MITE infestation of germline accommodated by genome editing in *Blepharisma*

Brandon Kwee Boon Seah[1], Minakshi Singh[1], Christiane Emmerich[1], Aditi Singh[1], Christian Woehle[2], Bruno Huettel[2], Adam Byerly[3], Naomi Stover[4], Mayumi Sugiura[5], Terue Harumoto[5], Estienne Carl Swart[1,*]

[1] Max Planck Institute for Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany

[2] Max Planck Genome Center Cologne, Max Planck Institute for Plant Breeding, Building B, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

[3] Department of Computer Science and Information Systems, Bradley University, Peoria IL, USA

[4] Department of Biology, Bradley University, Peoria IL, USA

[5] Department of Chemistry, Biology, and Environmental Sciences, Faculty of Science, Nara Women's University, Nara 630-8506, Japan

* Correspondence: estienne.swart@tuebingen.mpg.de

Summary

During a sophisticated developmental process, ciliates excise numerous internally eliminated sequences (IESs) from a germline genome copy, producing a functional somatic genome. Most IESs ultimately originate from transposons but homology is obscured by sequence decay. To obtain more representative perspectives on ciliate genome editing, we assembled forty thousand IESs of *Blepharisma stoltei*, from a much earlier-diverging lineage than existing models. Short IESs (< 115 bp) were largely non-repetitive, with a pronounced ~10 bp length periodicity, whereas longer IESs (max 7 kbp) were non-periodic and contained abundant interspersed repeats. Contrary to current models, the *Blepharisma* germline genome encodes few transposases. Instead, its most abundant repeat (8000 copies) was a Miniature Inverted-repeat Transposable Element (MITE), apparently a deletion derivative of a germline-limited Pogo-family transposon. We propose MITEs as an important and eventually self-limiting IES source. Rather than defending germline genomes against mobile elements, we argue that transposase domestication actually facilitates junk DNA accumulation.

Keywords

micronucleus, macronucleus, DNA elimination, mobile element, selfish gene, nuclear dualism, sRNA, Ciliophora, protist

Abbreviations

- IES - internally eliminated sequence
- LTR - long terminal repeat
- MAC - macronucleus
- MIC - micronucleus
- MITE - miniature inverted-repeat transposable element
- MITIES - miniature inverted-repeat transposable internally eliminated sequences
- TDR - terminal direct repeat
- TIR - terminal inverted repeat
- TSD - target site duplication

Introduction

Ciliates are microbial eukaryotes that maintain separate germline and somatic genomes in each cell, housed in two types of nuclei. During the sexual life cycle, germline micronuclei (MICs) develop via a process of small RNA (sRNA)-assisted DNA elimination and DNA amplification into new somatic macronuclei (MACs), which are the site of most gene expression in vegetative cells. Germline-limited genome segments, called internally eliminated sequences (IESs), are excised during development from MIC to MAC. The MAC genome content is hence a subset of the germline MIC. Each of the few taxa studied so far has its own peculiarities. For example, typical *Paramecium* IESs are short, have unique sequence content, and are precisely excised, while *Tetrahymena* IESs are longer, more repetitive, and imprecisely excised (Arnaiz et al., 2012; Feng et al., 2017; Hamilton et al., 2016).

Ciliate IESs are thought to originate from cut-and-paste DNA transposons (Klobutcher and Herrick, 1997) (Figure 1B), because: (i) 5’-TA-3’ motifs at IES boundaries (*Euplotes*, *Paramecium*) resemble the terminal direct repeats of Tc1/Mariner-superfamily transposons (Klobutcher and Herrick, 1995); (ii) transposon-derived “domesticated” excisases are used to remove IESs (Baudry et al., 2009; Cheng et al., 2010; Nowacki et al., 2009); and (iii) intact transposons encoding transposases are mostly germline-limited (Arnaiz et al., 2012; Herrick et al., 1985; Jahn et al., 1993; Le Mouël et al., 2003). Recently, IESs with non-autonomous mobile elements that resemble miniature inverted-repeat transposable elements (MITEs) have been reported in *Paramecium* spp. (Sellis et al., 2021). MITEs are deletion derivatives of Tc1/Mariner transposons, generally short (<500 bp), lacking coding sequences, bounded by terminal repeats, and are common in plants and animals (Feschotte et al., 2002). However, the autonomous counterparts of most *Paramecium* putative MITEs, including the most abundant ones with thousands of copies, have not been identified.

Developmental DNA elimination has been called “genome defense” because the process removes IESs, which not only derive from selfish genetic elements (transposons), but are often intragenic and hence deleterious if not removed (Yao et al., 2003). The “defense” analogy was popularized due to parallels to other eukaryotes where small RNA-mediated DNA heterochromatinization is thought to suppress mobile element proliferation (Coyne et al., 2012; Grewal and Jia, 2007; Vogt and Mochizuki, 2013). Ciliates have been proposed to use development-specific sRNAs to guide DNA elimination; in oligohymenophoreans, they mark sequences for elimination (Mochizuki et al., 2002; Sandoval et al., 2014; Yao et al., 2003), whereas spirotrich sRNAs mark sequences to be retained (Fang et al., 2012; Zahler et al., 2012). Histone modifications are also required for elimination (Liu et al., 2007; Taverna et al., 2002). sRNAs may not always be strictly necessary: in *Paramecium*, knockdown of key sRNA biogenesis enzymes had a smaller effect on shorter IESs, and were only weakly correlated with the more potent effects of knocking down the main IES excisase (Sandoval et al., 2014; Swart et al., 2014).

Other phenomena during genome editing vary markedly between the few model species studied in detail (reviews: (Chalker et al., 2013; Coyne et al., 2012; Rzeszutek et al., 2020)). For example, in all species, germline chromosomes are fragmented into smaller somatic ones to some degree, but spirotrichs produce extremely short somatic “nanochromosomes” with only one or a few genes. “Unscrambling” of nonsequential MAC-destined sequences into the correct order in the somatic genome occurs frequently in some spirotrichs, e.g. *Oxytricha* and *Stylonychia* (Prescott and Greslin, 1992), infrequently in *Tetrahymena* (Hamilton et al., 2016), and has not been reported in other ciliates (e.g. *Paramecium and Euplotes*). Draft-quality germline genomes are available from only two out of eleven class-level taxa (following taxonomy of Lynn, 2010): Oligohymenophorea (Arnaiz et al., 2012; Guérin et al., 2017; Hamilton et al., 2016; Sellis et al., 2021) and Spirotrichea (Chen et al., 2014) (Figure 1C).

Since it is not apparent which genome editing elements are common to all ciliates, we targeted the heterotrich *Blepharisma stoltei* (class Heterotrichea), whose last common ancestor with other ciliates with sequenced germline genomes is the last common ancestor of all ciliates (Gao and Katz, 2014). *Blepharisma* has been a laboratory model for photobiology (Giese, 1973) and mating factors (Kubota et al., 1973; Miyake and Beyer, 1974; Miyake et al., 1991; Sugiura and Harumoto, 2001), so cultivated strains and protocols for inducing conjugation and development are available, and now too an accurate, highly contiguous draft somatic genome (Singh et al., 2021). The somatic genome encodes a likely IES excisase, *Blepharisma* PiggyMac (BPgm), most closely related to the main IES excisases of *Paramecium* (PiggyMac) and *Tetrahymena* (Tpb2). Other somatic PiggyBac paralogs are also present but lack a complete “catalytic triad”, similar to the situation in *Paramecium* (Bischerour et al., 2018). BPgm is upregulated during formation of the new somatic MAC alongside other development-specific genes, including homologs of sRNA biogenesis proteins implicated in genome editing (Singh et al., 2021).

In this study, we assembled a draft germline genome for *Blepharisma stoltei*. Through single molecule long read sequencing and targeted assembly, we assembled IESs including many with long, repetitive elements, which is not feasible with short read shotgun sequencing alone. We found about ten thousand short (≤115 bp), precisely excised IESs with a periodic length distribution like *Paramecium*’s. However most IESs (about thirty thousand) were longer, up to several kbp, and, importantly, also include a Tc1/Mariner transposon whose non-autonomous MITE was also the most abundant repeat in the genome. Complementing the genomic analyses, we also identified small RNAs expressed during sexual development with characteristics of scnRNAs that guide DNA elimination in other ciliates. These results show that characteristics of germline-limited DNA in ciliates may be disjunct to phylogeny, and also illustrate how MITEs could be an intermediate stage in the origin and proliferation of IESs.

Results

Detection and targeted assembly of ca. forty thousand germline-limited IESs

To investigate the *Blepharisma* germline genome we enriched germline micronuclei from *B. stoltei* strain ATCC 30299, and reconstructed 39799 IESs (13.2 Mbp total, average coverage ~45x) scaffolded on the previously assembled 41 Mbp somatic genome (Singh et al., 2021) using a mapping and targeted assembly approach for PacBio long reads (Seah and Swart, 2021). This MAC-scaffolded germline assembly is here referred to as the “MAC+IES” assembly. About 70% of all predicted IESs were intragenic (within coding sequences or introns), implying precise excision of IESs, as they would otherwise cause deleterious translation frameshifts. However, genes occupied 77% of the somatic assembly (excluding telomeres), so there was a small but statistically significant ($p = 3 \times 10^{-269}$) relative depletion of intragenic IESs.

A “hybrid” IES length distribution with periodic length peaks for short IESs

Most IESs were short (median 255 bp, mean 331 bp), but the distribution was long-tailed (90th percentile 603 bp, max 7251 bp). The length distribution was not unimodal, but had multiple peaks at specific length values (Figure 1A, Table S1). It appeared to be a “hybrid” distribution composed of two ranges: a “periodic” range, from ~65 to 115 bp (10778 IESs), and a “non-periodic” range, >115 bp (29021).

The “periodic” IES size range contained sharp peaks every 10 to 11 bp, similar to the periodicity of IESs in *Paramecium tetraurelia* (Arnaiz et al., 2012; Guérin et al., 2017). The first peak in *B. stoltei* was centered at 65 bp, compared to 28 bp in *P. tetraurelia*, and there was no “forbidden” peak. The most abundant “periodic” length peaks were at 72 bp and 110 bp. The “non-periodic” range (≥115 bp) contained isolated peaks at 153, 174, 228, and 389 bp, which has no obvious periodicity. Only 9701 IESs (total 1.36 Mbp) were contained within the size classes represented by the above peaks (both periodic and non-periodic) (Table S1), meaning that most IESs had lengths outside the peak values.

IESs are bounded by heterogeneous direct and inverted terminal repeats

In other ciliates, IES boundaries often have conserved terminal repeat motifs that could reflect excisase cut site preferences or IES origins from specific classes of transposons (Klobutcher and Herrick, 1997). We therefore searched for both direct and inverted terminal repeats in *Blepharisma* IESs.

About three quarters of IESs (30212, 9.43 Mbp) were bounded by terminal direct repeats (TDRs) that contained the subsequence TA (“TA-bound”). Other non-TA TDRs accounted for another 6566 (2.85 Mbp); the remainder were not TDR-bound, though some may be assembly errors (Figure 1A). Like most ciliates, *B. stoltei* genomes were AT-rich (somatic 33.5% GC, IESs 33.3% GC) but the number of TA- and TDR-bound sequences was unlikely to be due to nucleotide composition alone (Figure 2A, 2B). The most common TDRs were simple alternations of T and A (TA, TAT/ATA, TATA), especially in IESs up to 228 bp (Figure 2C), with the exception of TAA/TTA (see below).

Erroneous, low-frequency excision of MAC-destined sequences (MDSs) by the excision machinery (“cryptic” IESs) was also detected in MAC DNA libraries, with a slight peak at 72 bp (Figure S1C). Of 10048 cryptic IESs, 56% were TA-bound; TAA/TTA-bound IESs were also common, which suggests that the observed TDRs, including TAA/TTA, represented intrinsic cut site preferences of the domesticated excisase(s) (Figure S1C to F).

Terminal inverted repeats (TIRs) at IES junctions were heterogeneous among IES size classes (Figure 1D, Figure 2F), and no single TIR motif was generally conserved across all *Blepharisma* IESs, unlike the common 5’-TAYNR-3’ motif of *Paramecium* IESs. Considering only TA-bound IESs, boundaries of “periodic” IESs had a weak consensus 5’-TAT rrn ttt t-3’ (weakly conserved bases in lowercase), whereas IES from “non-periodic” peaks had other signatures, e.g. 5’-TAT Agn nnT TT-3’ for both ~153 and ~174 bp IESs. Despite their heterogeneity, TIRs were more common and longer than expected by chance, even with a strict criterion of no gaps or mismatches (Figure 2D to F). Sequence clustering of long (≥10 bp) TIRs showed distinct TIRs associated with specific IES lengths. Additionally, 376 completely palindromic IESs were identified, of which 153 (40.7%) fell within the same ~228 bp length peak, despite comprising several apparently unrelated palindrome sequences (Figure S2, Supplemental Information).

IESs in the ~389 bp size peak had distinctive TDRs and TIRs, suggesting they are a family of “mobile IESs”, i.e. homologous IESs inserted at nonhomologous genomic sites (Sellis et al., 2021), described further below (see “Pogo/Tigger-family transposon with abundant MITEs”).

Repeat elements are abundant in long, non-periodic IESs

Mobile elements that have recently proliferated should appear as interspersed repeat elements in the genome. As identified by RepeatModeler, a quarter of the assembly (12.7 Mbp, 23.3%) was composed of such interspersed repeats; like in other model ciliates (Chen et al., 2014; Hamilton et al., 2016), they made up a greater proportion of germline-limited

IESs (71.0%) than the somatic genome (8.12%) (Figure 3A). The majority of sequence in IESs ≥115 bp was annotated as repetitive, whereas the converse was true for shorter "periodic" IESs (Figure 3C), paralleling short IESs in *Paramecium* which are mostly unique sequences (Arnaiz et al., 2012).

Most interspersed repeats could not be classified to a known transposable element class by RepeatClassifier (Figure 3B, Table S2). The most abundant classifiable type was "DNA/TcMar-Tc2", all of which actually belonged to a single repeat family rnd-1_family-1, followed by "LINE/RTE-X". The most abundant family, rnd-1_family-0, was unclassified and made up 21.2% (2.69 Mbp) of total repeats. Families rnd-1_family-0 and rnd-1_family-1 were related to each other and are discussed further below ("Pogo/Tigger-family transposon with abundant MITEs").

