal1.2 reference genome[79], to the 79 individuals used in **Figure 2C** (**Supplemental Table 1**). The subset was generated with bcftools view[83], v.1.16-9-g99f3988) (arguments --types snps -m2 -M2 -f PASS -S $sample_list) to retain exclusively biallelic SNPs that passed all filters. Chromosome-scale VCFs along with the four largest contigs (> 1 Mbp) were concatenated into a single VCF using bcftools concat and served as the input for the selection analysis. A selection scan was performed using Raised Accuracy in Sweep Detection (RAiSD) v2.9[84] (with arguments: -f -M 3 -y 2 -m 0 -R -I). PiggyBac-1 sequences adjacent to *piwil1* genes were extracted according to their annotation coordinates, and aligned with the PiggyBac-1 family consensus from the curated TE library using MAFFT v7.475[85] with option --auto. L-INS-i was the alignment method automatically selected. Alignment visualisation was optimised in Jalview v2.11.2.7[86]. To expand the analysis, we extracted all the PiggyBac-1 sequences annotated in the *A. calliptera* reference genome (according to the curated TE annotation) with SWscore > 1000, and aligned them with MUSCLE v3.8.31[81]. We further filtered the alignment to contain only the region encompassed by the PiggyBac1 elements associated with *piwil1.2*, *piwil1.3*, and *piwil1.4*, and removed alignment columns consisting almost exclusively of missing data. A phylogenetic tree was constructed with IQ-TREE v2.1.2[87], option -B 1000. TPM2+F+R2 was the best fit model. Trees were visualised and annotated in FigTree v1.4.4 (https://github.com/rambaut/figtree).

The sequences of Piwil1 protein orthologs were collected from Ensembl. For *Piwil1* genes encoding more than one protein isoform, the longest isoform was chosen for analysis. As *A. calliptera piwil1.2* may be a pseudogene, we did not include its predicted protein sequence in the subsequent analysis. Fish Piwil1 proteins were aligned with MAFFT v7.475[85], using option --auto, and L-INS-i was the alignment method automatically selected. We trimmed the alignment manually, keeping only 296 sites corresponding to the C-terminal region of the proteins with excellent alignment score, which includes the PIWI domain. Original protein sequences and alignment files can be found at https://github.com/migueldvalmeida/Cichlid_TEs_piRNAs2024. IQ-TREE v2.1.2[87] was used to construct phylogenetic trees from these two alignments with options -B 10000 -o {medaka and zebrafish Piwil1 proteins were defined as outgroups}. -B parameter refers to ultrafast bootstrap approximation[88]. PMB+G4 was the best fit model. To test for selection, we redid the alignment using a smaller subset of the proteins, including only Piwil1 proteins of African cichlids. Alignment of protein sequences was performed with MAFFT v7.475[85], using option --auto. L-INS-i was the chosen alignment model. Next, we used pal2nal v14[89] to produce a reverse alignment from an alignment of the protein sequences to an alignment of the coding sequences. The resulting reverse alignment was used as input for selection tests in Datamonkey[90]. A gene-wide test was first performed using Branch-site Unrestricted Statistical Test for Episodic Diversification (BUSTED)[91]. We conducted the test in two ways, testing for selection across all branches and testing for selection only in radiating cichlids, with

O. niloticus as an outgroup. BUSTED reported very strong support for positive selection in both cases (p-value = 0). **Figure S4A** shows the subsequent analysis to identify residues very likely to be under positive selection according to Mixed Effects Model of Evolution (MEME)[92].

To pinpoint catalytic residues of cichlid Piwil1 proteins, we first added the sequence of human PIWIL1 (HIWI) to the list of fish Piwil1 proteins used in the alignments above, and redid the alignment using MAFFT v7.475[85] with option --auto (L-INS-i was the model automatically chosen). The alignment was visualised in Jalview v2.11.2.7[86] and the catalytic residues were manually pinpointed based on their known positions in HIWI[52,53]. Structural alignments were performed with open-source PyMOL v2.5.0 using the align command. We aligned AlphaFold predictions of Piwil1.1 (Uniprot ID A0A3P8PWP0), Piwil1.3 (Uniprot ID A0A3P8NS09), and Piwil1.4 (Uniprot ID A0A3P8NRZ4) of *A. calliptera*, downloaded from AlphaFold Protein Structure Database[93,94], with crystal structures of *bombyx mori* Siwi (PDB ID 5GUH)[50] and *Drosophila melanogaster* Piwi (PDB ID 6KR6)[51]. As we focus on the PIWI domain, we aligned only the PIWI domains of *A. calliptera* Piwil1.1 (residues 550-856), *D. melanogaster* Piwi (residues 537-843), and *B. mori* Siwi (residues 593-899). As Piwil1.3 and Piwil1.4 of *A. calliptera* are truncations encompassing only the Piwi domain, we used their full-length structure for the alignments.

RNA extractions

Frozen brain, muscle, and gonad tissues were partitioned on a mortar positioned on dry ice, quickly to avoid thawing, and weighed. Biological replicates were created by collecting a similar mass of the same organ/tissue from size-matched individuals of the same species. 15-30 mgs of brain tissue, 26 mg of dorsal muscle tissue, and 14-144 mgs of gonad tissue were used, according to the specific tissue, tissue availability, and size of the specimen, which varied per species. Tissue pieces were transferred to BeadBug tubes prefilled with 0.5 mm Zirconium beads (Merck, #Z763772) and 500-600 µl of TRIzol (Life Technologies, #15596026) was added to the tubes and mixed vigorously. Afterwards, we conducted the homogenisation using a BeadBug microtube homogeniser (Sigma, #Z764140) at approximately 4°C (in cold room). Each sample was homogenised with five BeadBug runs at maximum speed (4,000 rpm) for 60 seconds each. No sample was run on BeadBug more than two consecutive times to avoid overheating. Other than the run time inside the BeadBug, samples were left on ice. After homogenisation, lysates were centrifuged for 5 minutes at 18,000 G at 4°C. Supernatant was then removed into a clean 1.5 mL tube. Centrifuged the lysates again, this time at maximum speed (approximately 21,000 G) for 5 minutes at 4°C. Transferred supernatant into a clean tube without disturbing the pellet and tissue debris. Mixed supernatant thoroughly 1:1 with 100% ethanol, pipetted the mix into a column provided in the Direct-zol RNA Miniprep Plus kit (Zymo Research, #R2072) and followed manufacturer's instructions, using the recommended in-column DNase I treatment.

Library preparation and sequencing

mRNA sequencing. Library preparation (directional, with poly-A enrichment) and sequencing (Illumina, PE150) of *A. calliptera*, *M. zebra*, *T.* sp. 'mauve', *A. burtoni*, and *O. niloticus* gonads was performed by Novogene. Libraries of *P. nyererei* gonads and *A. calliptera* brain tissues were prepared and sequenced as follows. Initial quality control was done using a Qubit Fluorometer (Invitrogen) and Qubit RNA HS Assay Kit (Invitrogen, #Q32855), and Agilent RNA TapeStation reagents (Agilent, #5067-5576; #5067-5577; #5067-5578). 50-250 ng of total RNA were used for library production with the NEBNext® Poly(A) mRNA Magnetic Isolation Module (NEB, #E7490), in conjunction with the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina® (NEB, #E7760) and the NEBNext® Multiplex Oligos for Illumina® (96 Unique Dual Index Primer Pairs, NEB #E6440). Quality control of the libraries was done with the Qubit dsDNA HS Assay Kit (Invitrogen, #Q32854) and Agilent DNA 5000 TapeStation reagents (Agilent, #5067-5588; #5067-5589). Samples were then pooled in equimolar amounts according to the TapeStation results and sequenced on a NovaSeq 6000 system (PE150 on one lane of an S1 Flowcell).

Small RNA sequencing. Initial quality control was conducted using a Qubit Fluorometer (Invitrogen) and the Qubit RNA HS Assay Kit (Invitrogen, #Q32855), and Agilent RNA TapeStation reagents (Agilent, #5067-5576; #5067-5577; #5067-5578). Samples were processed according to the NEXTFLEX® Small RNA-Seq Kit v4 with UDIs (PerkinElmer, #NOVA-5132-32) with a 1 µg starting input and 12 cycles of PCR. Quality control of the libraries was done with Qubit dsDNA HS Assay Kit (Invitrogen, #Q32854) and Agilent DNA 5000 TapeStation reagents (Agilent, #5067-5588; #5067-5589). Samples were then pooled in equimolar amounts according to the TapeStation results and sequenced on a Novaseq 6000 system (PE50 on one lane of an SP Flowcell).

Transposable element annotations

In each respective cichlid genome, transposable elements and repeats were first modelled and identified using RepeatModeler v1.0.11 in combination with the recommended programmes RECON v1.08, RepeatScout v1.0.6, TRF v4.0.9 and NCBI-RMBlast v2.14, and then annotated using RepeatMasker v4.0.9 in combination with NCBI-RMBlast v2.14, TRF v4.0.9 and the custom libraries of modelled repeats, Dfam3.0 and Giri-Repbase-20170127[95]. The curated TE library for Lake Malawi cichlids was created following a previously described protocol[96] and will be described in detail elsewhere (P. Sierra & R. Durbin, unpublished results). This library was used as input to RepeatMasker v4.1.2-p1[95] with options -e rmblast -no_is -gff -lib -a to generate a final TE annotation for the *A. calliptera* genome fAstCal1.2. GTF files with TE annotations amenable to be used for TEtranscripts (see below **Bioinformatic analysis**, mRNA-sequencing analysis section) were created using custom scripts (available at https://github.com/migueldvalmeida/Cichlid_TEs_piRNAs2024).

Bioinformatic analysis

mRNA-sequencing analysis. Illumina adapters and reads with low-quality calls were filtered out with Trimmomatic v0.39[97] using options SLIDINGWINDOW:4:28 MINLEN:36. Quality of raw and trimmed fastq files was assessed with fastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and summarised with multiQC v1.11[98]. Gene expression was quantified from trimmed reads using salmon v1.5.1[99], with options --seqBias --gcBias - validateMappings -l A. Salmon indexes were prepared for each species separately, and used as input (in the -i option) for gene expression quantification in the respective species. DESeq2[100] and custom scripts (available at https://github.com/migueldvalmeida/Cichlid_TEs_piRNAs2024) were used to calculate normalised and TPM counts, generate plots and conduct statistical tests on an R framework[72]. See R packages used below, in the end of this section.

Trimmed fastq files were mapped to the cichlid genomes using HISAT2 v2.2.1[101] with options -x -1 -2 -S. Reads from *A. burtoni*, *P. nyererei* and *O. niloticus* were mapped to their respective Ensembl genomes (AstBur1.0, GCA_000239415.1; PunNye1.0, GCA_000239375.1; O_niloticus_UMD_NMBU, GCA_001858045.3). Reads from all

Lake Malawi cichlid species used (*A. calliptera*, *M. zebra* and *T.* sp. ‘mauve’) were mapped to *A. calliptera* Ensembl genome fAstCal1.2 (GCA_900246225.3). SAM alignment files were converted to BAM format, sorted and indexed with samtools v1.10[76]: 1) samtools view -bS ; 2) samtools sort ; and 3) samtools index. To create bigwig files, the BAM alignment files were used as input to bamCoverage v3.5.1, part of the deepTools package[102], using options --normalizeUsing CPM -of bigwig --binSize 10. Bigwig files of biological replicates of same organ were combined using WiggleTools[103] mean and wigToBigWig v4[104]. Genome tracks were plotted with custom scripts (available at https://github.com/migueldvalmeida/Cichlid_TEs_piRNAs2024) using the Gviz[105] and GenomicFeatures[106] packages on an R framework[72].

To quantify TE expression at the TE family level, we mapped trimmed reads using STAR v2.5.4b[107] with options --readFilesCommand zcat --outSAMtype BAM SortedByCoordinate --outFilterType BySJout --outFilterMultimapNmax 150 --winAnchorMultimapNmax 150 --alignSJoverhangMin 8 --alignSJDBoverhangMin 3 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 10000000 --alignMatesGapMax 100000000. As above, reads from *A. burtoni*, *P. nyererei* and *O. niloticus* were mapped to their respective Ensembl genomes (AstBur1.0, GCA_000239415.1; PunNye1.0, GCA_000239375.1; O_niloticus_UMD_NMBU, GCA_001858045.3) and reads from all Lake Malawi cichlid species used (*A. calliptera*, *M. zebra* and *T.* sp. ‘mauve’) were mapped to *A. calliptera* Ensembl genome fAstCal1.2 (GCA_900246225.3). The resulting BAM files were used as inputs for TEtranscripts v2.2.1[108] with options --stranded reverse --SortByPos. TEtranscripts was run separately for each species, using gene annotations of the respective species downloaded from Ensembl (March 2021) and TE annotations described above (see **Transposable element annotations** section). For Lake Malawi cichlids, TEtranscripts was ran using *A. calliptera* gene and TE annotations (both default and curated versions). A TE family was defined as expressed if it had >10 counts in at least 2 samples. DESeq2[100] and custom scripts (available at https://github.com/migueldvalmeida/Cichlid_TEs_piRNAs2024) were used to calculate normalised counts, generate plots and conduct statistical tests on an R framework[72]. We have used the following R packages: tidyverse[74], lattice[109], eulerr[110], genefilter[111], pheatmap[112], reshape2[113], ggrepel[114], biomaRt[115], tximport[116], RColorBrewer[117], ashr[118], ggpubr[119], GenomicFeatures[106], patchwork[120].

mRNA-sequencing analysis of Lake Malawi cichlid embryogenesis datasets. The embryogenesis dataset collection and experimental design will be reported in detail elsewhere (Chengwei Ulrika Yuan & Eric A. Miska, unpublished results). Trimmomatic-0.39[97] was used to trim the Illumina adapters. Salmon v0.14.2[99] was used to quantify expression of protein-coding genes (--seqBias --validateMappings --gcBias). TEtranscripts analysis on embryo samples was performed as described above (mRNA-sequencing analysis subsection), with one exception: option --stranded no. Locus-specific TE expression levels were analysed with SQuIRE (v0.9.9.9a-beta)[121]. For squire Count the option --strandness ‘0’ was run as default for unstranded Illumina data. Reads were mapped to the *A. calliptera* genome (Ensembl, fAstCal1.2), and the TE annotation created from the curated TE library was used (see above, Transposable element annotations section). Tot_counts was used in downstream analysis from the Squire output. Only expressed TEs were kept (defined as >5 reads in at least 2 samples). Heatmap and enrichment plots were made from SQuIRE output with code adapted from Chang et al., 2022[13].

Small RNA-sequencing analysis. CutAdapt v1.15[122] was used to remove adapters and reads shorter than 18 nucleotides with options -a TGGAATTCTCGGGTGCCAAGG --minimum-length 18. Quality of raw and trimmed fastq files was assessed with fastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and summarised with multiQC v1.11[98]. Next, we mapped the trimmed reads to the genome using STAR v2.5.4b[107], with options readFilesCommand zcat --outMultimapperOrder Random --outFilterMultimapNmax 100 --outFilterMismatchNmax 2 --alignIntronMax 1 --outSAMtype BAM SortedByCoordinate --outFilterType BySJout --winAnchorMultimapNmax 100 --alignEndsType EndToEnd --scoreDelOpen -10000 --scoreInsOpen -10000 --outSAMmultNmax 1 --outFileNamePrefix. As above, reads from *A. burtoni*, *P. nyererei* and *O. niloticus* were mapped to their respective Ensembl genomes (AstBur1.0, GCA_000239415.1; PunNye1.0, GCA_000239375.1; O_niloticus_UMD_NMBU, GCA_001858045.3) and reads from all Lake Malawi cichlid species used (*A. calliptera*, *M. zebra* and *T.* sp. ‘mauve’) were mapped to *A. calliptera* Ensembl genome fAstCal1.2 (GCA_900246225.3). An in-house custom script[123,124] was used, with the BAM files of the alignment as inputs, to create sRNA length distribution profiles in the range of 18-36 nucleotides, and to report 5'-nucleotide frequency, normalised to all mapping reads. The script creates separate sRNA length distribution profiles for 1) collapsed and 2) uncollapsed reads. The first profile keeps only one read of each unique sequence to remove abundance bias, while the second profile keeps all reads. Lastly, the script also produces a FASTA file with the collapsed sequences. With the outputs of the scripts, plots of sRNA length distribution profiles and first nucleotide composition plots were created on an R framework[72] with the packages tidyverse[74], reshape2[113], and RColorBrewer[117].

Next, we selected sRNAs in the piRNA size range, between 24 and 35 nucleotides long, for further analysis. We have done this size selection on the trimmed reads using CutAdapt v1.15[122] with options --minimum-length 24 --maximum-length 35. We mapped 24-35 nucleotides long sRNAs to the genome with the same settings as discriminated in the previous paragraph. Next, we used “Small RNA Signatures” v3.5.0[125] of the Mississippi Tool Suite from the web-based analysis tool Galaxy to calculate z-scores of overlapping sRNA pairs. For this analysis, alignment BAM files of 24-35 nucleotide long reads were used as input, along with following options: Min size of query sRNAs 24, Max size of query sRNAs 35, Min size of target sRNAs 24, Max size of target sRNAs 35, Minimal relative overlap analyzed 1, Maximal relative overlap analyzed 26. To find signatures of phased piRNA biogenesis, BAM files of 24-35 nucleotide long reads were loaded into R as Genomic Ranges[106] and using RSamtools[126], the Follow function was used to identify the next mapping piRNA pair and distances between the 5' and 3' were calculated for plotting. To create sequence logos, we first ran the custom script described above[123,124] to produce a FASTA file with the 24-35 nucleotide long collapsed reads (unique sequences). Then, we created a new FASTA file with all these reads trimmed from the 3' end to a total length of 20 nucleotides, and concatenated together the FASTA files of the biological replicates for each species and organ. The FASTA file with the concatenated and trimmed sequences was in turn used to generate sequence logos in R (scripts available at https://github.com/migueldvalmeida/Cichlid_TEs_piRNAs2024), with packages ggseqlogo[127], phylotools[128], and tidyverse[74]. This process was repeated to generate sequence logos of piRNAs mapping sense or antisense in regard to TE orientation using BAM files with 24-35 nucleotide long reads, which were created as follows: 1) samtools view -b -f (16 or 0); 2) bedtools intersect (-s or -S); 3) samtools merge; 4) samtools sort; 5) samtools index.