Three non-periodic IES length peaks (153, 174, 389 bp) could be attributed to specific repeat families, suggesting that they proliferated recently (Table S3, Figure 3C, S3B). This was most pronounced for the ~389 bp peak, where 68.5% of the sequence content belonged to rnd-1_family-0, whereas about a quarter of the ~153 and ~174 bp peaks was composed of repeat families rnd-1_family-87 (palindromic) and rnd-1_family-82 respectively.

Germline-limited repeats include few autonomous transposons but many MITEs

Unlike *Tetrahymena* and *Oxytricha* where transposases are abundant in the germline-limited IESs but rare in the somatic genome (Chen et al., 2014; Hamilton et al., 2016), *Blepharisma* encoded only a few dozen identifiable transposase domains in either the germline-limited or somatic genomes. Cut-and-paste DNA transposase domains of the DDE/D superfamily identified in *Blepharisma* included DDE_1 and DDE_3 (Tc1/Mariner family), DDE_Tnp_1_7 (PiggyBac), DDE_Tnp_IS1595 (Merlin), and MULE (Mutator) (Figure 4E, Table S4). Not all copies of DDE/D transposase domains in *Blepharisma* contained an intact catalytic triad, suggesting that some may be inactive fragments or pseudogenes. Nonetheless, domains with an intact triad were found in both germline-limited and somatic sequences. In general, the expression level of somatic transposase genes was substantially higher than germline-limited ones (Figure S4). This contrasts with the observations in *Oxytricha* of abundant germline-limited transposase gene expression (Chen et al., 2014).

To identify intact transposon units, we examined the seven repeat families in the MAC+IES assembly classified by RepeatClassifier (Figure 3B). Of these, only two were predominantly germline-limited and represented by more than one full-length copy, namely rnd-1_family-1 and rnd-1_family-73 (Table S5). They contained distinct transposases from those found in the MAC genome (Figure 4).

Pogo/Tigger-family transposon with abundant MITEs

Repeat elements of rnd-1_family-1 were bound by a ~30 bp terminal inverted repeat (TIR) 5’-CTC CCC CCC CCC CTC CGT GAG CGA ACA AAA-3’ whose poly-C run length was variable, possibly from assembly errors, and were flanked by a putative target site duplication (TSD) 5’-TAA-3’ (or its reverse complement 5’-TTA-3’) (Figure 4B). All thirty intact (≥95% of consensus length) copies of this family were found within IESs, and had high sequence identity, with median 0.5% divergence from consensus.

The encoded transposase contained two domains characteristic of Pogo transposases from the Tc1/Mariner superfamily: a DDE/D superfamily endonuclease domain (Pfam domain DDE_1) and a helix-turn-helix domain (HTH_Tnp_Tc5) (Gao et al., 2020). The conserved acidic residues (“catalytic triad”) characteristic of DDE/D transposases (Yuan and Wessler, 2011) were also present, with the motif DD35D, i.e. all three residues were Asp, 35 a.a. between the second and third conserved Asp. A phylogeny of the DDE_1 domain placed the transposase in the Pogo/Tigger family, most closely related to the Tc2 subfamily and a sequence from the oyster *Crassostrea*, all of which also had the DD35D motif (Figure 4A). The transposase appeared to be germline-limited, with only ten partial Tblastn hits in the somatic MAC genome (seven on “cruft” contigs) mostly overlapping the HTH_Tnp_Tc5 domain (17 to 84 a.a., E-values 2.3 x 10^{-12} to 1.4 x 10^{-6}) and no matches to the DDE_1 domain. However, the TIR did not match previously characterized TIR signatures for the Tc2, Fot, and Pogo subfamilies. A search of all *B. stoltei* IES sequences against HMMs for known DNA transposon TIRs in the Dfam database found only three matches with E-value < 0.01, none from the above subfamilies.

The same TIR and TSD were also found in another repeat family rnd-1_family-0, which was the most abundant repeat in the genome (Figure 4F), but these were short elements without any predicted coding sequences. rnd-1_family-0 elements often constituted most of the ~389 bp IES size class (Figure 3C): the TSDs bounding the repeats (TAA/TTA) were the TDRs for most of these IESs (Figure 2C), and the C-rich TIR motif corresponded to the C-rich IES junctions (Figure 1D, Figure 2F). Copies of rnd-1_family-0 were also found nested in longer IESs, suggesting recent proliferation (Figure S3C). Degenerated or partial copies were found in shorter IESs (Figure 3C), with copies >5% divergence from consensus having median length 308 bp, vs. 388 bp for copies <5% divergence (Figure 4D).

Therefore, we interpreted rnd-1_family-1 as a new Pogo/Tigger transposon, with a non-autonomous derivative MITE, rnd-1_family-0. We propose the names Bogo for the transposon and BogoMITE for its MITE, as well as the new term “MITIES” (miniature inverted-repeat transposable internally eliminated sequences) to reflect their dual nature as

MITEs and IESs. Given their palindromic nature, sequences underlying rnd-1_family-87 and rnd-1_family-160 repeats may also be MITIES.

Tc1-family transposon with microsatellites

Another IES-limited repeat family, rnd-1_family-73, also contained a DDE/D-type transposase coding sequence . Twenty-two copies were >80% of the consensus length with low sequence divergence (median 0.6% vs. consensus). A putative complete transposon bounded by a TSD 5'-TATA-3' and a 38 bp TIR 5'-GTA CCC CCC CCC TCG TTT GTC GCA TTT TCT AGT TTT TT-3' could be defined after manual curation of repeat boundaries (Figure 4C). Nine of these were mobile IESs, with the TSDs corresponding to the IES junctions. The remaining cases were nested in larger IESs alongside other repeat elements. Ten repeats also contained a microsatellite with ~5 to 42 copies of its 10 bp repeat unit 5'-GGG AAG GAC T-3' (Figure 4C) not found elsewhere in the genome. We propose the name BstTc1 for this putative transposon.

The transposase encoded in full-length copies of BstTc1 contained a conserved DDE/D superfamily domain DDE_3, phylogenetically affiliated to the Tc1 family although the exact placement is unclear, grouping with only moderate support with Tc1 elements from *Crassostrea* and *Hydra* (Figure 4A). Its catalytic triad motif DD34E differed from previously reported motifs for the Tc1 family, DD41D, DD37D or DD36E (Dupeyron et al., 2020), so it may be a novel subfamily.

Non-LTR retrotransposon sequences in both the somatic and germline genomes

Three retrotransposon repeat families in the MAC+IES assembly were classified by RepeatClassifier, i.e. "LINE" or "LINE/RTE-X" (Table S5). Two of these were more closely related with numerous very high identity sequences (>97%) (Figure 5A), suggesting recent radiation of two related retrotransposon elements, while the third was more divergent (Figure 5B; Supplemental Information). Unlike the Bogo and BstTC1-derived elements, more retrotransposon-derived sequences were detected in the *B. stoltei* somatic MAC genome than in assembled IESs (Figure 4E, Table S5). However genes in IESs may be undercounted because of (i) lower completeness of germline vs. somatic assembly; (ii) indels caused by the lower accuracy of the uncorrected long reads used to assemble IESs that prevent prediction; and (iii) shorter total length of IESs than somatic sequence. Consistent with them being true somatic sequences, mappings of error-corrected long reads from a MAC-enrichment library spanned well into flanking regions (Figure 5C; Figure S5A, S5B). In each repeat family, some loci showed sharp dips in coverage suggesting partial excision as IESs while other loci did not (Figure S5B). In MAC-enriched DNA, coverage of such sequences is well above residual IES coverage (Figure S1B).

Twenty-nine genes in the main somatic assembly encoded full or partial copies of reverse transcriptase domain RVT_1 (Singh et al., 2021). The four longest retrotransposon genes also encoded an N-terminal apurinic/apyrimidinic endonuclease (Exo_endo_phos_2) domain upstream of RVT_1. This domain pair is characteristic of some proteins from non-LTR retrotransposons/LINE-like transposable elements, e.g. the BS element from *Drosophila melanogaster* (UniProt Q95SX7) (Han, 2010; Udomkit et al., 1995). In contrast to the development-specific upregulation of retrotransposon genes in *Tetrahymena* (Fillingham et al., 2004) and *Oxytricha* (Chen et al., 2014), expression of *Blepharisma* genes encoding proteins containing RVT_1 or Exo_endo_phos_2 domains was negligible in starved cells and throughout a post-conjugation developmental time series, for both germline-limited and somatic copies (Figure S4) (Singh et al., 2021). The only exception was a somatic APEX1 protein homolog (BSTOLATCC_MAC3189). APEX1 is involved in DNA repair (Fritz, 2000), and Blastp best matches of this *Blepharisma* protein to GenBank's NR database are other similarly annotated proteins.

Six retrotransposon-derived sequences from repeat family rnd-1_family-273 contained a central IES that encoded almost half the amino acids of an Exo_endo_phos_2 endonuclease domain (Figure 5D). Excision of the IES during development thus knocks out the endonuclease domain in the somatic version of the gene. Furthermore, the repeat units as a whole had >99% identity to each other over their ~4.1 kbp length, and were flanked by dissimilar sequences (Figure 5D). The similar lengths of these IESs (173 to 182 bp), their homologous location relative to the coding sequence, and their high sequence identity (>96%) all point to a replication of an ancestral retrotransposon which coincidentally contained a sequence recognized and excised as an IES. In two of these cases, the endonuclease and reverse transcriptase domains can be linked into a single reading frame when the IES is present (Figure 5D). None of *Blepharisma*'s putative domesticated transposases are anywhere near as abundant as the retrotransposon repeats in the somatic genome, let alone show signs of substantial recent replication.

Development-specific 24 nt small RNAs are likely scnRNAs in Blepharisma stoltei

Small RNA (sRNA) libraries were sequenced from a developmental time series, where two complementary mating types of *B. stoltei* (strains ATCC 30299 and HT-IV) were separately gamone-treated and mixed to initiate conjugation. Expression patterns of somatic genes from mRNA-seq and the morphological staging have been reported in our sister report on the MAC genome (Singh et al., 2021). Briefly: after mating types were mixed (0 h), cells paired, produced gametic nuclei by meiosis and exchanged them (2 to 18 h), followed by karyogamy (18 to 22 h) and development of the zygotic nuclei to new macronuclei (22 h onwards). At 38 h, about a third of observed cells were exconjugants.

The most abundant sRNA length classes were 22 and 24 nt, comprising 32% and 30% of the total reads respectively (Figure 6A). This is consistent with model ciliates, where Dicer-generated, mRNA-derived siRNAs employed in gene silencing are typically 21 or 22 nt long, whereas development-specific sRNAs are distinct and consistently ≥2 bp longer (Lepère et al., 2009; Mochizuki et al., 2002).

Developmental dynamics of the 24 nt *Blepharisma* sRNAs resembled scnRNAs of other species. Coverage of 24 nt sRNAs mapping to all feature types initially increased from 2 to 6 h and plateaued until 14 h. Coverage over IESs increased further from 14 h to 22 h, reaching ~25 RPKM by the last time point (38 h), whereas coverage declined over coding sequences (CDSs) and other genomic regions (“NON”) after 14 h. The initial increase across all feature types coincided with meiotic stages iv to viii of (Miyake et al., 1991) (Singh et al., 2021), whereas the divergence between IESs and the rest of the genome corresponded to the onset of karyogamy (Figure 6B). In contrast, 22 nt sRNAs were initially abundant (albeit with high variance) at CDS and NON regions but low (<1 RPKM) at IESs, and declined sharply to <5 RPKM in all features from 6 h onwards (Figure 6B).

Blepharisma 24 nt sRNAs had a strongly conserved 5’-U base preference, like scnRNAs in other ciliates (Lepère et al., 2009; Mochizuki and Kurth, 2013; Zahler et al., 2012). For 24 nt sRNAs mapping to IESs, all time points showed conserved 5’-U except for a slight decrease at 6 h (Figure 6D, S6). 24 nt sRNAs mapping to CDSs only showed 5’-U bias after 6 h. We interpret this to mean that 24 nt sRNAs mapping to IESs were predominantly scnRNAs at all time points, whereas those mapping to CDSs initially comprised siRNAs and other types of small RNAs, before being dominated by scnRNAs from 6 h onwards. In contrast, 22 nt sRNAs mapping to CDSs showed no base biases at any time point, whereas 22 nt reads mapping to IESs had a moderate 5’-U bias only from 6 h onwards. The latter may represent true 22 nt scnRNAs, or fragments of originally 24 nt scnRNAs.

Putative scnRNAs have lower coverage over periodic IESs and BogoMITE IESs

Relative expression levels of putative scnRNAs differed between IES size classes. Based on the IES length distribution and repeat content, we divided IESs into five groups: (1) short “periodic” IESs (≤115 bp), (2) BogoMITEs, because that was the most abundant family, (3) IESs with full-length Bogo transposons, (4) IESs with full-length BstTc1 transposons, and (5) all other IESs ("non-periodic"). BogoMITEs and periodic IESs had lower scnRNA coverage (max ~5 and 10 RPKM respectively) compared with nonperiodic IESs (~30 RPKM). The former were comparable to or even lower than expression levels over non-IES features (Figure 6C). Nonetheless, scnRNA coverage of BogoMITEs and periodic IESs showed an initial increase then plateau, without the subsequent decline seen in non-IES regions. Bogo-containing IESs had similar scnRNA coverage to other non-periodic IESs, but BstTc1-containing IESs had higher coverage (Figure 6C).

Because of the repetitive sequence content in IESs and the short sRNA length, it is possible that the expression levels calculated could be affected by mis-mapping. We reason that such mismapping would not influence the results described above, because "periodic" IESs (group 1) had low repetitive content, whereas the transposon-containing IESs (groups 2, 3, 4) each represented a single repeat family so any mismappings would be contained within the same group and count towards the same RPKM value.

Discussion

Despite belonging to the earliest diverging lineage of ciliates sequenced to date, the germline genome of *Blepharisma stoltei* has similarities to established model species, especially the periodic lengths of short IESs like in *Paramecium*. It also provides fresh observations, notably recent proliferation of non-autonomous MITEs that have autonomous counterparts in the same genome, and retroelements in the somatic genome. Parallels between *Paramecium* and *Blepharisma* suggest that ciliate germline characteristics may be relatively plastic over evolutionary time and not strongly phylogenetically constrained.