To quantify piRNA counts associated with TEs, we used featureCounts v1.6.0[129] with options -t exon -M. The 24-35 nucleotide long BAM file was used as input. The TEtranscripts-compatible TE annotations described above (see **Transposable element annotations**) were provided as the intersecting features. For Lake Malawi cichlids, featureCounts analysis was performed twice, using *A. calliptera* default and curated TE annotations. After obtaining the tables of counts, DESeq2[100] and custom scripts (available at https://github.com/migueldvalmeida/Cichlid_TEs_piRNAs2024) were used to calculate normalised counts, generate plots and conduct statistical tests on an R framework[72], with packages tidyverse[74], lattice[109], eulerr[110], genefilter[111], pheatmap[112], reshape2[113], ggrepel[114], biomaRt[115], tximport[116], RColorBrewer[117], ashr[118], ggpubr[119], GenomicFeatures[106], patchwork[120]. To create bigwig files, the 24-35 nucleotide long BAM alignment files were used as inputs to bamCoverage v3.5.1, part of the deepTools package[102], using options --normalizeUsing CPM -of bigwig --binSize 5. Bigwig files of biological replicates of same organ were combined using WiggleTools[103] mean and wigToBigWig v4[104]. Genome tracks were plotted with custom scripts (available at https://github.com/migueldvalmeida/Cichlid_TEs_piRNAs2024) using the Gviz[105] and GenomicFeatures[106] packages on an R framework[72]. We used these bigwig files to produce sRNA metagene profiles with deepTools[102] computeMatrix scale-regions v3.5.1 (options -b 1000 -a 1000 --regionBodyLength 2000 --averageTypeBins median --missingDataAsZero --binSize 5) and plotProfile v3.5.1 (--plotType se --averageType mean --perGroup). To generate metagene profiles against particular TE classes or superfamilies, TE annotations were subsetted by TE class or superfamily and converted to bed format with grep and awk utilities. The resulting bed files contained the regions to plot and were used as input for computeMatrix.

To define piRNA clusters, we first re-mapped trimmed reads 24-35 nucleotides long to the *A. calliptera* (fAstCal1.2) or *O. niloticus* (O_niloticus_UMD_NMBU) genomes using STAR v2.5.4b[107], with options: --readFilesCommand zcat --outFilterMultimapNmax 100 --outFilterMismatchNmax 2 --alignIntronMax 1 --outSAMtype BAM SortedByCoordinate --outFilterType BySJout --alignSoftClipAtReferenceEnds No --winAnchorMultimapNmax 100 --alignEndsType EndToEnd --scoreDelOpen -10000 --scoreInsOpen -10000 --outSAMmultNmax 100 --outSAMattributes All. Method and code for the approach below will be detailed elsewhere (A. Friman and A. Haase, unpublished results). The resulting BAM files were loaded into R environment using GenomicAlignments package[106]. For each BAM file the alignments were sorted into three categories: unique mapping alignments, primary multimapping alignments, and secondary multimapping alignments (method by A. Friman and A. Haase, unpublished results). The reference genome was split into sliding windows[106] with size and step between starting position depending on the alignments category. For unique mapping alignments the windows were 350 nt (window size) starting at every 35 nt (window step) of genome length. For each of these windows the number of overlapping unique mapping alignments was counted. If the number was at least 2 FPKM (RPKM), the window was called. The called windows were reduced into genomic intervals named “seeds”, indicating the genomic origin of uniquely mapping piRNAs. Seeds that were shorter than 800 nt were discarded to reduce false positives, which can be caused by individual degradation fragments of abundant structural RNAs or other cellular transcripts. Next, we incorporated multimapping piRNA reads, considering first their primary alignments and then all possible alignments (up to 1000 according to the parameters used for genome mapping). We counted primary multimapping alignments using 350 nt long sliding windows (window size) located at every 35 nt (window step) of genome length. Windows overlapping with more than 4 FPKM (RPKM) with each other and with previously established ‘seeds’, were reduced into intervals named ‘cores’. Each ‘core’ was required to overlap with at least one seed. Finally, we integrated all secondary multimapping alignments using 1000 nt long sliding windows (window size) with 100 nt step (window step). We requested read coverage greater or equal 0.2 FPKM (RPKM) as threshold. Overlapping windows were reduced into 'clusters' when they overlapped with at least one 'core'. All clusters contain strand information and predict one or multiple piRNA precursor transcripts from a defined genomic strand. Intersection[106] of genomic 'cluster' coordinates from different samples or biological replicates take strand information into account. Results were plotted on a R framework[72], using packages: tidyverse (Wickham et al., 2019), reshape2[113], and ggpubr[119]. Circos plots were created with Circos v0.69-8[130]. Density tracks are displayed on the circos plots as the number of features per mega-base.

Protein preparations and mass spectrometry

Frozen brain and gonad tissues were partitioned on a mortar positioned on dry ice, and weighed. This was done quickly to avoid thawing. A similar mass of the same tissue was collected from size-matched individuals of the same species to create biological replicates. 6-50 mgs of brain tissue, and 8-120 mgs of gonad tissue were used, according to the specific tissue, tissue availability, and size of the specimen, which varied per species. Partitioned tissues were transferred to BeadBug tubes prefilled with 0.5 mm Zirconium beads (Merck, #Z763772) together with 150 µl (if using 6-20 mg of tissue) or 250 µl (if using >20 mg of tissue) of modified RIPA buffer (50 mM Tris HCl pH 7.5, 150 mM NaCl, 1% IGEPAL CA-630, 1% Sodium Deoxycholate, supplemented with cOmplete EDTA-free protease inhibitor cocktail tablets, Roche #4693132001). Next, homogenisation was conducted using a BeadBug microtube homogeniser (Sigma, #Z764140) at approximately 4°C (conducted in cold room). Each sample was homogenised with five BeadBug runs at maximum speed (4,000 rpm) for 60 seconds each. Did not run any sample more than two consecutive times to avoid overheating. Other than the run time inside the BeadBug, samples were left on ice. After homogenisation, lysates were centrifuged for 5 minutes at 18,000 G at 4°C. Supernatant was then removed into a clean 1.5 mL tube. Centrifuge the lysates again, this time at maximum speed (approximately 21,000 G) for 5 minutes at 4°C. Transfer supernatant into a clean tube without disturbing the pellet and tissue debris. Measured protein concentration using Bradford (Bio-Rad, Protein Assay Dye Reagent Concentrate, #5000006) and prepared a final sample by combining 150 µg of lysate, 1x LDS (prepared from NuPAGE LDS Sample Buffer 4x, Thermo Scientific, #NP0007) and 100 mM DTT and boiling for 10 minutes at 95°C. Half of the sample was sent for mass spectrometry.

In-gel digestion for mass spectrometry was performed as previously described[131]. Samples were boiled at 70°C for 10 minutes prior to loading on a 4%-12% NuPAGE Bis-Tris gel (Thermo Scientific, #NP0321). The gel was run in 1x MOPS buffer at 180V for 10 minutes and subsequently fixed and stained with Coomassie G250 (Carl Roth). Each lane was minced and transferred to a 1.5 mL reaction tube, destained with 50% EtOH in 50 mM ammonium bicarbonate buffer (pH 8.0). Gel pieces were dehydrated with 100% acetonitrile and dried in a Concentrator Plus (Eppendorf, #5305000304). Then, samples were reduced with 10 mM DTT / 50 mM ABC buffer (pH 8.0) at 56°C and alkylated with 50 mM iodoacetamide / 50 mM ABC buffer (pH 8.0) in the dark. After washing with ABC buffer (pH 8.0) and dehydration with acetonitrile the proteins were digested with 1 µg mass spectrometry-grade Trypsin (Serva) at 37°C overnight. The peptides were purified on stage tips as previously described[132]. Peptides were analysed by nanoflow liquid chromatography using an EASYnLC 1200 system (Thermo Scientific) coupled to an Exploris 480 (Thermo Scientific). Peptides were separated on a C18-reversed phase column (60 cm, 75µm diameter), packed in-house with Reprosil aq1.9 (Dr. Maisch GmbH), mounted on the electrospray ion source of the mass spectrometer. Peptides were eluted from the column with an optimized 103-min gradient from 2% to 40% of a mixture of 80% acetonitrile/0.1% formic acid at a flow rate of 250 nL/min. The Exploris was operated in positive ion mode with a data-dependent acquisition strategy of one mass spectrometry full scan (scan range 300–1650 m/z; 60,000 resolution; normalised AGC target 300%; max IT 28 ms) and up to 20 MS/MS scans (15,000 resolution; AGC target 100%, max IT 28 ms; isolation window 1.4 m/z) with peptide match preferred using HCD fragmentation. Mass spectrometry measurements were analysed with MaxQuant v1.6.10.43[133] with the following protein databases (downloaded from Ensembl): Haplochromis_burtoni.AstBur1.0.pep.all.fa (35,619 entries, from *A. burtoni*), Oreochromis_niloticus.O_niloticus_UMD_NMBU.pep.all.fa (75,555 entries, from *O. niloticus*), Astatotilapia_calliptera.fAstCal1.2.pep.all.fa (41,597 entries, from *A. calliptera*), and Pundamilia_nyererei.PunNye1.0.pep.all.fa (32,153 entries, from *P. nyererei*). Missing values were imputed at the lower end of LFQ values using random values from a beta distribution fitted at 0.2-2.5%. Prior to further analysis, protein groups with contaminants, reverse hits and only identified by site were removed.

Data accessibility

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[134] partner repository with the dataset identifier PXD047439. The mRNA and sRNA sequencing data generated in this study have been deposited to GEO under accession numbers GSE252804 and GSE252805. The genomic data of Lake Malawi cichlids used in this work is available on SRA, bioproject PRJEB1254 (see a list of samples in **Supplemental Table 1**), on an open access basis for research use only. Any person who wishes to use this data for any form of commercial purpose must first enter into a commercial licensing and benefit sharing arrangement with the Government of Malawi.

Acknowledgements

We are grateful to all members of the Miska, Durbin, and Santos groups for discussion and suggestions. We are especially grateful to Chenxi Zhou for input on confirmation of *piwil1* duplications, and Navin B. Ramakrishna for analytical input and critically reading the manuscript. We thank all who provided fish or fish samples: Kevs Rifts for *M. zebra* fishes; Hans Hofmann and Caitlin Friesen (University of Texas at Austin, TX, USA) for *A. burtoni* samples; David Penman, Alastair McPhee, and James F. Turnbull (Institute of Aquaculture, University of Stirling, Scotland, UK) for *O. niloticus* samples. We are thankful to the animal caretakers of the cichlid facility of the University of Cambridge for diligently ensuring the maintenance of our *A. calliptera* and *T.* sp. 'mauve' stocks. The Malawi cichlid samples were collected ethically under prescribed permits, and the results and data are published under an Access and Benefit Sharing agreement with the Government of Malawi. We acknowledge the contributions of the Malawi Department of Fisheries and the people and Government of Malawi for their assistance in the collection of samples and the generation of data and results. We thank Maike Paramor and Vicki Murray (Wellcome-MRC Cambridge Stem Cell Institute) for NGS library preparation. For the purpose of open access, the lead author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Funding

M.V.A. is funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101027241. M.B. is funded by a Harding Distinguished Postgraduate Scholarship of the University of Cambridge. C.U.Y., F.X.Q., and A.L.K.P. were funded by the Cambridge Commonwealth, European and International Trust. P.S. is funded by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 956229 (ALPACA). F.X.Q. is supported by a Wellcome studentship (108864/B/15/Z). A.D. is supported by the Royal Botanic Gardens Kew. G.V. acknowledges funding from Wolfson College - University of Cambridge and the Genetics Society London. A.D.H.’s research group is supported by the intramural research program of the National Institute of Diabetes and Digestive and Kidney Diseases (ZIA DK075111-07). R.D. is supported by a Wellcome Investigator Award (207492/Z/17/Z). M.E.S is supported by NERC IRF NE/R01504X/1. This work was supported by the following grants to E.A.M.: Wellcome Trust Senior Investigator Award (219475/Z/19/Z) and CRUK award (C13474/A27826). The authors also acknowledge core funding to the Gurdon Institute from Wellcome (092096/Z/10/Z, 203144/Z/16/Z) and CRUK (C6946/A24843).

Author contributions

Conceptualisation: M.V.A. and E.A.M.; Data curation: M.V.A., M.B., and J.L.P.; Formal analysis: M.V.A., M.B., C.U.Y., P.S., J.L.P., and F.X.Q.; Funding acquisition: M.V.A., A.D.H., R.D. and E.A.M.; Investigation: M.V.A., and C.U.Y.; Project administration: M.V.A. and E.A.M.; Resources: P.S., A.F., A.D., G.V., A.L.K.P., A.M.S., D.A.J., F.B., A.D.H., R.D., M.E.S.; Software: A.F. and A.D.H.; Supervision: M.V.A., A.D.H., R.D., M.E.S., and E.A.M; Visualisation: M.V.A., M.B., C.U.Y., and J.L.P.; Writing – original draft: M.V.A.; Writing – review & editing: all authors contributed.

Competing interests

The authors declare no competing interests.

References

1. Salzburger, W. Understanding explosive diversification through cichlid fish genomics. Nature Reviews Genetics 19, 705 (2018).

2. Santos, M. E., Lopes, J. F. & Kratochwil, C. F. East African cichlid fishes. EvoDevo 14, 1 (2023).

3. Svardal, H., Salzburger, W. & Malinsky, M. Genetic Variation and Hybridization in Evolutionary Radiations of Cichlid Fishes. Annual Review of Animal Biosciences 9, 55–79 (2021).

4. Malinsky, M. et al. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nature Ecology & Evolution 2, 1940 (2018).

5. Quah, F. X. et al. A pangenomic perspective of the Lake Malawi cichlid radiation reveals extensive structural variation driven by transposable elements. 2024.03.28.587230 Preprint at https://doi.org/10.1101/2024.03.28.587230 (2024).

6. Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. Annual Review of Genetics 54, 539–561 (2020).

7. Arkhipova, I. R. Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution. Molecular Biology and Evolution 35, 1332–1337 (2018).

8. Almeida, M. V., Vernaz, G., Putman, A. L. K. & Miska, E. A. Taming transposable elements in vertebrates: from epigenetic silencing to domestication. Trends in Genetics (2022) doi:10.1016/j.tig.2022.02.009.

9. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics 18, 71–86 (2017).

10. Fueyo, R., Judd, J., Feschotte, C. & Wysocka, J. Roles of transposable elements in the regulation of mammalian transcription. Nat Rev Mol Cell Biol 1–17 (2022) doi:10.1038/s41580-022-00457-y.

11. Carducci, F., Barucca, M., Canapa, A., Carotti, E. & Biscotti, M. A. Mobile Elements in Ray-Finned Fish Genomes. Life 10, 221 (2020).

12. Chalopin, D., Naville, M., Plard, F., Galiana, D. & Volff, J.-N. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. Genome Biology and Evolution 7, 567–580 (2015).

13. Chang, N.-C., Rovira, Q., Wells, J. N., Feschotte, C. & Vaquerizas, J. M. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. Genome Res. gr.275655.121 (2022) doi:10.1101/gr.275655.121.

14. Gao, B. et al. The contribution of transposable elements to size variations between four teleost genomes. Mobile DNA 7, 4 (2016).

15. Reinar, W. B. et al. Teleost genomic repeat landscapes in light of diversification rates and ecology. Mobile DNA 14, 14 (2023).

16. Shao, F., Han, M. & Peng, Z. Evolution and diversity of transposable elements in fish genomes. Sci Rep 9, 15399 (2019).

17. Yuan, Z. et al. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics 19, 141 (2018).

18. Brawand, D. et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature 513, 375–381 (2014).

19. Kratochwil, C. F. et al. An intronic transposon insertion associates with a trans-species color polymorphism in Midas cichlid fishes. Nat Commun 13, 296 (2022).

20. Santos, M. E. et al. The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. Nature Communications 5, 5149 (2014).

21. Munby, H. et al. Differential use of multiple genetic sex determination systems in divergent ecomorphs of an African crater lake cichlid. bioRxiv https://doi.org/10.1101/2021.08.05.455235, 2021.08.05.455235 (2021).

22. Carleton, K. L. et al. Movement of transposable elements contributes to cichlid diversity. Molecular Ecology 29, 4956–4969 (2020).

23. Vernaz, G. et al. Mapping epigenetic divergence in the massive radiation of Lake Malawi cichlid fishes. Nat Commun 12, 5870 (2021).

24. Bruno, M., Mahgoub, M. & Macfarlan, T. S. The Arms Race Between KRAB–Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. Annual Review of Genetics 53, 393–416 (2019).

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

26. Loubalova, Z., Konstantinidou, P. & Haase, A. D. Themes and variations on piRNA-guided transposon control. Mobile DNA 14, 10 (2023).

27. Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. & Zamore, P. D. PIWI-interacting RNAs: small RNAs with big functions. Nat Rev Genet 20, 89–108 (2019).

28. Seczynska, M., Bloor, S., Cuesta, S. M. & Lehner, P. J. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. Nature 1–9 (2021) doi:10.1038/s41586-021-04228-1.

29. Tchasovnikarova, I. A. et al. Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. Science 348, 1481–1485 (2015).

30. Wells, J. N. et al. Transposable elements drive the evolution of metazoan zinc finger genes. Genome Res. (2023) doi:10.1101/gr.277966.123.

31. Aravin, A. et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442, 203–207 (2006).

32. Brennecke, J. et al. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell 128, 1089–1103 (2007).

33. Houwing, S. et al. A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. Cell 129, 69–82 (2007).

34. Gainetdinov, I., Colpan, C., Arif, A., Cecchini, K. & Zamore, P. D. A Single Mechanism of Biogenesis, Initiated and Directed by PIWI Proteins, Explains piRNA Production in Most Animals. Molecular Cell 71, 775-790.e5 (2018).