Comparison to IESs in other ciliates

Most *Blepharisma* IESs are short, TA-bound, and intragenic, more similar to *Paramecium* than *Tetrahymena* or spirotrichs. The most striking parallel is the sharply periodic length distribution of short IESs with peaks every ~10 bp, coinciding with the DNA helical turn, implying that the *Blepharisma* excisase complex has similar geometric constraints as proposed for *Paramecium* (Arnaiz et al., 2012). *Blepharisma* “periodic” IESs are longer on average and do not have a “forbidden” second peak, but the last peak (~110 bp; Figure 1A) is still below the upper limit where such periodicity would be expected given the properties of DNA (Figure 7 of (Arnaiz et al., 2012)). In contrast, *Tetrahymena thermophila* has a continuous distribution (average length ~3 kbp) (Hamilton et al., 2016; Seah and Swart, 2021), while *Oxytricha trifallax* non-scrambled IESs (length ~20-100 bp) have weak periodicity (Chen et al., 2014). Periodicity is consistent with a single primary IES excisase, rather than multiple excisase families, which would smooth the length distribution.

Longer, nonperiodic IESs of *Blepharisma* contain more repeats, including whole transposons, than short IESs. Unlike *Tetrahymena*, where 41.7% of high-confidence IESs comprise putative autonomous transposons (Hamilton et al., 2016), some of which can be grouped into families (Fillingham et al., 2004; Wuitschick et al., 2002), only a small fraction of *Blepharisma*’s long IESs encode transposases, and their length distribution is not unimodal, but long-tailed, with distinct peaks representing individual abundant families (Figure 3). Germline-specific repeats and transposons across *Paramecium* spp. have recently been surveyed (Sellis et al., 2021), but were likely underestimated because such repeats are difficult to assemble from short-read data even with high coverage, as we saw with *Blepharisma* BogoMITE elements, (Supplemental Information, Figure S1A).

The dynamics of *Blepharisma* 24 nt sRNAs are consistent with the scnRNA turnover model, where RNA intermediates are produced from both IESs and MDSs (Malone et al., 2005; Mochizuki and Gorovsky, 2005), but those from MDSs are selectively degraded, allowing the remaining scnRNAs to mark IESs for excision. *Blepharisma* 24 nt sRNAs mapping to IESs increase more than those mapping to CDSs during post-conjugation development (Figure 6B), complementing our finding that homologs of scnRNA biogenesis proteins, Dicer-like (Dcl) and Piwi proteins, are highly upregulated during development (Singh et al., 2021). Furthermore, higher coverage of *Blepharisma* scnRNAs in longer (presumably younger) IESs than in short (~older) periodic IESs mirrors that of *Paramecium*, where younger IESs are more likely to require scnRNAs for efficient excision (Lhuillier-Akakpo et al., 2014; Sellis et al., 2021).

The longer an IES, the more likely it will contain a promoter by chance or contain one from a transposase gene, thus giving rise to such sRNAs. This would explain the low 24 nt sRNA level of BogoMITE IESs in contrast to their autonomous counterparts (Figure 6C), though removal of both is essential. In contrast to the abundant Bogo transposon 24 nt sRNAs, expression of these and other transposase genes in RNA-seq is negligible (Figure S4). This raises the possibility that active, transcribed *Blepharisma* transposons are in fact silenced, turning most of their transcripts into 24 nt sRNAs. This is contrary to the role of scnRNAs proposed to target DNA for excision, but congruent with the role of sRNAs in transposon silencing in other eukaryotes, from which the scnRNA biosynthesis enzymes originated (Sandoval et al., 2014).

Are MITEs a missing link in the IBAF model?

The prevailing Invasion-Bloom-Abdication-Fade (IBAF) model for the evolution of IESs hypothesizes that they originate from cut-and-paste DNA transposons that invade and proliferate (“bloom”) in the germline genome (Klobutcher and Herrick, 1997). Transposon proliferation stops (“abdication”) when its transposase is domesticated by a host promoter, releasing the transposons from purifying selection, whereupon their sequences erode by drift (“fade”). Depictions of the IBAF model usually show all the transposons expressing transposases during “bloom”, i.e. functioning as autonomous transposons (Feng and Landweber, 2021; Klobutcher and Herrick, 1997). This is reasonable for *Tetrahymena* and *Oxytricha*, which have hundreds of germline-encoded transposases that vastly outnumber those in the somatic genome (Table S4). However, *Blepharisma* and *Paramecium* only have a few dozen transposases, although germline-limited transposases may be underestimated, especially for short-read assemblies.

This discrepancy can be resolved by taking MITIESs (MITE IESs) into account. In *Blepharisma* this is best exemplified by the few autonomous Bogo transposon copies compared to thousands of non-autonomous BogoMITEs. The narrow length distribution of

BogoMITEs, their high sequence identity, and occasional nested insertion inside unrelated IESs are the clearest illustrations to date of recent MITE proliferation. Bogo is also the first Pogo/Tigger transposon found in a ciliate germline genome; this subfamily is known to be especially prone to MITE formation (Feschotte and Mouchès, 2000; Guermonprez et al., 2008). The prevalence of IESs bound by terminal inverted repeats, including numerous palindromic IESs (Figure 2D, S2), also suggest many more *Blepharisma* IESs are MITE derivatives.

In *Paramecium* spp., MITEs of the Thon and Merou transposons have been identified but only numbered about a dozen copies per genome, and their transposases belong to a different transposase family than Bogo (Figure 4). The most abundant mobile IES family in *Paramecium*, FAM_2183, is probably a MITE but its autonomous counterpart was not reported (Sellis et al., 2021). MITEs as transposon/IES life cycle intermediates can hence explain why *Blepharisma* and *Paramecium* have few MIC-encoded transposases compared to *Oxytricha* and *Tetrahymena*, but nevertheless tens of thousands of IESs.

MITEs also provide a mechanism for transposon/IES proliferation self limitation (Figure 7A). When MITEs outnumber the autonomous transposon, active transposase protein is more likely to bind to target sites in MITEs than the full length transposon (“titration”), hindering the replication of the autonomous version, giving time for loss-of-function mutations to inactivate the transposases (“fade”). This “vertical inactivation” scenario (Hartl et al., 1997) was already discussed in the original IBAF proposal (Klobutcher and Herrick, 1997), but no plausible examples from ciliates were then known.

Is “genome defense” a flawed analogy?

The IBAF model also does not explain how ciliates can consistently and precisely excise novel mobile elements from different transposon families that invade the germline genome. The domesticated excisases of *Paramecium* (Baudry et al., 2009), *Tetrahymena* (Cheng et al., 2010), and *Blepharisma* (Singh et al., 2021) belong to the PiggyBac family. Except for *Tetrahymena* Tpb2, PiggyBacs are known to perform seamless excision, where the host sequence after transposon excision is identical to that before insertion (Chen et al., 2020). This would make them the ideal progenitor for IESs within coding sequences; indeed, PiggyBac transposons are also known to produce MITEs (Mitra et al., 2013; Wang et al., 2010). By extension, the first IESs probably originated from PiggyBac transposons. But what about subsequent invasions by other transposons that leave behind “scars” upon excision? Such imprecision would cause deleterious frameshift mutations in coding regions. How can they invade the germline genome and yet avoid deleterious effects?

Part of the answer lies in the “hijacking” model proposed from *Paramecium* (Arnaiz et al., 2012; Sellis et al., 2021), whereby the domestication of PiggyBac transposase changed the dynamic for subsequent transposon invasions. New transposons would persist as IESs only if they also encode a seamless excisase, or if they can also be recognized and cut by the exapted PiggyBac transposase. The latter favors the invasion of transposons that produce a TSD containing a submotif recognized as a cut site by PiggyBac (Figure 7B). The similarity between IES and transposon boundaries would hence not be due to common origin or sequence evolution after IES fixation in the germline (Klobutcher and Herrick, 1997), but rather because of selection for transposons whose TSDs already match the excision site preferences of domesticated PiggyBac. Analogous exaptation of TSDs for excision has been demonstrated in another context: independent origin of introns from MITEs in at least two different eukaryotes, where one of the TSDs produced upon MITE insertion was co-opted as an intron splice site (Huff et al., 2016). Cross-talk between different (albeit related) transposases for MITE transposition has also been documented (Feschotte et al., 2005).

We further argue that “genome defense” is a teleological expression that confuses cause and effect. Domesticated excisases actually facilitate mobile element accumulation in the germline, by shielding them from selection by effective exclusion from the somatic genome. *Tetrahymena* is the exception that proves the rule: its domesticated excisase appears to be imprecise; correspondingly, most of its IESs are intergenic, because intragenic IESs have been efficiently removed by selection (Cheng et al., 2016; Feng et al., 2017). The origins of gene silencing by DNA methylation in vertebrates have also been reinterpreted with similar reasoning. Vertebrates have high levels of CpG methylation that inactivates transposons, which was thus proposed to “compensate for” transposon proliferation in eukaryotic genomes (Bestor, 1990). When seen from a non-teleological perspective, it is precisely because CpG-mediated transposon inactivation is so effective, preventing exposure to selection, that transposons persist, leading to larger genomes (Zhou et al., 2020).

Why does the Blepharisma somatic genome contain retrotransposon sequences?

Transposon-related sequences are typically germline-limited in other model ciliates, which was formerly interpreted as successful “genome defense” keeping them out of the somatic MAC genome (Chen et al., 2014; Fillingham et al., 2004; Guérin et al., 2017; Hamilton et al., 2016; Swart et al., 2013). Counter to this, we found several retrotransposon-derived sequences in the *Blepharisma* MAC genome (Figure 5; Table S2). Some show signs of partial excision or possible absence of the locus in part of the population, but plenty have uniform coverage typical of somatic sequences.

Recent retrotransposon proliferation in the soma and patchy distribution of different somatic transposase classes across ciliates (Table S4) (Singh et al., 2021) suggest that "genome defense" is at best leaky. We conjecture that if foreign DNA lacks suitable target sites recognized by the excisase, it might still be marked by scnRNAs but fail to be excised or only be partially excised (e.g. the IESs in Figure 5C). Such DNA would still be deleterious if inserted intragenically.

Somatic MACs may be unable to repress mobile elements by heterochromatinization like germline MICs and other eukaryotic nuclei. In *Tetrahymena*, most MAC DNA is not associated with classical heterochromatin marks (Liu et al., 2007), while in *Paramecium* MACs, H3K27me3 is not associated with transcription repression, despite being a classic heterochromatin mark in multicellular eukaryotes (Drews et al., 2021). In such a permissive expression environment, selection against mobile elements that are not already excised as IESs may be especially effective, unless they are relatively transcriptionally inactive like the *Blepharisma* retroelements. On the other hand, regular *Blepharisma* stock culture passaging maintains a small effective population size, which would counteract selection against mobile element accumulation in the soma.

The genome defense model may lead one to dismiss IES retention in the somatic genome as excisase inefficiency or MIC contamination of the library, however, IES excision is not all-or-nothing but a continuum. Experimental evolution experiments in *Paramecium* suggest IES retention variability is itself a plastic and evolvable trait with consequences for somatic genotypic diversity (Catania et al., 2021; Vitali et al., 2019). Assembly algorithms tend to present an oversimplified, "pristine" view of somatic genomes, because they collapse repetitive and lower-coverage regions, which are characteristic of mobile elements and partially retained IESs. Accurate long read sequencing, haplotype-aware assemblers, and sequence graphs will all play a role in building a more realistic picture of somatic genome heterogeneity.

Conclusion

Why do we credit developmental DNA elimination with defending the genome, when natural selection has been doing the hard work? Apart from technical biases during genome assembly, there is also sampling bias by using lab strains. These are often clonal and largely homozygous; if so, we would not observe accumulation of strongly deleterious foreign DNA that actually needs defending against, but only IESs that have reached fixation and that are already efficiently excised and non-deleterious. Purifying selection against deleterious IESs has had to be indirectly observed, e.g. in the lack of intragenic IESs in *Tetrahymena*, where excision is imprecise (Hamilton et al., 2016), and the statistical depletion of IES-like sequences in the *Paramecium* somatic genome (Swart et al., 2014). Similar evolutionary logic applies to prokaryotic CRISPR defense systems, where hidden fitness costs (autoimmunity) have been underestimated because those individuals are removed by selection (Stern et al., 2010), hence the phenomenon is easily misinterpreted as inheritance of acquired traits (Weiss, 2015). Most studies on ciliate developmental DNA elimination to date have focussed on the underlying molecular mechanisms, but to understand its origins and evolution we should expand our view to diverse ciliates and their germline genomes from natural populations.

Figure Legends

Figure 1. A “hybrid” IES length distribution with periodic length peaks for short IESs. (**A**) IES length histogram (0 to 500 bp (inset: full range), stacked bars for types of terminal direct repeats (TDRs) at IES boundaries. Peaks for IES size classes discussed are marked. (**B**) Comparison of cut-and-paste DNA transposons (above) and ciliate genome editing (below), showing parallels between target site duplications (TSD) of transposons and terminal direct repeats (TDRs) bounding IESs, and effects of precise vs. imprecise excision. (**C**) Diagrammatic tree of ciliates (following Lynn, 2008), branch lengths arbitrary. Genera with draft MIC genomes listed on right. (**D**) Sequence logos for MDS-IES junctions for TA-bound IESs of specific size classes, centered on the “TA”. See also Figure S1.

Figure 2. IESs are bounded by heterogeneous direct and inverted terminal repeats. (**A**) Numbers of terminal direct repeats (TDRs) per TDR length observed (blue) vs. number expected by random chance if bases were independently distributed (orange). (**B**) Ratio of observed to expected numbers of TDRs by length. (**C**) Length distributions of IESs containing TDRs of lengths 2, 3, 4, and 5 bp; the most abundant TDR sequences per TDR length are shown in color (sequences and their reverse complements are counted together, because TDRs could be encountered in either orientation, e.g. TAA/TTA), simple T/A alternations are in shades of blue. NB: plots in panel C have different vertical axis scales. (**D**) Observed IESs per terminal inverted repeat (TIR) length vs. expected number by chance alone. (**E**) Same as panel D but for *P. tetraurelia*. (**F**) Lengths (scatter-overlaid boxplot) of IESs containing long TIRs (≥10 bp), grouped by their TIR sequence (rows). Each TIR-cluster is annotated with the median IES length (bp), cluster size (n), TDR consensus sequence, and TIR representative sequence. See also Figure S2.

Figure 3. Repeat elements are abundant in long, non-periodic IESs. (**A**) Total sequence length annotated as interspersed repeats vs. non-repetitive , in germline-limited vs. somatic parts of the genome. (**B**) Classification of repeat families by RepeatClassifier, and total annotated length per repeat class. (**C**) Total sequence length (vertical axis) per IES size class (horizontal axis), stacked plot of non-repetitive fraction vs. interspersed repeats , with the most abundant repeat families in the four non-periodic peaks overlaid in color. Inset: Distribution to 1000 bp. See also Figure S3.