35. Gunawardane, L. S. et al. A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5' End Formation in Drosophila. Science 315, 1587–1590 (2007).

36. Han, B. W., Wang, W., Li, C., Weng, Z. & Zamore, P. D. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science 348, 817–821 (2015).

37. Mohn, F., Handler, D. & Brennecke, J. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. Science 348, 812–817 (2015).

38. Lewis, S. H. et al. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. Nat Ecol Evol 2, 174–181 (2018).

39. Lewis, S. H., Salmela, H. & Obbard, D. J. Duplication and Diversification of Dipteran Argonaute Genes, and the Evolutionary Divergence of Piwi and Aubergine. Genome Biol Evol 8, 507–518 (2016).

40. Palmer, W. H., Hadfield, J. D. & Obbard, D. J. RNA-Interference Pathways Display High Rates of Adaptive Protein Evolution in Multiple Invertebrates. Genetics 208, 1585–1599 (2018).

41. Parhad, S. S., Tu, S., Weng, Z. & Theurkauf, W. E. Adaptive Evolution Leads to Cross-Species Incompatibility in the piRNA Transposon Silencing Machinery. Developmental Cell 43, 60-70.e5 (2017).

42. Fryer, G. & Iles, T. D. The Cichlid Fishes of the Great Lakes of Africa: Their Biology and Evolution. (Oliver and Boyd, Edinburgh, 1972).

43. Marconi, A., Yang, C. Z., McKay, S. & Santos, M. E. Morphological and temporal variation in early embryogenesis contributes to species divergence in Malawi cichlid fishes. Evolution & Development 25, 170–193 (2023).

44. Wang, X., Ramat, A., Simonelig, M. & Liu, M.-F. Emerging roles and functional mechanisms of PIWI-interacting RNAs. Nat Rev Mol Cell Biol 24, 123–141 (2023).

45. Li, M. A. et al. Mobilization of giant piggyBac transposons in the mouse genome. Nucleic Acids Research 39, e148 (2011).

46. Li, X. et al. piggyBac transposase tools for genome engineering. Proceedings of the National Academy of Sciences 110, E2279–E2287 (2013).

47. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12, 2 (2021).

48. Cary, L. C. et al. Transposon mutagenesis of baculoviruses: Analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology 172, 156–169 (1989).

49. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics 11, 97–108 (2010).

50. Matsumoto, N. et al. Crystal Structure of Silkworm PIWI-Clade Argonaute Siwi Bound to piRNA. Cell 167, 484-497.e9 (2016).

51. Yamaguchi, S. et al. Crystal structure of Drosophila Piwi. Nat Commun 11, 858 (2020).

52. Parker, J. S., Roe, S. M. & Barford, D. Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. The EMBO Journal 23, 4727–4737 (2004).

53. Yuan, Y.-R. et al. Crystal Structure of A. aeolicus Argonaute, a Site-Specific DNA-Guided Endoribonuclease, Provides Insights into RISC-Mediated mRNA Cleavage. Molecular Cell 19, 405–419 (2005).

54. Stein, C. B. et al. Decoding the 5′ nucleotide bias of PIWI-interacting RNAs. Nat Commun 10, 828 (2019).

55. Irisarri, I. et al. Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. Nat Commun 9, 3159 (2018).

56. Meier, J. I. et al. Cycles of fusion and fission enabled rapid parallel adaptive radiations in African cichlids. Science 381, eade2833 (2023).

57. Meier, J. I. et al. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat Commun 8, 14363 (2017).

58. Svardal, H. et al. Ancestral Hybridization Facilitated Species Diversification in the Lake Malawi Cichlid Fish Adaptive Radiation. Molecular Biology and Evolution 37, 1100–1113 (2020).

59. Brawand, D. et al. The evolution of gene expression levels in mammalian organs. Nature 478, 343–348 (2011).

60. Soumillon, M. et al. Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Reports 3, 2179–2190 (2013).

61. Sakashita, A. et al. Endogenous retroviruses drive species-specific germline transcriptomes in mammals. Nat Struct Mol Biol 27, 967–977 (2020).

62. Lee, H. J. et al. Epigenomic analysis reveals prevalent contribution of transposable elements to cis-regulatory elements, tissue-specific expression, and alternative promoters in zebrafish. Genome Res. 32, 1424–1436 (2022).

63. El Taher, A. et al. Gene expression dynamics during rapid organismal diversification in African cichlid fishes. Nature Ecology & Evolution 1–8 (2020) doi:10.1038/s41559-020-01354-3.

64. Yi, M. et al. Rapid Evolution of piRNA Pathway in the Teleost Fish: Implication for an Adaptation to Transposon Diversity. Genome Biology and Evolution 6, 1393–1407 (2014).

65. Frank, F., Sonenberg, N. & Nagar, B. Structural basis for 5′-nucleotide base-specific recognition of guide RNA by human AGO2. Nature 465, 818–822 (2010).

66. Song, J.-J., Smith, S. K., Hannon, G. J. & Joshua-Tor, L. Crystal Structure of Argonaute and Its Implications for RISC Slicer Activity. Science 305, 1434–1437 (2004).

67. Yan, K. S. et al. Structure and conserved RNA binding of the PAZ domain. Nature 426, 469–474 (2003).

68. Genner, M. J. Cichlid fish seized an ecological opportunity to diversify. Nature (2023) doi:10.1038/d41586-023-03014-5.

69. Ngoepe, N. et al. A continuous fish fossil record reveals key insights into adaptive radiation. Nature 1–6 (2023) doi:10.1038/s41586-023-06603-6.

70. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20, 238 (2019).

71. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16, 157 (2015).

72. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2021).

73. Brennan, P. drawProteins: a Bioconductor/R package for reproducible and programmatic generation of protein schematics. Preprint at https://doi.org/10.12688/f1000research.14541.1 (2018).

74. Wickham, H. et al. Welcome to the Tidyverse. Journal of Open Source Software 4, 1686 (2019).

75. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

76. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

77. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).

78. Robinson, J. T. et al. Integrative genomics viewer. Nat Biotechnol 29, 24–26 (2011).

79. Martin, F. J. et al. Ensembl 2023. Nucleic Acids Research 51, D933–D941 (2023).

80. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. Nat Methods 19, 705–710 (2022).

81. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792–1797 (2004).

82. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30, 3276–3278 (2014).

83. Danecek, P. et al. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008 (2021).

84. Alachiotis, N. & Pavlidis, P. RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. Commun Biol 1, 1–11 (2018).

85. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30, 772–780 (2013).

86. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191 (2009).

87. Minh, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37, 1530–1534 (2020).

88. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35, 518–522 (2018).

89. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research 34, W609–W612 (2006).

90. Weaver, S. et al. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Molecular Biology and Evolution 35, 773–777 (2018).

91. Murrell, B. et al. Gene-Wide Identification of Episodic Selection. Molecular Biology and Evolution 32, 1365–1371 (2015).

92. Murrell, B. et al. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLOS Genetics 8, e1002764 (2012).

93. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

94. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research 50, D439–D444 (2022).

95. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).

96. Goubert, C. et al. A beginner's guide to manual curation of transposable elements. Mobile DNA 13, 7 (2022).

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

98. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048 (2016).

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

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

101. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11, 1650–1667 (2016).

102. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 44, W160–W165 (2016).

103. Zerbino, D. R., Johnson, N., Juettemann, T., Wilder, S. P. & Flicek, P. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. Bioinformatics 30, 1008–1009 (2014).

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

105. Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. in Statistical Genomics: Methods and Protocols (eds. Mathé, E. & Davis, S.) 335–351 (Springer, New York, NY, 2016). doi:10.1007/978-1-4939-3578-9_16.

106. Lawrence, M. et al. Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology 9, e1003118 (2013).

107. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).

108. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics 31, 3593–3599 (2015).

109. Sarkar, D. et al. lattice: Trellis Graphics for R. (2023).

110. Larsson, J. et al. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. (2022).

111. Gentleman, R. et al. genefilter: genefilter: methods for filtering genes from high-throughput experiments. Bioconductor version: Release (3.17) https://doi.org/10.18129/B9.bioc.genefilter (2023).

112. Kolde, R. pheatmap: Pretty Heatmaps. (2019).

113. Wickham, H. reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package. (2020).

114. Slowikowski, K. et al. ggrepel: Automatically Position Non-Overlapping Text Labels with ‘ggplot2’. (2023).

115. Durinck, S. et al. biomaRt: Interface to BioMart databases (i.e. Ensembl). Bioconductor version: Release (3.17) https://doi.org/10.18129/B9.bioc.biomaRt (2023).

116. Love, M. et al. tximport: Import and summarize transcript-level estimates for transcript- and gene-level analysis. Bioconductor version: Release (3.17) https://doi.org/10.18129/B9.bioc.tximport (2023).

117. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2022).