Figure 4. Germline-limited repeats include few autonomous transposons but many MITEs. (**A**) Phylogenetic tree of DDE/D domains for Tc1/Mariner superfamily, including *B. stoltei* germline-limited (Bogo and BstTc1) and somatic transposases. (**B**) Diagram of features in Bogo and BogoMITE; TSD – target site duplications, TIR – terminal inverted

repeats, HTH_Tnp_Tc5, DDE_1 – conserved domains. (**C**) Diagram of features in BstTc1: DDE_3 – conserved domain. (**D**) Histograms of sequence divergence from repeat family consensus for copies of the Bogo and BogoMITE repeat families annotated by RepeatMasker; for rnd-1_family-1, most low-divergence copies (<5% divergence) were short fragments, but all full-length copies were low-divergence. (**E**) Counts of transposase-related domains in different ciliates from six-frame translations of somatic vs. germline-limited genome sequence. See also Figure S4. (**F**) Sequence logos for Bogo and BogoMITE repeat boundaries, aligned on the terminal inverted repeats (TIRs) and terminal direct repeats (TDRs). 3'-boundaries have been reverse complemented to show the TIRs. Sequence logos were generated from alignments of full-length, intact Bogo elements (>1.8 kbp) and BogoMITEs (between 385-395 bp), with columns comprising >90% gaps removed.

Figure 5. Non-LTR retrotransposon sequences in both somatic and germline genomes. (**A**) Phylogeny of rnd-1_family-273 and rnd-1_family-276 retrotransposon sequences. (**B**) Phylogeny of rnd-4_family-193 retrotransposon sequences. (**C**) Window of mapped HiFi reads from sucrose gradient-purified MACs (grey) spanning a retrotransposon gene with both an AP endonuclease domain and a reverse transcriptase domain (from rnd-4_family-193). Only sequence columns with < 90% gaps are shown. (**D**) Multiple sequence alignment of non-LTR retrotransposon copies from rnd-1_family-273. Schematic for consequences of IES excision (Contig_45). Identity scale: green=100%; gold=30-99.9%; red=0-29.9%. See also Figure S5.

Figure 6. Development-specific 24 nt small RNAs are likely scnRNAs in *B. stoltei*. (**A**) Read length histogram for all sRNAs in the time series. (**B**) Relative expression (RPKM units, vertical axis) of 22 and 24 nt sRNAs mapping to different feature types across time series: blue - IES, orange - CDS, green - all other regions not annotated as IES or CDS (including UTRs and intergenic regions which are difficult to delimit exactly with available data). Timing of developmental stages inferred from morphology are labeled below (Singh et al., 2021). (**C**) Relative expression of 22 and 24 nt sRNAs mapping to different categories of IESs: containing full-length copies of BstTc1 and Bogo transposons, at least 90% covered by BogoMITE elements, IESs in the periodic length range (< 115 bp), and all other IESs ("non-periodic"). (**D**) Sequence logos for 22 and 24 sRNAs mapping to CDS and IES features in controls and different time points (rows). See also Figure S6.

Figure 7. Model for IES evolution in a ciliate genome with an existing domesticated excisase. (**A**) Graphs depict IES length distribution. (**1**) Invasion of germline genome by full length transposon (green); existing IESs (blue) are excised by domesticated excisase. (**2**) New transposon produces MITIES which are both MITES and IESs. (**3a**) If MITIES can be

excised by domesticated excisase, they proliferate and titrate the progenitor transposase. (**4**) Proliferation of MITIES favors vertical inactivation of the full length transposon; loss of function stops production of new MITIES, leading to eventual decay. (**3b**) If the MITE cannot be excised by domesticated excisase (i.e. it is not an IES), it is more likely to cause deleterious mutations upon insertion, and is therefore selected against and does not reach fixation. (**B**) If a transposon TSD contains a submotif that can be recognized by the domesticated excisase, it can theoretically be excised cleanly without leaving a “footprint”, avoiding potential frameshift mutations.

Supplemental Figure Legends

Figure S1. Length distributions and retention scores for different IES assembly methods, MAC library, and cryptic IESs. (**A**) Comparison of IES reconstructions from MIC-enrichment library sequenced with short reads by ParTIES (above) vs. from long reads by BleTIES (below). Main panels: IES length histograms up to 500 bp, insets: IES retention scores colored by TDR sequence type. Length peak at ~390 bp representing BogoMITE element is present in BleTIES reconstruction but not ParTIES. (**B**) Conventional IESs: retention scores computed from MAC-enriched library, sequenced with PacBio HiFi reads. (**C**) “Cryptic” IESs from MAC read library: length histogram, colored by TDR sequence type. (**D**) Retention scores of “cryptic IESs”. (**E**) Length distribution of “cryptic” IESs that contain “TTA” or “TAA” in their TDR, detail <500 bp, inset detail <150 bp. (**F**) Sequence logos of TA-bound “cryptic” IES junctions centered on the TA motif, for all cryptic IESs (above) and the subset in the ~72 bp size class (below). (**G**) Mapping pileup at IES with TA-containing TDR. For aligned reads in panels E and F, dots: bases identical to reference, dashes: gaps relative to reference, red bar: read clipping. (**H**) Mapping pileup at IES with non-TA-containing TDR.

Figure S2. Palindromic IESs clustering and length distribution. Strip plots of IES lengths for palindromic IESs (≥90% self-alignment identity), after they have been clustered by sequence identity (rows represent clusters). Each cluster is annotated with the median IES length and the cluster size. Insets: (**A**) Overall sequence length distribution histogram for all palindromic IESs. The most common length of palindromic IESs is ~230 bp. (**B, C**) Dendrogram of sequence distance and multiple sequence alignment of palindromic IESs with ~230 bp length to illustrate that they comprise several distinct clusters of sequences.

Figure S3. Most abundant repeat families in non-periodic IES size classes. (**A**) Total lengths (horizontal axis) of the top ten repeat families per IES size class (panel rows). (**B)** Top repeat family (by sequence length) for each IES size class (panel rows); the total length covered by that repeat family within IESs vs. the lengths of those IESs is shown in red, superimposed on the total sequence vs. IES length distribution of IESs in general (grey). Arrowheads mark centers of the size classes. (**C**) Examples of nested repeats within IESs. Nested elements can be recognized when the two outer repeat elements belong to the same family and align to consecutive parts of its family’s consensus sequence, implying that the inner element has likely been inserted into the middle of an existing element. Coordinates of the split segments are relative to the repeat family consensus.

Figure S4. Expression of genes with transposase domains. Comparison of expression levels for MAC- vs. MIC-limited transposase-related domains across developmental time series; heatmap color scaled to log(transcripts per million). Domain architecture shown diagrammatically.

Figure S5. Non-LTR retrotransposon sequences in both somatic and germline genomes. (**A**) As in Figure 5A. (**B**) As in Figure 5A. Inset shows coverage across the entire contig and position of the retrotransposon gene. **(C**) Alignment of MAC+IES and somatic genomic sequences for Contig_44 retroelement genes from Figure 5A, showing how excision of the central IES deletes part of the endonuclease domain and produces a premature stop codon.

Figure S6. Per-position base entropy of 22 nt and 24 nt sRNAs from developmental time series. Plots show conservation of 5’-U in 24 nt sRNAs. Each plot symbol represents positional sequence entropy (symbol size) for a given nucleotide base (columns) and position in the sRNA sequence (vertical axis) and time point (horizontal axis), in sRNAs mapping to different feature types (rows).

Methods

General reagents were analytical grade and purchased from Sigma-Aldrich or Merck unless otherwise indicated.

Ciliate strains origin and cultivation

The strains used and their original isolation localities were: *Blepharisma stoltei* ATCC 30299, Lake Federsee, Germany (Repak, 1968); *Blepharisma stoltei* HT-IV, Aichi prefecture, Japan (Harumoto et al., 1998). Methods for cell cultivation and harvesting of material for sequencing are described in our sister report (Singh et al., 2021).

Enrichment of micronuclei, isolation and sequencing of genomic DNA

B. stoltei ATCC 30299 cells were harvested and cleaned to yield 400 mL of cell suspension (1600 cells/mL). This suspension was twice concentrated by centrifugation (100 g; 2 min; room temperature) in pear-shaped flasks and in 50 mL tubes to ~8 mL. 10 mL chilled Qiagen Buffer C1 (from the Qiagen Genomic DNA Buffer Set, Qiagen no. 19060) and 30 mL chilled, autoclaved deionized water were added. The suspension was mixed by gently inverting the tube until no clumps of cells were visible, and then centrifuged (1300 g; 15 min; 4˚C). The pellet was washed with chilled 2 mL Buffer C1 and 6 mL water, mixed by pipetting gently with a wide-bore pipette tip, centrifuged (1300 g; 15 min; 4˚C), and resuspended with chilled 2 mL Buffer C1 and 6 mL water by pipetting gently with a wide-bore pipette tip.

The nuclei suspension was layered over a discrete sucrose gradient of 20 mL 10% (w/v) sucrose in TSC medium (0.1% (v/v) Triton X-100, 0.01% (w/v) spermidine trihydrochloride and 5mM $CaCl_2$) on top of 40% (w/v) sucrose in TSC medium (Lauth et al., 1976). Gradients were centrifuged (250 g; 10 min; 4˚C). 10 to 12 mL fractions were collected by careful pipetting from above, and the nuclei were pelleted by centrifugation (3000 g; 10 min; 4˚C). DNA was extracted from pelleted nuclei with the Qiagen Genomic tips 20/G and HMW DNA extraction buffer set (Qiagen no. 19060) according to the manufacturer's instructions. DNA concentration was measured by the Qubit dsDNA High-Sensitivity assay kit. Fragment size distribution in each sample was assessed by a Femto Pulse analyzer.

B. stoltei ATCC 30299 DNA isolated from the MIC-enriched fraction on two separate occasions was used to prepare two sets of DNA sequencing libraries. A low-input PacBio SMRTbell library was prepared without shearing the DNA and was sequenced in the CLR-(continuous long read) sequencing mode on a PacBio Sequel II instrument. Paired-end short-read libraries were prepared for four sucrose gradient fractions (top (T), middle (M), middle lower (ML), bottom (B)) and sequenced with 100 bp BGI-Seq paired-end reads on a BGI-Seq instrument.

IES prediction from PacBio subreads

PacBio subreads (CLR reads) from a MIC-enriched sample (ENA accession ERR6548140) were aligned to the somatic genome reference assembly (accession PRJEB40285) (Singh et al., 2021) with minimap2 v2.17-r941 (Li, 2018), with options: -ax map-pb --secondary=no --MD. Mapped reads were sorted and indexed with samtools v1.10 (Li et al., 2009), and then used for predicting IESs with BleTIES MILRAA v0.1.9, with options: --type subreads --junction_flank 5 --min_ies_length 15 --min_break_coverage 10 --subreads_pos_max_cluster_dist 5. The BleTIES pipeline has been previously described (Seah and Swart, 2021) and uses spoa v4.0.3 (Vaser et al., 2017) for assembly. After inspecting the initial IES predictions, we removed IES predictions with length <50 bp and retention score <0.075, which we judged to be more likely to be spurious or to have insufficient coverage for an accurate assembly.

Terminal direct repeats (TDRs) at the boundary of a given IES were defined as a sequence of any length that was exactly repeated on both ends of the IES, such that one copy lies within the IES, and the other in the MAC-destined sequence. Because the sequence is identical, it is not possible to determine from sequencing data alone where the physical excision of the IES would occur; such ambiguous excision junctions have been termed “floating IESs” (Sellis et al., 2021). Therefore, TDRs were always reported starting from the left-most coordinate. If the TDR sequence contained 5’-TA-3’, the corresponding IES was also considered to be “TA-bound”, even if the TDR was longer than the 2 bp 5’-TA-3’ sequence.

Reconstructed IES sequences were computationally inserted into the MAC assembly with BleTIES Insert, to produce a hybrid MAC+IES assembly, which approximates the part of the MIC genome that is collinear with the MAC.

Identification and comparison of IES length classes

Visual inspection of the length distribution of BleTIES-predicted IESs showed sharp peaks every ~10 bp between ~65 and 115 bp. Peak calling on the graph of number of IESs (TA-bound only) vs. length (bp) was performed with the function find_peaks from the Python package scipy.signal v1.3.1 (Virtanen et al., 2020), with height cutoff 100. The ranges for each IES size class were defined with the width at half peak height. In *Paramecium tetraurelia*, where most IESs are TA-bound, the IES termini have a short, weakly conserved inverted repeat (Arnaiz et al., 2012; Klobutcher and Herrick, 1995). To search for similar motifs in *B. stoltei*, sequences flanking TA-bound IES junctions were extracted, with one from each pair reverse-complemented so that the sequences were always in the orientation 5'-(MDS segment)-TA-(IES segment)-3'. Sequence logos of the junctions (10 bp MDS, 14 bp within IES, not including the TA itself) were drawn for each IES length class with Weblogo (Crooks et al., 2004). Only TA-bound IESs were used for the sequence logos because they could be aligned relative to the 5'-TA-3' repeat, whereas for IESs bound by other types of junctions there is no common reference point to align the boundaries of the IES.

Probability of a pair of repeated sequences

Under a null model where all bases in a sequence are independently and identically distributed, the probability P_n of having any possible sequence of length *n* bounding a given sequence feature (either a TDR or a TIR) is the sum of probabilities of all possible sequences (each of which notated as *k*) of length *n*, squared: $P_n = \sum_{k \in K} {p_k}^2$, which can be transformed to $P_n = (\sum_{b \in B} {p_b}^2)^n$, where *B* is the alphabet of bases and p_b is the individual probability of each base. The number of possible sequences *k* of length *n* is simply $|K| = |B|^n$.

The probability of having a repeat of length at least 2 is equal to the probability of having a repeat of length 2, because all cases of repeat length > 2 implicitly have a repeat of length = 2. Therefore the probability of having a repeat of length exactly *n*, i.e. match in bases 1 to *n*, and mismatch on base *n*+1 is $P_n \times Pr\ (mismatch) = P_n \times (1 - \sum_{b \in B} {p_b}^2)$. The expected number of TDRs in *Blepharisma* were calculated by using the empirical base frequencies of the MAC+IES genome assembly for p_b, and multiplying this probability by the number of IESs.

Identification of terminal inverted repeats (TIRs) and palindromes in IESs

The BleTIES-assembled IES sequences for *Blepharisma* were used to identify exact, ungapped terminal inverted repeats (TIRs). Starting from the ends of the IES sequence immediately within the flanking TDRs, each base was compared to the reverse complement of the corresponding base on the opposite end for a match, extending the TIR until a mismatch was encountered, up to a maximum length of 25 bp. The same procedure was used for *Paramecium tetraurelia* using IESs sequences downloaded from ParameciumDB (https://paramecium.i2bc.paris-saclay.fr/files/Paramecium/tetraurelia/51/annotations/ptetraurelia_mac_51_with_ies, accessed 14 October 2021), except that the coordinates of TDRs were first renumbered and extended beyond the "TA" motif if possible, following the BleTIES coordinate numbering convention, in case there are potential TDRs that are longer than a simple TA. The expected number of TIRs of given lengths under a null model was computed as described in “Probability of a pair of sequences”.

Long TIRs (≥10 bp) were clustered by sequence identity to look for IESs of potentially related origin, using the cluster_fast algorithm (Edgar, 2010) implemented in Vsearch v2.13.6 (Rognes et al., 2016) at 80% identity and the CD-HIT definition of sequence identity (-iddef 0). For each resulting cluster of similar TIRs, the cluster centroid was used as the representative sequence shown in Figure TIRS. TDRs associated with each cluster’s IESs were grouped by length, and for each TDR length a degenerate consensus was reported with the degenerate_consensus function of the Bio.motifs module in Biopython v1.74.

Palindromic IESs were defined as IESs that align to their own reverse complement with a sequence identity ≥90% (matching columns over sequence length); this definition was less strict and permitted inexact matches unlike the TIR search, to allow for sequence divergence and assembly errors. IES sequences were aligned with the PairwiseAligner function from Bio.Align in BioPython v1.74, using global mode and parameter match_score = 1.0, with all other scores set to zero.

Palindromic IESs were clustered with Vsearch cluster_fast as described above, except that one sequence (BSTOLATCC_IES35757) was manually removed after inspection of results because it appears to contain two different nested palindromic sequences. Cluster centroids were aligned pairwise as above and used to calculate a matrix of edit distances (matching columns / alignment length). The distance matrix was clustered with average linkage clustering to produce a sequence distance dendrogram with the functions average and dendrogram from scipy.cluster.hierarchy v1.3.1 (Virtanen et al., 2020).

Comparison of intragenic:intergenic IES ratios

Intragenic vs. intergenic IESs were defined by overlap of predicted IES annotations with “gene” feature annotations on the MAC reference (ENA accession GCA_905310155), using Bedtools v2.30.0 (Quinlan and Hall, 2010) and pybedtools v0.8.1 (Dale et al., 2011).

To test whether the underrepresentation of IESs within gene features was statistically significant, compared to the null hypothesis of IESs and gene feature locations being independently distributed, we assumed that the number of intragenic IESs would follow a binomial distribution with individual probability equal to the fraction of the genome that is covered by gene features. The p-value of the observed number of intragenic IESs would then be equal to the cumulative probability density up to and including the observed value.

Developmental time series small RNA-seq

Complementary mating strains *B. stoltei* ATCC 30299 and HT-IV were pre-treated with Gamone 2 and Gamone 1 respectively, and then mixed to initiate conjugation as described previously; sRNA and mRNA were isolated from total RNA at the same time points (“Conjugation time course”, (Singh et al., 2021)). sRNA libraries were prepared with the BGISeq-500 Small RNA Library protocol, which selects 18 to 30 nt sRNAs by polyacrylamide gel electrophoresis, and sequenced on a BGISeq 500 instrument.

Small RNA libraries mapping and comparison

Small RNA libraries were mapped to the MAC+IES assembly with bowtie2 v2.4.2 (Langmead and Salzberg, 2012) using default parameters. Total reads mapping to CDS vs. IES features were counted with featureCounts v2.0.1 (Liao et al., 2014). To account for different total sequence lengths represented by CDSs, IESs, and intergenic regions, the read counts were converted to relative expression values (reads per kbp transcript per million reads mapped, RPKM (Mortazavi et al., 2008)) using the total lengths of each feature type in place of transcript length in the original definition of RPKM, with the following formula:

10^9 × (reads mapped to feature type) / (total reads mapped × total length of feature type).

Reads mapping to CDSs, IESs, or neither (but excluding tRNA and rRNA features) were extracted with samtools view, with 22 and 24 nt reads extracted to separate files. Read length distributions for each sequence length and feature type were summarized with samtools stats.

mRNA-seq read mapping

To permit correct mapping of tiny introns RNA-seq data was mapped to the MAC genome using a version of Hisat2 (Kim et al., 2019) with the static variable minIntronLen in hisat2.cpp in the source code lowered to 9 from 20 (https://github.com/Swart-lab/hisat2/; commit hash 86527b9). Hisat2 was run with default parameters and parameters --min-intronlen 9 --max-intronlen 30. It should be noted that spliced-reads do not span introns that are interrupted by an IES due to the low maximum length, however such cases are not expected to occur often.

Gene prediction and domain annotation

To predict protein-coding genes in IESs, non-IES nucleotides in the MAC+IES assembly were first masked with ‘N’s. The Intronarrator pipeline (https://github.com/Swart-lab/Intronarrator), a wrapper around Augustus (Stanke and Waack, 2003), was run with the same parameters as for the *B. stoltei* MAC genome, i.e. a cut-off of 0.2 for the fraction of spliced reads covering a potential intron, and ≥10 reads to call an intron (Singh et al., 2021). Without masking, gene predictions around IESs were poor, with genuine MDS-limited genes (with high RNA-seq coverage) frequently incorrectly extended into IES regions. The possibility of genes spanning IES boundaries was not catered for.

Domain annotations for diagrams were generated with the InterproScan 5.44-79.0 pipeline (Jones et al., 2014) incorporating HMMER (v3.3, Nov 2019, hmmscan) (Eddy, 2011).

For comparison of transposase-related domain content in MAC vs. MIC, reference sequences were obtained from public databases for *Paramecium tetraurelia* (https://paramecium.i2bc.paris-saclay.fr/files/Paramecium/tetraurelia/51/annotations/ptetraurelia_mac_51_with_ies/), *Tetrahymena thermophila* (http://www.ciliate.org/system/downloads/3-upd-cds-fasta-2021.fasta), and *Oxytricha trifallax* (https://oxy.ciliate.org/common/downloads/oxy/Oxy2020_CDS.fasta, https://knot.math.usf.edu/mds_ies_db/data/gff/oxytri_mic_non_mds.gff). IES gene prediction in *Blepharisma* was hampered by intermittent polynucleotide tract length errors, due to the assembly of IESs from PacBio CLR reads. To mitigate this, a six-frame translation of the MIC-limited genome regions was performed using a custom script, then scanned against the Pfam-A database 32.0 (release 9) (Mistry et al., 2021) with hmmscan (HMMER), with i-E-value cutoff $\leq 10^{-6}$. Domains were annotated from the MAC genome with three different methods: using published coding sequences (“cds” in Table S4), six-frame translations (“6ft”), and six-frame translations split on stop codons (“6ft_split”).

Repeat annotation and clustering

To evaluate the repetitive sequence content in IESs, we applied a repeat prediction and annotation to the combined MAC+IES assembly, instead of clustering whole IESs by sequence similarity. This was so that: (i) Repeats shared between the MDS and IES could be identified. (ii) Complex structures such as nested repeats could be detected. (iii) Repeat families were predicted *de novo*, permitting discovery of novel elements. (iv) Repeats did not have to be strictly identical to be grouped into a family.

Interspersed repeat element families were predicted from the MAC+IES genome assembly with RepeatModeler v2.0.1 (default settings, random number seed 12345) with the following dependencies: rmblast v2.9.0+ (http://www.repeatmasker.org/RMBlast.html), TRF 4.09 (Benson, 1999), RECON (Bao and Eddy, 2002), RepeatScout 1.0.6 (Price et al., 2005), RepeatMasker v4.1.1 (http://www.repeatmasker.org/RMDownload.html). Repeat families were also classified in the pipeline by RepeatClassifier v2.0.1 through comparison against RepeatMasker's repeat protein database and the Dfam database. Consensus sequences of the predicted repeat families, produced by RepeatModeler, were then used to annotate repeats in the MAC+IES assembly with RepeatMasker, using rmblast as the search engine.

The consensus sequences for rnd-1_family-0 and rnd-1_family-73 were manually curated for downstream analyses. For rnd-1_family-0 (BogoMITE) the original consensus predicted by RepeatModeler for rnd-1_family-0 was 784 bp long, but this was a spurious inverted duplication of the basic ~390 bp unit; the duplication had been favored in the construction of the consensus because RepeatModeler attempts to find the longest possible match to represent each family. For family rnd-1_family-73 (containing BstTc1 transposon), the actual repeat unit was longer than the boundaries predicted by RepeatModeler. In most IESs that contain this repeat (19 of 22), it was flanked by and partially overlapping with short repeat elements from families rnd-4_family-1308 and rnd-1_family-117, which are spurious predictions. Repeat unit boundaries were manually defined by alignment of full length repeats and their flanking regions.

Terminal inverted repeats of selected repeat element families were identified by aligning the consensus sequence from RepeatModeler, and/or selected full-length elements, with their respective reverse complements using MAFFT (Katoh and Standley, 2013) (plugin version distributed with Geneious).

TIRs from the Dfam DNA transposon termini signatures database (v1.1, https://www.dfam.org/releases/dna_termini_1.1/dna_termini_1.1.hmm.gz) (Storer et al., 2021) were searched with hmmsearch (HMMer v3.2.1) against the IES sequences, to identify matches to TIR signatures of major transposon subfamilies.

Phylogenetic analysis of Tc1/Mariner-superfamily transposases

Repeat family rnd-1_family-1 was initially classified as a "TcMar/Tc2" family transposable element by RepeatClassifier. 30 full length copies (>95% of the consensus length) were annotated by RepeatMasker, all of which fell within IESs and contained CDS predictions. However, CDSs were of varying lengths because of frameshifts caused by indels, which may be biological or due to assembly error; nonetheless, the nucleotide sequences had high pairwise identity (about 98%, except for one outlier). We chose BSTOLATCC_MIC4025 as the representative CDS sequence for phylogenetic analysis because it was one of the longest predicted and both predicted Pfam domains (HTH_Tnp_Tc5 and DDE_1) appeared to be intact.

For repeat family rnd-1_family-73, the initial classification was "DNA/TcMar-Tc1". As described above, CDS predictions were of variable lengths, and the longest CDSs were not necessarily the best versions of the sequence because of potential frameshift errors. For phylogenetic analysis, we chose BSTOLATCC_MIC48344 as the representative copy, because a complete *DDE_3* Pfam domain was predicted by HMMER that could align with other DDE/D domains from reference alignments described below.

The representative CDSs of the rnd-1_family-1 and rnd-1_family-73 transposases were aligned with MAFFT (E-INS-i mode) against a published DDE/D domain reference alignment (Supporting Information Dataset_S01 of (Yuan and Wessler, 2011)) to identify the residues at the conserved catalytic triad and the amino acid distance between the conserved residues.

For the phylogenetic analysis of the DDE/D domains in the Tc1/Mariner superfamily, both MAC- and MIC-limited genes containing DDE_1 and DDE_3 domains were separately aligned for each Pfam domain with MAFFT v7.450 (algorithm: E-INS-i, scoring matrix: BLOSUM62, Gap open penalty: 1.53) and trimmed to the DDE/D domain with Geneious and incomplete domains were removed. As reference, 204 sequences from a published alignment (Additional File 4 of (Dupeyron et al., 2020)) were selected to represent the 53 groups defined in that study, choosing only complete domains (with all three conserved catalytic residues) and all *Oxytricha trifallax* TBE and *Euplotes crassus* Tec transposase sequences. Thirteen *Paramecium* Tc1/Mariner DDE/D domain consensus sequences were added (Additional File 4 of (Guérin et al., 2017)). Sequences were aligned with MAFFT (E-INS-i mode) and trimmed to only the DDE/D domain boundaries with Geneious. Phylogeny was inferred with FastTree2 v2.1.11 (Price et al., 2010) using the WAG substitution model. The tree was visualized with Dendroscope v3.5.10 (Huson and Scornavacca, 2012), rooted with bacterial IS630 sequences as outgroup

Phylogenetic analysis of retrotransposon-derived sequences

All the nucleotide sequences ≥500 bp for the repeat families identified by RepeatClassifier as LINE or LINE/RTE-x: rnd-1_family-273, rnd-1_family-276 and rnd-4_family-193 were aligned to one another with MAFFT v7.450 (automatic algorithm) (Katoh and Standley, 2013), with the option to automatically determine sequence direction (via the MAFFT plugin for Geneious Prime (Kearse et al., 2012)). Since the alignment appeared to be poor between the rnd-4-family-193 sequences and the rest, we generated separate alignments for this family from the other two, also with MAFFT (E-INS-i mode). Maximum likelihood phylogenies were generated by PhyML (Guindon et al., 2010) version 3.3.20180621 with the HKY85 substitution model.

Data availability

Annotated draft MAC+IES genome for *Blepharisma stoltei* strain ATCC 30299 (European Nucleotide Archive (ENA) Bioproject PRJEB46944 under accession GCA_914767885). IES sequences and annotations, MAC gene predictions with intervening IESs, and gene predictions within IESs (EDMOND, doi:10.17617/3.83; genome browser, https://bleph.ciliate.org. Sequencing data for the MIC-enriched nuclear fractions (PacBio CLR reads: ENA accession ERR6510520 and ERR6548140; BGI-seq reads: ENA accessions ERR6474675, ERR6496962, ERR6497067, ERR6501836). Small RNA libraries from developmental time series (ENA Bioproject PRJEB47200 under accessions ERR6565537-ERR6565561). Repeat family predictions and annotations by RepeatModeler and RepeatMasker (EDMOND, doi:10.17617/3.82). Alignment and phylogeny of Tc1/Mariner superfamily transposase domains (EDMOND, doi:10.17617/3.JLWBFM)

Acknowledgements

We thank C. Lanz for assistance with DNA quality control, A. Noll for computer system administration, L.A. Klobutcher, G. Herrick, O. Weichenrieder and A. Streit for helpful discussions. Research reported in this publication was supported by the National Institutes of Health (award No. P40OD010964) to N.A.S and the Max Planck Society.

Author contributions

Data curation: B.K.B.S., E.C.S., M.S.1. Formal analysis: B.K.B.S., M.S.1, E.C.S., C.W. Funding acquisition: N.S., E.C.S. Investigation: B.K.B.S., M.S.1., E.C.S. Methodology: B.K.B.S., M.S.1., E.C.S., M.S.2, T.H., A.S., C.E. Resources: M.S.2., T.H., C.W., B.H., A.B., N.S. Software: B.K.B.S., E.C.S., M.S.1, A.B., N.S. Supervision: M.S.2, T.H., E.C.S. Visualization: B.K.B.S., M.S.1., E.C.S. Writing – original draft: B.K.B.S., M.S.1., E.C.S. Writing – review & editing: B.K.B.S., M.S.1, E.C.S., M.S.2, T.H., A.S., C.W., N.S.