118. Stephens, M. et al. ashr: Methods for Adaptive Shrinkage, using Empirical Bayes. (2023).

119. Kassambara, A. ggpubr: ‘ggplot2’ Based Publication Ready Plots. (2023).

120. Pedersen, T. L. patchwork: The Composer of Plots. (2023).

121. Yang, W. R., Ardeljan, D., Pacyna, C. N., Payer, L. M. & Burns, K. H. SQuIRE reveals locus-specific regulation of interspersed repeat expression. Nucleic Acids Research 47, e27 (2019).

122. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12 (2011).

123. Di Domenico, T. tstk/peterplot.py · master · Tomás Di Domenico / tstk · GitLab. GitLab https://gitlab.com/tdido/tstk/-/blob/master/tstk/peterplot.py (2022).

124. Ramakrishna, N. B., Battistoni, G., Surani, M. A., Hannon, G. J. & Miska, E. A. Mouse primordial germ-cell-like cells lack piRNAs. Developmental Cell 57, 2661-2668.e5 (2022).

125. Antoniewski, C. Computing siRNA and piRNA Overlap Signatures. in Animal Endo-SiRNAs: Methods and Protocols (ed. Werner, A.) 135–146 (Springer, New York, NY, 2014). doi:10.1007/978-1-4939-0931-5_12.

126. Morgan, M., Pagès, H., Obenchain, V. & Hayden, N. Rsamtools. Bioconductor http://bioconductor.org/packages/Rsamtools/ (2023).

127. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics 33, 3645–3647 (2017).

128. Zhang, J. helixcn/phylotools. (2023).

129. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014).

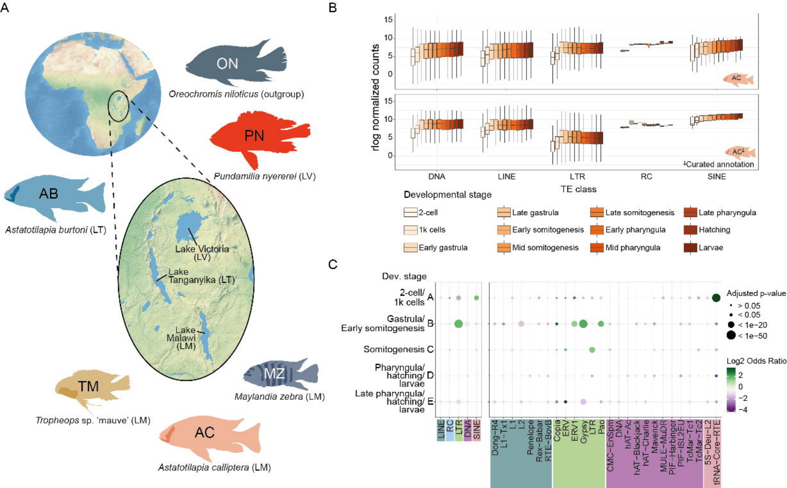
130. Krzywinski, M. I. et al. Circos: An information aesthetic for comparative genomics. Genome Res. (2009) doi:10.1101/gr.092759.109.

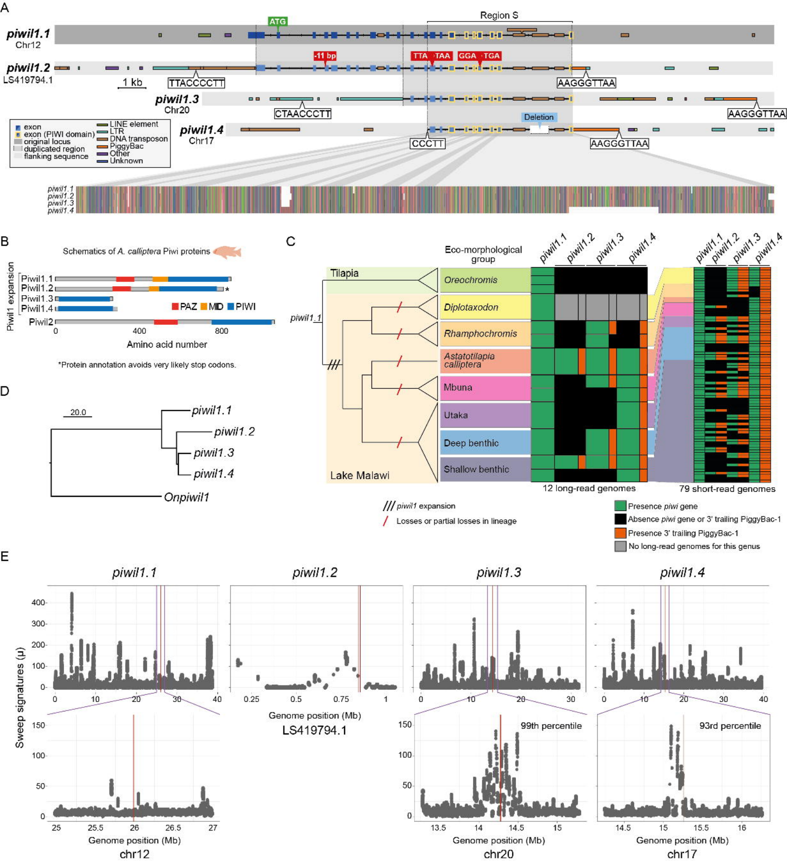
131. Shevchenko, A., Tomas, H., Havli, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat. Protocols 1, 2856–2860 (2007).

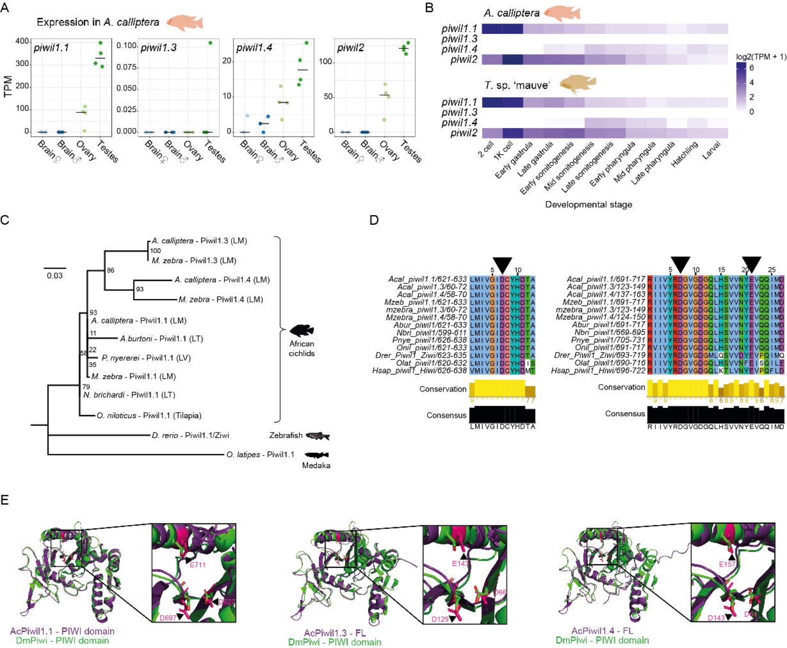
132. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat. Protocols 2, 1896–1906 (2007).

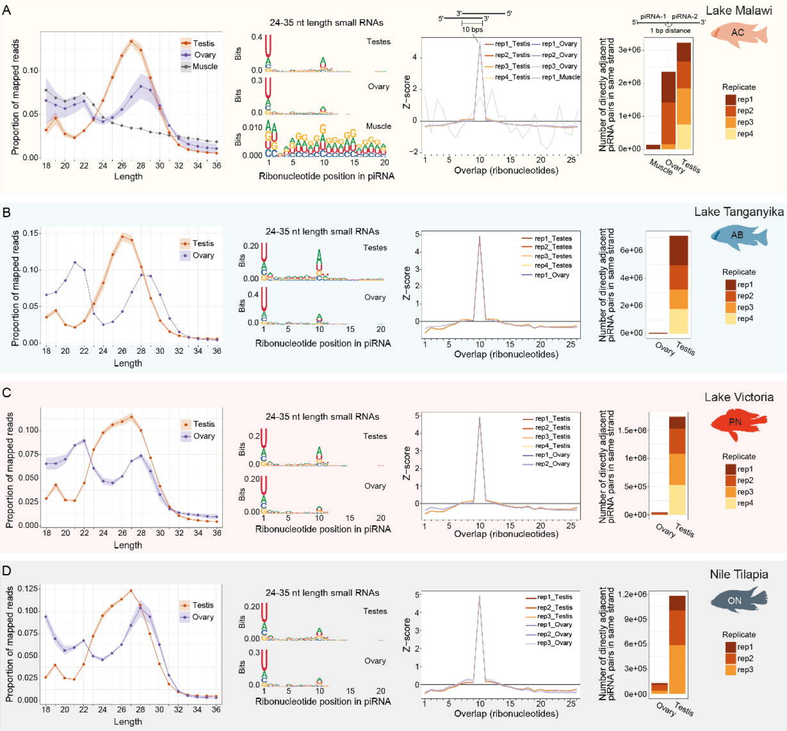
133. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotech 26, 1367–1372 (2008).