Declaration of interests

The authors declare no competing interests.

References

Arnaiz, O., Mathy, N., Baudry, C., Malinsky, S., Aury, J.-M., Denby Wilkes, C., Garnier, O., Labadie, K., Lauderdale, B.E., Le Mouël, A., et al. (2012). The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. PLoS Genet. *8*, e1002984.

Bao, Z., and Eddy, S.R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. *12*, 1269–1276.

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., and Bétermier, M. (2009). PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev. *23*, 2478–2483.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Bestor, T.H. (1990). DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. Philos. Trans. R. Soc. Lond. B Biol. Sci. *326*, 179–187.

Bischerour, J., Bhullar, S., Denby Wilkes, C., Régnier, V., Mathy, N., Dubois, E., Singh, A., Swart, E., Arnaiz, O., Sperling, L., et al. (2018). Six domesticated PiggyBac transposases together carry out programmed DNA elimination in Paramecium. ELife *7*.

Catania, F., Rothering, R., and Vitali, V. (2021). One cell, two gears: extensive somatic genome plasticity accompanies high germline genome stability in *Paramecium*. Genome Biol. Evol. *13*.

Chalker, D.L., Meyer, E., and Mochizuki, K. (2013). Epigenetics of ciliates. Cold Spring Harb. Perspect. Biol. *5*, a017764.

Cheng, C.-Y., Vogt, A., Mochizuki, K., and Yao, M.-C. (2010). A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in Tetrahymena thermophila. Mol. Biol. Cell *21*, 1753–1762.

Cheng, C.-Y., Young, J.M., Lin, C.-Y.G., Chao, J.-L., Malik, H.S., and Yao, M.-C. (2016). The piggyBac transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during the developmental rearrangement of key genes in *Tetrahymena thermophila*. Genes Dev. *30*, 2724–2736.

Chen, Q., Luo, W., Veach, R.A., Hickman, A.B., Wilson, M.H., and Dyda, F. (2020). Structural basis of seamless excision and specific targeting by piggyBac transposase. Nat. Commun. *11*, 3446.

Chen, X., Bracht, J.R., Goldman, A.D., Dolzhenko, E., Clay, D.M., Swart, E.C., Perlman, D.H., Doak, T.G., Stuart, A., Amemiya, C.T., et al. (2014). The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell *158*, 1187–1198.

Coyne, R.S., Lhuillier-Akakpo, M., and Duharcourt, S. (2012). RNA-guided DNA rearrangements in ciliates: is the best genome defence a good offence? Biol. Cell *104*, 309–325.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Dale, R.K., Pedersen, B.S., and Quinlan, A.R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics *27*, 3423–3424.

Drews, F., Salhab, A., Karunanithi, S., Cheaib, M., Jung, M., Schulz, M.H., and Simon, M. (2021). Broad domains of histone marks in the highly compact *Paramecium* macronuclear genome. BioRxiv.

Dupeyron, M., Baril, T., Bass, C., and Hayward, A. (2020). Phylogenetic analysis of the Tc1/mariner superfamily reveals the unexplored diversity of pogo-like elements. Mob. DNA *11*, 21.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Fang, W., Wang, X., Bracht, J.R., Nowacki, M., and Landweber, L.F. (2012). Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement. Cell *151*, 1243–1255.

Feng, Y., and Landweber, L.F. (2021). Transposon debris in ciliate genomes. PLoS Biol. *19*, e3001354.

Feng, L., Wang, G., Hamilton, E.P., Xiong, J., Yan, G., Chen, K., Chen, X., Dui, W., Plemens, A., Khadr, L., et al. (2017). A germline-limited piggyBac transposase gene is required for precise excision in *Tetrahymena* genome rearrangement. Nucleic Acids Res. *45*, 9481–9502.

Feschotte, C., and Mouchès, C. (2000). Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. Mol. Biol. Evol. *17*, 730–737.

Feschotte, C., Zhang, X., and Wessler, S.R. (2002). Miniature inverted-repeat trasnposable elements and their relationship to established DNA transposons. In Mobile DNA II, N.L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz, eds. (Washington, D.C.: ASM Press), pp. 1147–1158.

Feschotte, C., Osterlund, M.T., Peeler, R., and Wessler, S.R. (2005). DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. Nucleic Acids Res. *33*, 2153–2165.

Fillingham, J.S., Thing, T.A., Vythilingum, N., Keuroghlian, A., Bruno, D., Golding, G.B., and Pearlman, R.E. (2004). A non-long terminal repeat retrotransposon family is restricted to the germ line micronucleus of the ciliated protozoan *Tetrahymena thermophila*. Eukaryotic Cell *3*, 157–169.

Fritz, G. (2000). Human APE/Ref-1 protein. Int. J. Biochem. Cell Biol. *32*, 925–929.

Gao, F., and Katz, L.A. (2014). Phylogenomic analyses support the bifurcation of ciliates into two major clades that differ in properties of nuclear division. Mol. Phylogenet. Evol. *70*, 240–243.

Gao, B., Wang, Y., Diaby, M., Zong, W., Shen, D., Wang, S., Chen, C., Wang, X., and Song, C. (2020). Evolution of pogo, a separate superfamily of IS630-Tc1-mariner transposons, revealing recurrent domestication events in vertebrates. Mob. DNA *11*, 25.

Giese, A.C. (1973). *Blepharisma*: The Biology of a Light-sensitive Protozoan (Stanford University Press).

Grewal, S.I.S., and Jia, S. (2007). Heterochromatin revisited. Nat. Rev. Genet. *8*, 35–46.

Guérin, F., Arnaiz, O., Boggetto, N., Denby Wilkes, C., Meyer, E., Sperling, L., and Duharcourt, S. (2017). Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements. BMC Genomics *18*, 327.

Guermonprez, H., Loot, C., and Casacuberta, J.M. (2008). Different strategies to persist: the pogo-like Lemi1 transposon produces miniature inverted-repeat transposable elements or typical defective elements in different plant genomes. Genetics *180*, 83–92.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Hamilton, E.P., Kapusta, A., Huvos, P.E., Bidwell, S.L., Zafar, N., Tang, H., Hadjithomas, M., Krishnakumar, V., Badger, J.H., Caler, E.V., et al. (2016). Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome.

ELife *5*.

Han, J.S. (2010). Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mob. DNA *1*, 15.

Hartl, D.L., Lohe, A.R., and Lozovskaya, E.R. (1997). Modern thoughts on an ancyent marinere: function, evolution, regulation. Annu. Rev. Genet. *31*, 337–358.

Harumoto, T., Miyake, A., Ishikawa, N., Sugibayashi, R., Zenfuku, K., and Iio, H. (1998). Chemical defense by means of pigmented extrusomes in the ciliate *Blepharisma japonicum*. Eur. J. Protistol. *34*, 458–470.

Herrick, G., Cartinhour, S., Dawson, D., Ang, D., Sheets, R., Lee, A., and Williams, K. (1985). Mobile elements bounded by C4A4 telomeric repeats in *Oxytricha fallax*. Cell *43*, 759–768.

Huff, J.T., Zilberman, D., and Roy, S.W. (2016). Mechanism for DNA transposons to generate introns on genomic scales. Nature *538*, 533–536.

Huson, D.H., and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. *61*, 1061–1067.

Jahn, C.L., Doktor, S.Z., Frels, J.S., Jaraczewski, J.W., and Krikau, M.F. (1993). Structures of the *Euplotes crassus* Tec1 and Tec2 elements: identification of putative transposase coding regions. Gene *133*, 71–78.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.

Klobutcher, L.A., and Herrick, G. (1995). Consensus inverted terminal repeat sequence of *Paramecium* IESs: resemblance to termini of Tc1-related and *Euplotes* Tec transposons. Nucleic Acids Res. *23*, 2006–2013.

Klobutcher, L.A., and Herrick, G. (1997). Developmental genome reorganization in ciliated protozoa: the transposon link. Prog. Nucleic Acid Res. Mol. Biol. *56*, 1–62.

Kubota, T., Tokoroyama, T., Tsukuda, Y., Koyama, H., and Miyake, A. (1973). Isolation and structure determination of blepharismin, a conjugation initiating gamone in the ciliate blepharisma. Science *179*, 400–402.

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

Lauth, M.R., Spear, B.B., Heumann, J., and Prescott, D.M. (1976). DNA of ciliated protozoa: DNA sequence diminution during macronuclear development of *Oxytricha*. Cell *7*, 67–74.

Le Mouël, A., Butler, A., Caron, F., and Meyer, E. (2003). Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in paramecia. Eukaryotic Cell *2*, 1076–1090.

Lepère, G., Nowacki, M., Serrano, V., Gout, J.-F., Guglielmi, G., Duharcourt, S., and Meyer, E. (2009). Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*. Nucleic Acids Res. *37*, 903–915.

Lhuillier-Akakpo, M., Frapporti, A., Denby Wilkes, C., Matelot, M., Vervoort, M., Sperling, L., and Duharcourt, S. (2014). Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements. PLoS Genet. *10*, e1004665.

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

Liu, Y., Taverna, S.D., Muratore, T.L., Shabanowitz, J., Hunt, D.F., and Allis, C.D. (2007). RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in Tetrahymena. Genes Dev. *21*, 1530–1545.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lynn, D.H. (2010). The Ciliated Protozoa (Dordrecht: Springer Netherlands).

Malone, C.D., Anderson, A.M., Motl, J.A., Rexer, C.H., and Chalker, D.L. (2005). Germ line transcripts are processed by a Dicer-like protein that is essential for developmentally programmed genome rearrangements of *Tetrahymena thermophila*. Mol. Cell. Biol. *25*, 9151–9164.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419.

Mitra, R., Li, X., Kapusta, A., Mayhew, D., Mitra, R.D., Feschotte, C., and Craig, N.L. (2013). Functional characterization of piggyBat from the bat *Myotis lucifugus* unveils an active mammalian DNA transposon. Proc Natl Acad Sci USA *110*, 234–239.

Miyake, A., and Beyer, J. (1974). Blepharmone: a conjugation-inducing glycoprotein in the ciliate blepharisma. Science *185*, 621–623.

Miyake, A., Rivola, V., and Harumoto, T. (1991). Double paths of macronucleus differentiation at conjugation in *Blepharisma japonicum*. Eur. J. Protistol. *27*, 178–200.

Mochizuki, K., and Gorovsky, M.A. (2005). A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. Genes Dev. *19*, 77–89.

Mochizuki, K., and Kurth, H.M. (2013). Loading and pre-loading processes generate a distinct siRNA population in *Tetrahymena*. Biochem. Biophys. Res. Commun. *436*, 497–502.

Mochizuki, K., Fine, N.A., Fujisawa, T., and Gorovsky, M.A. (2002). Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. Cell *110*, 689–699.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Nowacki, M., Higgins, B.P., Maquilan, G.M., Swart, E.C., Doak, T.G., and Landweber, L.F. (2009). A functional role for transposases in a large eukaryotic genome. Science *324*, 935–938.

Prescott, D.M., and Greslin, A.F. (1992). Scrambled actin I gene in the micronucleus of *Oxytricha nova*. Dev. Genet. *13*, 66–74.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21 Suppl 1*, i351-8.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 — approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

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

Repak, A.J. (1968). Encystment and excystment of the heterotrichous ciliate *Blepharisma stoltei* Isquith. Journal of Protozoology *5*, 407–412.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ *4*, e2584.

Rzeszutek, I., Maurer-Alcalá, X.X., and Nowacki, M. (2020). Programmed genome rearrangements in ciliates. Cell. Mol. Life Sci. *77*, 4615–4629.

Sandoval, P.Y., Swart, E.C., Arambasic, M., and Nowacki, M. (2014). Functional diversification of Dicer-like proteins and small RNAs required for genome sculpting. Dev. Cell *28*, 174–188.

Seah, B.K.B., and Swart, E.C. (2021). BleTIES: Annotation of natural genome editing in ciliates using long read sequencing. Bioinformatics *37*, 3929–3931.

Sellis, D., Guérin, F., Arnaiz, O., Pett, W., Lerat, E., Boggetto, N., Krenek, S., Berendonk, T., Couloux, A., Aury, J.-M., et al. (2021). Massive colonization of protein-coding exons by selfish genetic elements in *Paramecium* germline genomes. PLoS Biol. *19*, e3001309.

Singh, M., Seah, B.K.B., Emmerich, C., Singh, A., Woehle, C., Huettel, B., Byerly, A., Stover, N.A., Sugiura, M., Harumoto, T., et al. (2021). The *Blepharisma stoltei* macronuclear genome: towards the origins of whole genome reorganization. BioRxiv.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics *19 Suppl 2*, ii215-25.

Stern, A., Keren, L., Wurtzel, O., Amitai, G., and Sorek, R. (2010). Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet. *26*, 335–340.

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

Sugiura, M., and Harumoto, T. (2001). Identification, characterization, and complete amino acid sequence of the conjugation-inducing glycoprotein (blepharmone) in the ciliate Blepharisma japonicum. Proc Natl Acad Sci USA *98*, 14446–14451.

Swart, E.C., Bracht, J.R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J.S., Goldman, A.D., Nowacki, M., Schotanus, K., et al. (2013). The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol. *11*, e1001473.

Swart, E.C., Wilkes, C.D., Sandoval, P.Y., Arambasic, M., Sperling, L., and Nowacki, M. (2014). Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion. Nucleic Acids Res. *42*, 8970–8983.

Taverna, S.D., Coyne, R.S., and Allis, C.D. (2002). Methylation of histone h3 at lysine 9 targets programmed DNA elimination in tetrahymena. Cell *110*, 701–711.

Udomkit, A., Forbes, S., Dalgleish, G., and Finnegan, D.J. (1995). BS a novel LINE-like element in *Drosophila melanogaster*. Nucleic Acids Res. *23*, 1354–1358.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. *27*, 737–746.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

Vitali, V., Hagen, R., and Catania, F. (2019). Environmentally induced plasticity of programmed DNA elimination boosts somatic variability in *Paramecium tetraurelia*. Genome Res. *29*, 1693–1704.

Vogt, A., and Mochizuki, K. (2013). A domesticated PiggyBac transposase interacts with heterochromatin and catalyzes reproducible DNA elimination in *Tetrahymena*. PLoS Genet. *9*, e1004032.

Wang, S., Zhang, L., Meyer, E., and Matz, M.V. (2010). Characterization of a group of MITEs with unusual features from two coral genomes. PLoS ONE *5*, e10700.

Weiss, A. (2015). Lamarckian Illusions. Trends Ecol. Evol. *30*, 566–568.

Wuitschick, J.D., Gershan, J.A., Lochowicz, A.J., Li, S., and Karrer, K.M. (2002). A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*. Nucleic Acids Res. *30*, 2524–2537.

Yao, M.-C., Fuller, P., and Xi, X. (2003). Programmed DNA deletion as an RNA-guided system of genome defense. Science *300*, 1581–1584.

Yuan, Y.-W., and Wessler, S.R. (2011). The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci USA *108*, 7884–7889.

Zahler, A.M., Neeb, Z.T., Lin, A., and Katzman, S. (2012). Mating of the stichotrichous ciliate *Oxytricha trifallax* induces production of a class of 27 nt small RNAs derived from the parental macronucleus. PLoS ONE *7*, e42371.

Zhou, W., Liang, G., Molloy, P.L., and Jones, P.A. (2020). DNA methylation enables transposable element-driven genome expansion. Proc Natl Acad Sci USA *117*, 19359–19366.

Figure 1.

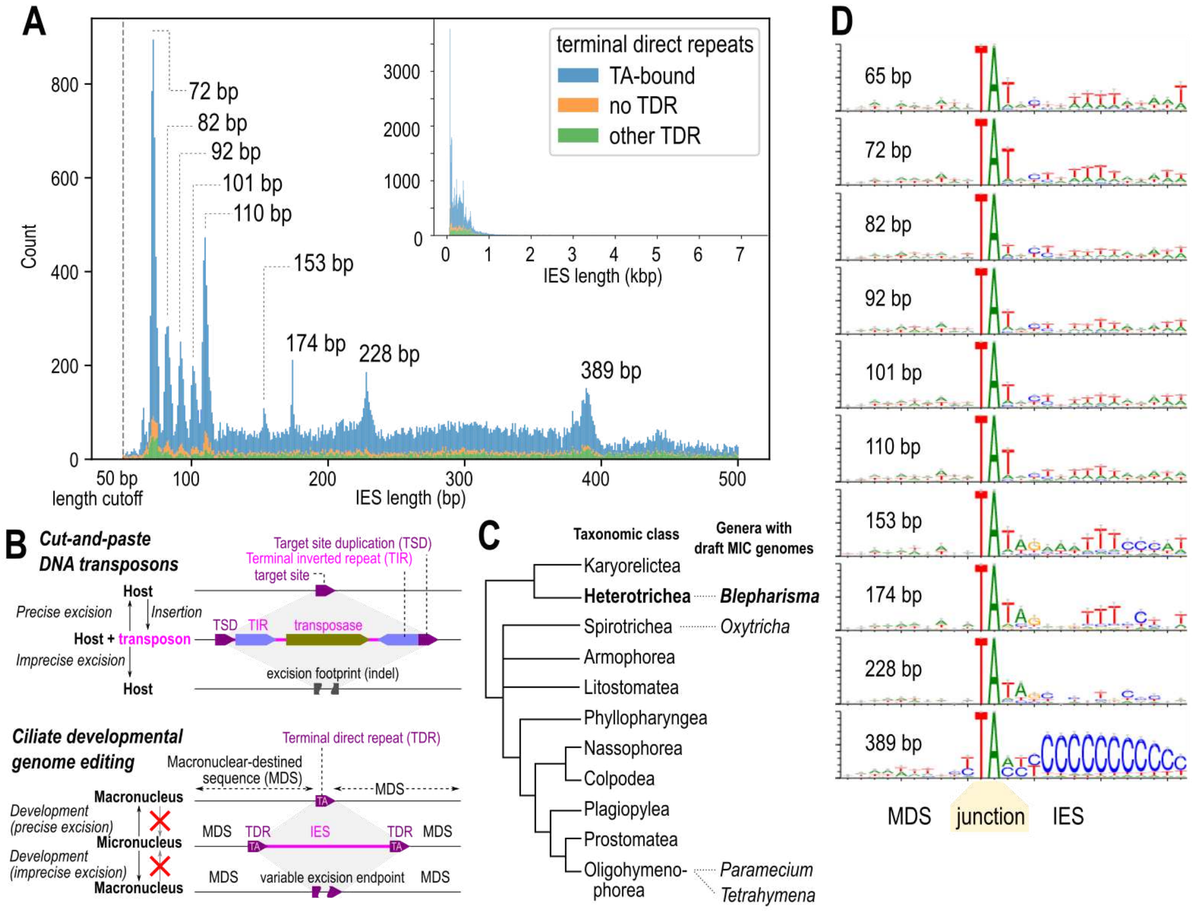

A
800
600
400
200
0
Count
72 bp
82 bp
92 bp
101 bp
110 bp
153 bp
174 bp
228 bp
389 bp
50 bp
length cutoff
100
200
300
400
500
IES length (bp)
3000
2000
1000
0
0 1 2 3 4 5 6 7
IES length (kbp)
terminal direct repeats
TA-bound
no TDR
other TDR
B
Cut-and-paste
DNA transposons
Target site duplication (TSD)
Terminal inverted repeat (TIR)
target site
Host
Precise excision
Insertion
Host + transposon
TSD TIR transposase
Imprecise excision
excision footprint (indel)
Host
Ciliate developmental
genome editing
Terminal direct repeat (TDR)
Macronuclear-destined
sequence (MDS)
MDS
Macronucleus
Development
(precise excision)
Micronucleus
MDS TDR IES TDR MDS
TA
Development
(imprecise excision)
Macronucleus
MDS variable excision endpoint MDS
C
Taxonomic class
Genera with
draft MIC genomes
Karyorelictea
Heterotrichea
Blepharisma
Spirotrichea
Oxytricha
Armophorea
Litostomatea
Phyllopharyngea
Nassophorea
Colpodea
Plagiopylea
Prostomatea
Oligohymeno-
phorea
Paramecium
Tetrahymena
D
65 bp
72 bp
82 bp
92 bp
101 bp
110 bp
153 bp
174 bp
228 bp
389 bp
MDS
junction
IES

Figure 2.

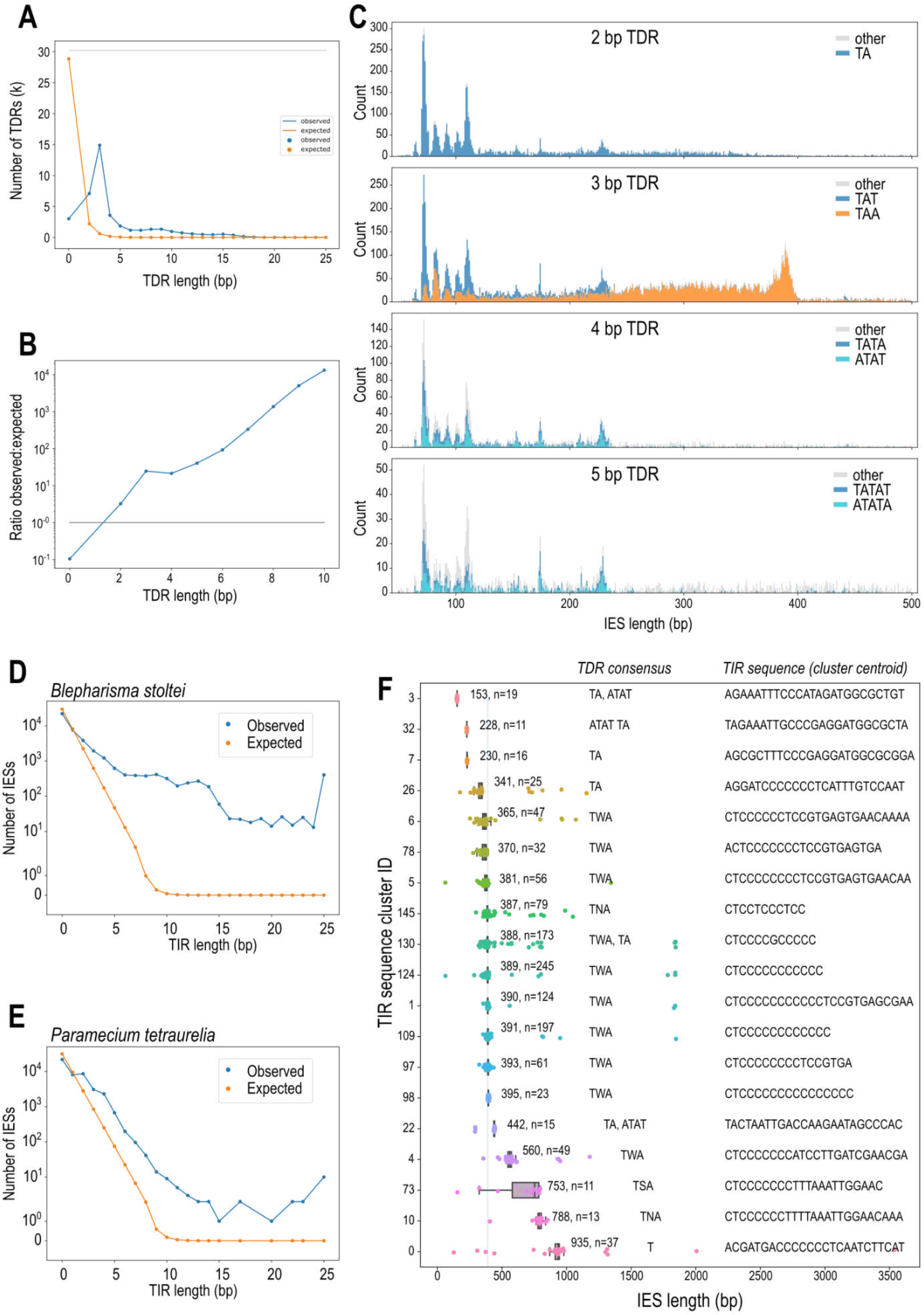

A
Number of TDRs (k)
30
25
20
15
10
5
0
TDR length (bp)
0
5
10
15
20
25
observed
expected
observed
expected
B
Ratio observed:expected
10^4
10^3
10^2
10^1
10^{-0}
10^{-1}
0
2
4
6
8
10
TDR length (bp)
C
2 bp TDR
other
TA
3 bp TDR
other
TAT
TAA
4 bp TDR
other
TATA
ATAT
5 bp TDR
other
TATAT
ATATA
Count
300
250
200
150
100
50
0
140
120
80
60
40
20
IES length (bp)
100
200
300
400
500
D
Blepharisma stoltei
Observed
Expected
Number of IESs
10^4
10^3
10^2
10^1
10^0
0
TIR length (bp)
0
5
10
15
20
25
E
Paramecium tetraurelia
Observed
Expected
Number of IESs
TIR length (bp)
F
TDR consensus
TIR sequence (cluster centroid)
TIR sequence cluster ID
3
153, n=19
TA, ATAT
AGAAATTTCCCATAGATGGCGCTGT
32
228, n=11
ATAT TA
TAGAAATTGCCCGAGGATGGCGCTA
7
230, n=16
TA
AGCGCTTTCCCGAGGATGGCGCGGA
26
341, n=25
TA
AGGATCCCCCCCTCATTTGTCCAAT
6
365, n=47
TWA
CTCCCCCCTCCGTGAGTGAACAAAA
78
370, n=32
TWA
ACTCCCCCCCTCCGTGAGTGA
5
381, n=56
TWA
CTCCCCCCCCTCCGTGAGTGAACAA
145
387, n=79
TNA
CTCCTCCCTCC
130
388, n=173
TWA, TA
CTCCCCGCCCCC
124
389, n=245
TWA
CTCCCCCCCCCCC
1
390, n=124
TWA
CTCCCCCCCCCCCTCCGTGAGCGAA
109
391, n=197
TWA
CTCCCCCCCCCCCC
97
393, n=61
TWA
CTCCCCCCCCTCCGTGA
98
395, n=23
TWA
CTCCCCCCCCCCCCCCC
22
442, n=15
TA, ATAT
TACTAATTGACCAAGAATAGCCCAC
4
560, n=49
TWA
CTCCCCCCCATCCTTGATCGAACGA
73
753, n=11
TSA
CTCCCCCCCTTTAAATTGGAAC
10
788, n=13
TNA
CTCCCCCCTTTTAAATTGGAACAAA
0
935, n=37
T
ACGATGACCCCCCCTCAATCTTCAT
0
500
1000
1500
2000
2500
3000
3500
IES length (bp)

Figure 3.