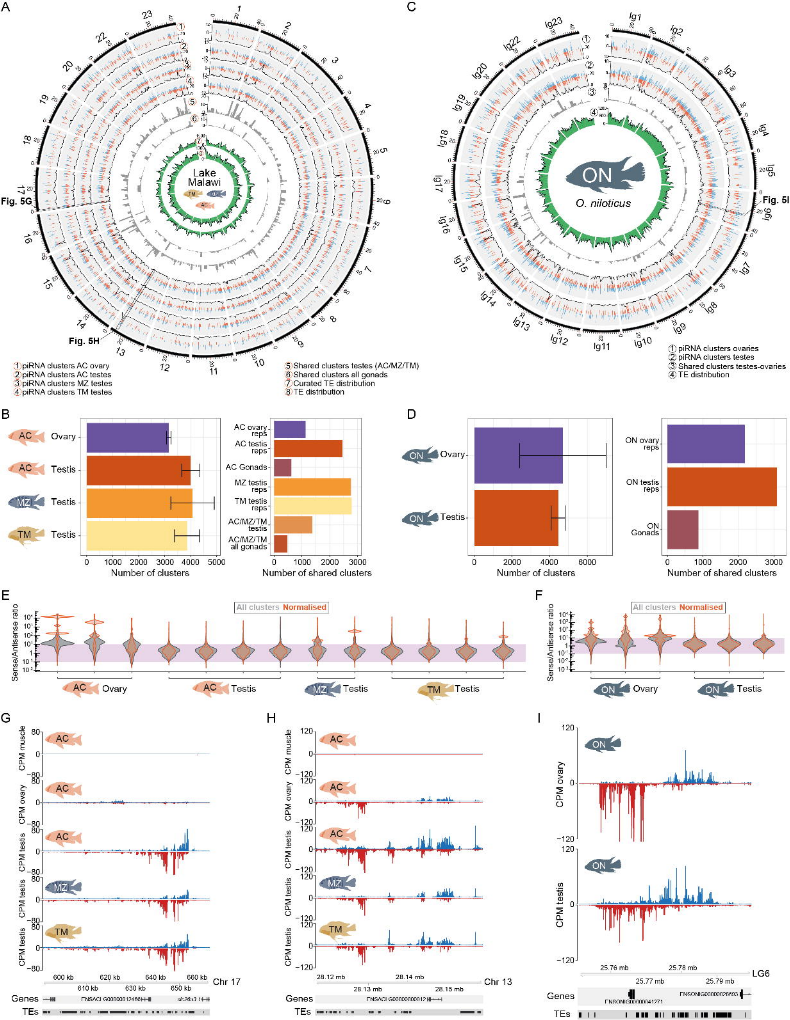
134. Perez-Riverol, Y. et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Research 50, D543–D552 (2022).

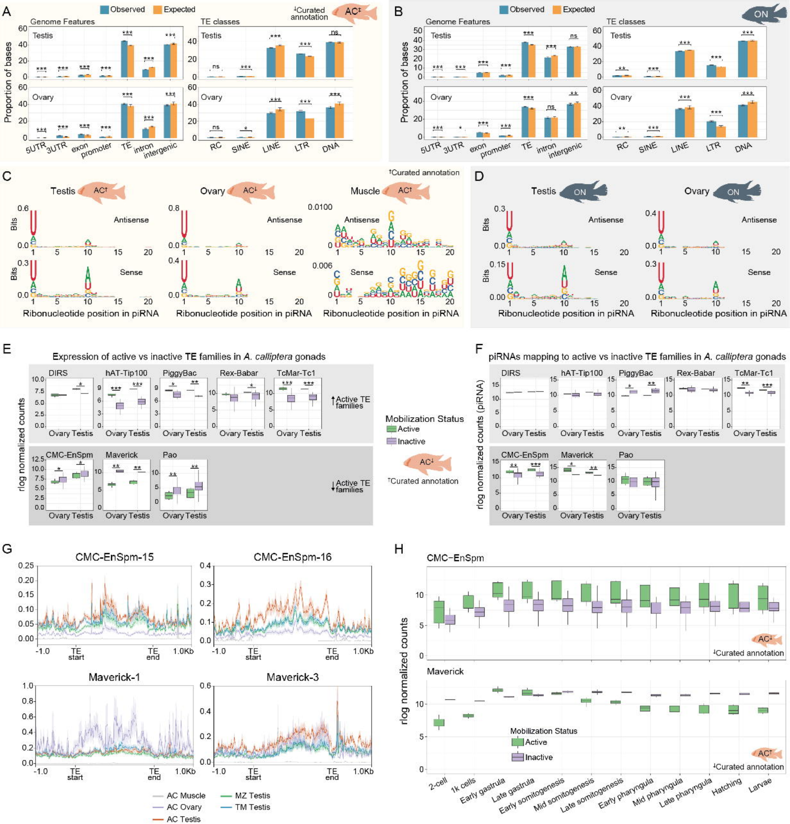

A
ON
Oreochromis niloticus (outgroup)
PN
Pundamilia nyererei (LV)
AB
Astatotilapia burtoni (LT)
Lake Victoria (LV)
Lake Tanganyika (LT)
Lake Malawi (LM)
MZ
Maylandia zebra (LM)
TM
Tropheops sp. 'mauve' (LM)
AC
Astatotilapia calliptera (LM)
B
rlog normalized counts
15
10
5
0
AC
AC‡
‡Curated annotation
DNA
LINE
LTR
RC
SINE
TE class
Developmental stage
2-cell
1k cells
Early gastrula
Late gastrula
Early somitogenesis
Mid somitogenesis
Late somitogenesis
Early pharyngula
Mid pharyngula
Late pharyngula
Hatching
Larvae
C
Dev. stage
2-cell/ 1k cells A
Gastrula/ Early somitogenesis B
Somitogenesis C
Pharyngula/ hatching/ larvae D
Late pharyngula/ hatching/ larvae E
LINE
RC
LTR
DNA
SINE
Dong-R4
L1-Tx1
L1
L2
Penelope
Rex-Babar
RTE-BovB
Copia
ERV
ERV1
Gypsy
LTR
Pao
CMC-EnSpm
DNA
hAT-Ac
hAT-Blackjack
hAT-Charlie
Maverick
MULE-MuDR
PIF-Harbinger
PIF-ISL2EU
TcMar-Tc1
TcMar-Tc2
5S-Deu-L2
tRNA-Core-RTE
Adjusted p-value
> 0.05
< 0.05
< 1e−20
< 1e−50
Log2 Odds Ratio
2
0
−2
−4


A
Region S
ATG
piwil1.1
Chr12
piwil1.2
LS419794.1
-11 bp
TTA→TAA
GGA→TGA
1 kb
TTACCCCTT
AAGGGTTAA
piwil1.3
Chr20
CTAACCCTT
AAGGGTTAA
Deletion
piwil1.4
Chr17
CCCTT
AAGGGTTAA
exon
exon (PIWI domain)
original locus
duplicated region
flanking sequence
LINE element
LTR
DNA transposon
PiggyBac
Other
Unknown
piwil1.1
piwil1.2
piwil1.3
piwil1.4
B
Schematics of A. calliptera Piwi proteins
Piwi1 expansion
Piwi1.1
Piwi1.2
Piwi1.3
Piwi1.4
Piwi2
PAZ
MID
PIWI
*
0
400
800
Amino acid number
*Protein annotation avoids very likely stop codons.
C
Eco-morphological group
piwil1.1
piwil1.2
piwil1.3
piwil1.4
Tilapia
Oreochromis
Diplotaxodon
Rhamphochromis
Astatotilapia calliptera
Mbuna
Utaka
Deep benthic
Shallow benthic
Lake Malawi
piwil1.1
12 long-read genomes
79 short-read genomes
/// piwil1 expansion
/ Losses or partial losses in lineage
Presence piwi gene
Absence piwi gene or 3' trailing PiggyBac-1
Presence 3' trailing PiggyBac-1
No long-read genomes for this genus
D
20.0
piwil1.1
piwil1.2
piwil1.3
piwil1.4
Onpiwil1
E
piwil1.1
piwil1.2
piwil1.3
piwil1.4
Sweep signatures (μ)
400
300
200
100
0
0
10
20
30
40
0.25
0.5
0.75
1
Genome position (Mb)
LS419794.1
0
10
20
30
0
10
20
30
40
150
100
50
0
25
25.5
26
26.5
27
Genome position (Mb)
chr12
150
100
50
0
13.5
14
14.5
15
Genome position (Mb)
chr20
99th percentile
14.5
15
15.5
16
Genome position (Mb)
chr17
93rd percentile


A
Expression in A. calliptera
piwil1.1
piwil1.3
piwil1.4
piwil2
TPM
0
100
200
300
400
0.000
0.025
0.050
0.075
0.100
0
10
20
0
50
100
Brain♀
Brain♂
Ovary
Testes
B
A. calliptera
T. sp. 'mauve'
piwil1.1
piwil1.3
piwil1.4
piwil2
log2(TPM + 1)
6
4
2
0
2 cell
1K cell
Early gastrula
Late gastrula
Early somitogenesis
Mid somitogenesis
Late somitogenesis
Early pharyngula
Mid pharyngula
Late pharyngula
Hatchling
Larval
Developmental stage
C
0.03
A. calliptera - Piwil1.3 (LM)
M. zebra - Piwil1.3 (LM)
A. calliptera - Piwil1.4 (LM)
M. zebra - Piwil1.4 (LM)
A. calliptera - Piwil1.1 (LM)
A.burtoni - Piwil1.1 (LT)
P. nyererei - Piwil1.1 (LV)
M. zebra - Piwil1.1 (LM)
N. brichardi - Piwil1.1 (LT)
O. niloticus - Piwil1.1 (Tilapia)
D. rerio - Piwil1.1/Ziwi
O. latipes - Piwil1.1
100
86
93
93
11
22
58
35
79
African cichlids
Zebrafish
Medaka
D
5
10
Acal_piwil1.1/621-633 LMIVGIDCYHDTA
Acal_piwil1.3/60-72 LMIVGIDCYHDTA
Acal_piwil1.4/58-70 LMIVGIDCYHDTA
Mzeb_piwil1.1/621-633 LMIVGIDCYHDTA
mzebra_piwil1.3/60-72 LMIVGIDCYHDTA
Mzebra_piwil1.4/58-70 LMIVGIDCYHDTA
Abur_piwil1/621-633 LMIVGIDCYHDTA
Nbri_piwil1/599-611 LMIVGIDCYHDTA
Pnye_piwil1/626-638 LMIVGIDCYHDTA
Onil_piwil1/621-633 LMIVGIDCYHDTA
Drer_Piwil1_Ziwi/623-635 LMIVGIDCYHDTA
Olat_piwil1/620-632 LMIVGIDCYHDIS
Hsap_piwil1_Hiwi/626-638 VMIVGIDCYHDMT
Conservation
9 * * * * * * * * * * 7 7
Consensus
LMIVGIDCYHDTA
5
10
15
20
25
Acal_piwil1.1/691-717 RIIVYRDGVGDGQLHSVVNYEVQQIMD
Acal_piwil1.3/123-149 RIIVYRDGVGDGQLHSVVNYEVQQIMD
Acal_piwil1.4/137-163 RIIVYRDGVGDGQLHSVVNYEVQQIMD
Mzeb_piwil1.1/691-717 RIIVYRDGVGDGQLHSVVNYEVQQIMD
mzebra_piwil1.3/123-149 RIIVYRDGVGDGQLHSVVNYEVQQIMD
Mzebra_piwil1.4/124-150 RIIVYRDGVGDGQLHSVVNYEVQQIMD
Abur_piwil1/691-717 RIIVYRDGVGDGQLHSVVNYEVQQIMD
Nbri_piwil1/669-695 RIIVYRDGVGDGQLHSVVNYEVQQIMD
Pnye_piwil1/705-731 RIIVYRDGVGDGQLHSVVNYEVQQIMD
Onil_piwil1/691-717 RVIVYRDGVGDGQLHSVVNYEVQQIMD
Drer_Piwil1_Ziwi/693-719 RIIVYRDGVGDGMLQSVVDYEVPQIMQ
Olat_piwil1/690-716 RIIVYRDGVGDGQLHSVVNFEISQILE
Hsap_piwil1_Hiwi/696-722 RIIVYRDGVGDGQLKTLVNYEVPQFLD
Conservation
*9*9* * * * * * *8*6 8 9* * 8 9* 9 6* 8 9 7
Consensus
RIIVYRDGVGDGQLHSVVNYEVQQIMD
E
E711
D627
D697
AcPiwil1.1 - PIWI domain
DmPiwi - PIWI domain
E143
D66
D129
AcPiwil1.3 - FL
DmPiwi - PIWI domain
E157
D143
D64
AcPiwil1.4 - FL
DmPiwi - PIWI domain


A
Lake Malawi
AC
Proportion of mapped reads
Length
Testis
Ovary
Muscle
24-35 nt length small RNAs
Bits
Testes
Ovary
Muscle
Ribonucleotide position in piRNA
5' 3'
3' 5'
10 bps
Z-score
Overlap (ribonucleotides)
rep1_Testis
rep2_Testis
rep3_Testis
rep4_Testis
rep1_Ovary
rep2_Ovary
rep3_Ovary
rep1_Muscle
piRNA-1
piRNA-2
1 bp distance
Number of directly adjacent piRNA pairs in same strand
Muscle
Ovary
Testis
Replicate
rep1
rep2
rep3
rep4
B
Lake Tanganyika
AB
Proportion of mapped reads
Length
Testis
Ovary
24-35 nt length small RNAs
Bits
Testes
Ovary
Ribonucleotide position in piRNA
Z-score
Overlap (ribonucleotides)
rep1_Testes
rep2_Testes
rep3_Testes
rep4_Testes
rep1_Ovary
Number of directly adjacent piRNA pairs in same strand
Ovary
Testis
Replicate
rep1
rep2
rep3
rep4
C
Lake Victoria
PN
Proportion of mapped reads
Length
Testis
Ovary
24-35 nt length small RNAs
Bits
Testes
Ovary
Ribonucleotide position in piRNA
Z-score
Overlap (ribonucleotides)
rep1_Testis
rep2_Testis
rep3_Testis
rep4_Testis
rep1_Ovary
rep2_Ovary
Number of directly adjacent piRNA pairs in same strand
Ovary
Testis
Replicate
rep1
rep2
rep3
rep4
D
Nile Tilapia
ON
Proportion of mapped reads
Length
Testis
Ovary
24-35 nt length small RNAs
Bits
Testes
Ovary
Ribonucleotide position in piRNA
Z-score
Overlap (ribonucleotides)
rep1_Testis
rep2_Testis
rep3_Testis
rep1_Ovary
rep2_Ovary
rep3_Ovary
Number of directly adjacent piRNA pairs in same strand
Ovary
Testis
Replicate
rep1
rep2
rep3


A
Lake Malawi
AC
MZ
TM
Fig. 5G
Fig. 5H
① piRNA clusters AC ovary
② piRNA clusters AC testes
③ piRNA clusters MZ testes
④ piRNA clusters TM testes
⑤ Shared clusters testes (AC/MZ/TM)
⑥ Shared clusters all gonads
⑦ Curated TE distribution
⑧ TE distribution
C
ON
O. niloticus
Fig. 5I
① piRNA clusters ovaries
② piRNA clusters testes
③ Shared clusters testes-ovaries
④ TE distribution
B
AC Ovary
AC Testis
MZ Testis
TM Testis
Number of clusters
0 1000 2000 3000 4000 5000
AC ovary reps
AC testis reps
AC Gonads
MZ testis reps
TM testis reps
AC/MZ/TM testis
AC/MZ/TM all gonads
Number of shared clusters
0 1000 2000 3000
D
ON Ovary
ON Testis
Number of clusters
0 2000 4000 6000
ON ovary reps
ON testis reps
ON Gonads
Number of shared clusters
0 1000 2000 3000
E
Sense/Antisense ratio
All clusters Normalised
AC Ovary
AC Testis
MZ Testis
TM Testis
F
Sense/Antisense ratio
All clusters Normalised
ON Ovary
ON Testis
G
CPM muscle
CPM ovary
CPM testis
CPM testis
CPM testis
80
0
-80
600 kb 620 kb 640 kb 660 kb
610 kb 630 kb 650 kb
Chr 17
Genes
ENSACLG00000012466
slc26a3.1
TEs
H
CPM muscle
CPM ovary
CPM testis
CPM testis
CPM testis
120
0
-120
28.12 mb 28.14 mb
28.13 mb 28.15 mb
Chr 13
Genes
ENSACLG00000000912
TEs
I
CPM ovary
CPM testis
120
0
-120
25.76 mb 25.78 mb
25.77 mb 25.79 mb
LG6
Genes
ENSONIG00000041271
ENSONIG00000026693
TEs


A
Observed
Expected
ꜜCurated annotation
AC‡
Genome Features
TE classes
Testis
Ovary
Proportion of bases
5UTR
3UTR
exon
promoter
TE
intron
intergenic
RC
SINE
LINE
LTR
DNA
ns
***
B
ON
C
Testis
Ovary
Muscle
AC†
†Curated annotation
Bits
Antisense
Sense
Ribonucleotide position in piRNA
D
E
Expression of active vs inactive TE families in A. calliptera gonads
DIRS
hAT-Tip100
PiggyBac
Rex-Babar
TcMar-Tc1
CMC-EnSpm
Maverick
Pao
Ovary Testis
rlog normalized counts
↑Active TE families
↓Active TE families
Mobilization Status
Active
Inactive
F
piRNAs mapping to active vs inactive TE families in A. calliptera gonads
rlog normalized counts (piRNA)
G
CMC-EnSpm-15
CMC-EnSpm-16
Maverick-1
Maverick-3
-1.0
TE start
TE end
1.0Kb
AC Muscle
AC Ovary
AC Testis
MZ Testis
TM Testis
H
CMC−EnSpm
Maverick
2-cell
1k cells
Early gastrula
Late gastrula
Early somitogenesis
Mid somitogenesis
Late somitogenesis
Early pharyngula
Mid pharyngula
Late pharyngula
Hatching
Larvae