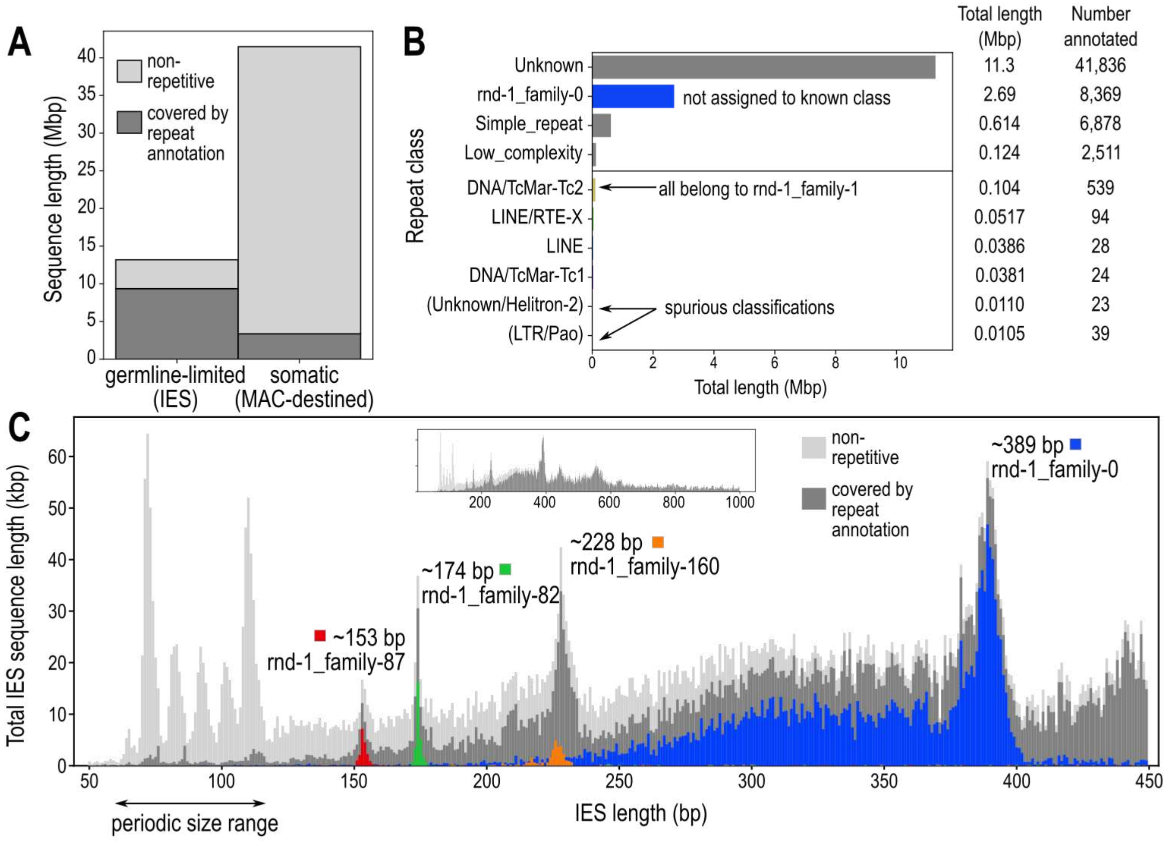

A
Sequence length (Mbp)
0
5
10
15
20
25
30
35
40
non-repetitive
covered by repeat annotation
germline-limited (IES)
somatic (MAC-destined)
B
Repeat class
Total length (Mbp)
Number annotated
Unknown 11.3 41,836
rnd-1_family-0 2.69 8,369
not assigned to known class
Simple_repeat 0.614 6,878
Low_complexity 0.124 2,511
DNA/TcMar-Tc2 0.104 539
all belong to rnd-1_family-1
LINE/RTE-X 0.0517 94
LINE 0.0386 28
DNA/TcMar-Tc1 0.0381 24
(Unknown/Helitron-2) 0.0110 23
(LTR/Pao) 0.0105 39
spurious classifications
0
2
4
6
8
10
Total length (Mbp)
C
Total IES sequence length (kbp)
0
10
20
30
40
50
60
200
400
600
800
1000
non-repetitive
covered by repeat annotation
~389 bp
rnd-1_family-0
~228 bp
rnd-1_family-160
~174 bp
rnd-1_family-82
~153 bp
rnd-1_family-87
50
100
150
200
250
300
350
400
450
IES length (bp)
periodic size range

Figure 4.

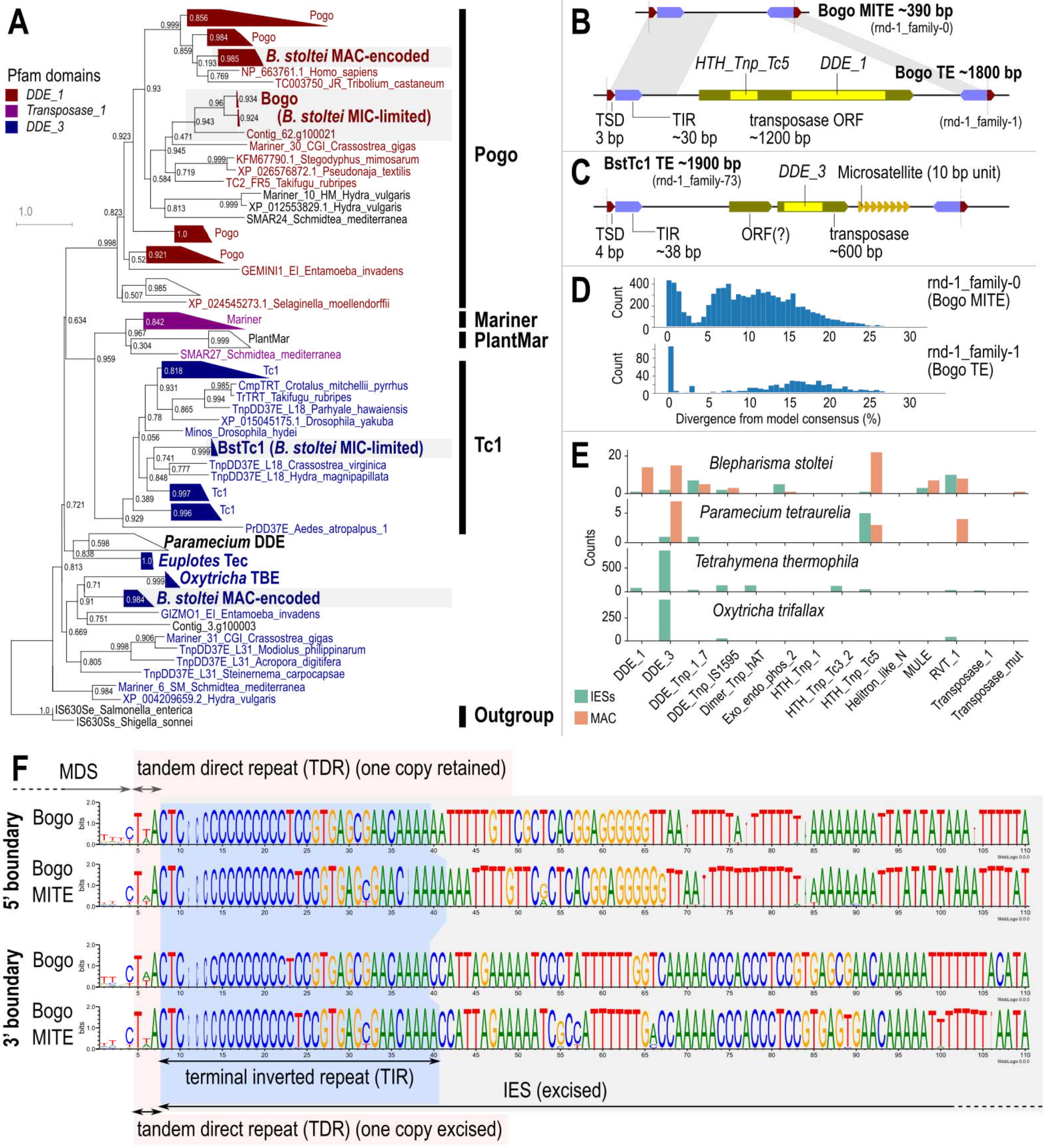

A
Pfam domains
DDE_1
Transposase_1
DDE_3
1.0
Pogo
B. stoltei MAC-encoded
NP_663761.1_Homo_sapiens
TC003750_JR_Tribolium_castaneum
Bogo
(B. stoltei MIC-limited)
Contig_62.g100021
Mariner_30_CGI_Crassostrea_gigas
KFM67790.1_Stegodyphus_mimosarum
XP_026576872.1_Pseudonaja_textilis
TC2_FR5_Takifugu_rubripes
Mariner_10_HM_Hydra_vulgaris
XP_012553829.1_Hydra_vulgaris
SMAR24_Schmidtea_mediterranea
GEMINI1_EI_Entamoeba_invadens
XP_024545273.1_Selaginella_moellendorffii
Mariner
PlantMar
SMAR27_Schmidtea_mediterranea
Tc1
CmpTRT_Crotalus_mitchellii_pyrrhus
TrTRT_Takifugu_rubripes
TnpDD37E_L18_Parhyale_hawaiensis
XP_015045175.1_Drosophila_yakuba
Minos_Drosophila_hydei
BstTc1 (B. stoltei MIC-limited)
TnpDD37E_L18_Crassostrea_virginica
TnpDD37E_L18_Hydra_magnipapillata
PrDD37E_Aedes_atropalpus_1
Paramecium DDE
Euplotes Tec
Oxytricha TBE
B. stoltei MAC-encoded
GIZMO1_EI_Entamoeba_invadens
Contig_3.g100003
Mariner_31_CGI_Crassostrea_gigas
TnpDD37E_L31_Modiolus_philippinarum
TnpDD37E_L31_Acropora_digitifera
TnpDD37E_L31_Steinernema_carpocapsae
Mariner_6_SM_Schmidtea_mediterranea
XP_004209659.2_Hydra_vulgaris
IS630Se_Salmonella_enterica
IS630Ss_Shigella_sonnei
Pogo
Mariner
PlantMar
Tc1
Outgroup
B
Bogo MITE ~390 bp
(rnd-1_family-0)
HTH_Tnp_Tc5
DDE_1
Bogo TE ~1800 bp
TSD
3 bp
TIR
~30 bp
transposase ORF
~1200 bp
(rnd-1_family-1)
C
BstTc1 TE ~1900 bp
(rnd-1_family-73)
DDE_3
Microsatellite (10 bp unit)
TSD
4 bp
TIR
~38 bp
ORF(?)
transposase
~600 bp
D
rnd-1_family-0
(Bogo MITE)
rnd-1_family-1
(Bogo TE)
Count
Divergence from model consensus (%)
E
Blepharisma stoltei
Paramecium tetraurelia
Tetrahymena thermophila
Oxytricha trifallax
Counts
IESs
MAC
DDE_1
DDE_3
DDE_Tnp_1_7
DDE_Tnp_IS1595
Dimer_Tnp_hAT
Exo_endo_phos_2
HTH_Tnp_1
HTH_Tnp_Tc3_2
HTH_Tnp_Tc5
Helitron_like_N
MULE
RVT_1
Transposase_1
Transposase_mut
F
MDS
tandem direct repeat (TDR) (one copy retained)
5' boundary
3' boundary
Bogo
Bogo MITE
bits
terminal inverted repeat (TIR)
IES (excised)
tandem direct repeat (TDR) (one copy excised)

Figure 5.

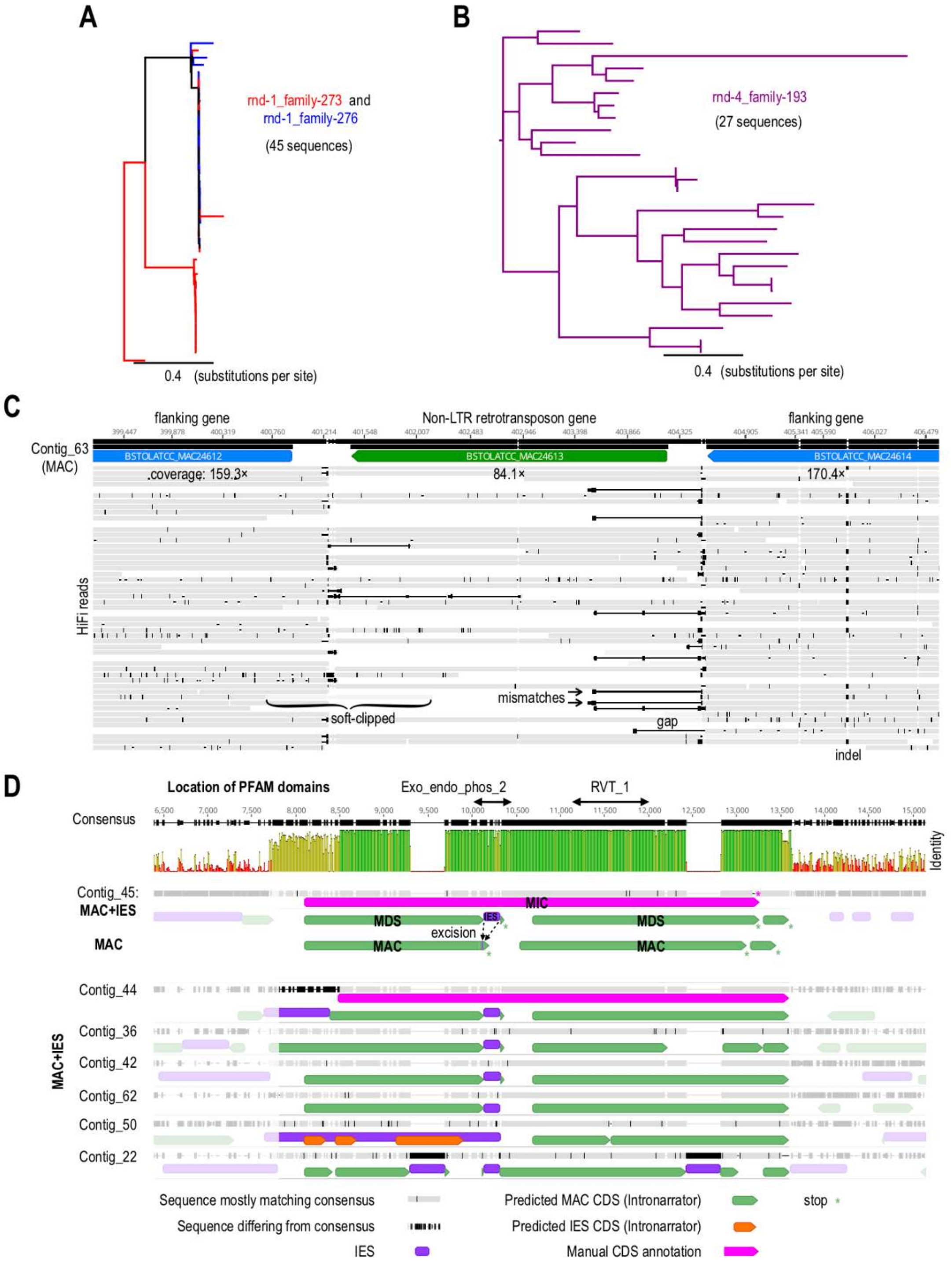

A
rnd-1_family-273 and
rnd-1_family-276
(45 sequences)
0.4 (substitutions per site)
B
rnd-4_family-193
(27 sequences)
0.4 (substitutions per site)
C
flanking gene
Non-LTR retrotransposon gene
flanking gene
399,447
399,878
400,319
400,760
401,214
401,548
402,007
402,483
402,946
403,398
403,866
404,325
404,905
405,341
405,590
406,027
406,479
Contig_63
(MAC)
BSTOLATCC_MAC24612
BSTOLATCC_MAC24613
BSTOLATCC_MAC24614
coverage: 159.3×
84.1×
170.4×
HiFi reads
mismatches
soft-clipped
gap
indel
D
Location of PFAM domains
Exo_endo_phos_2
RVT_1
Consensus
6,500
7,000
7,500
8,000
8,500
9,000
9,500
10,000
10,500
11,000
11,500
12,000
12,500
13,000
13,500
14,000
14,500
15,000
Identity
Contig_45:
MAC+IES
MIC
MDS
IES
MDS
excision
MAC
MAC
MAC
Contig_44
Contig_36
Contig_42
Contig_62
Contig_50
Contig_22
MAC+IES
Sequence mostly matching consensus
Sequence differing from consensus
IES
Predicted MAC CDS (Intronarrator)
Predicted IES CDS (Intronarrator)
Manual CDS annotation
stop

Figure 6.

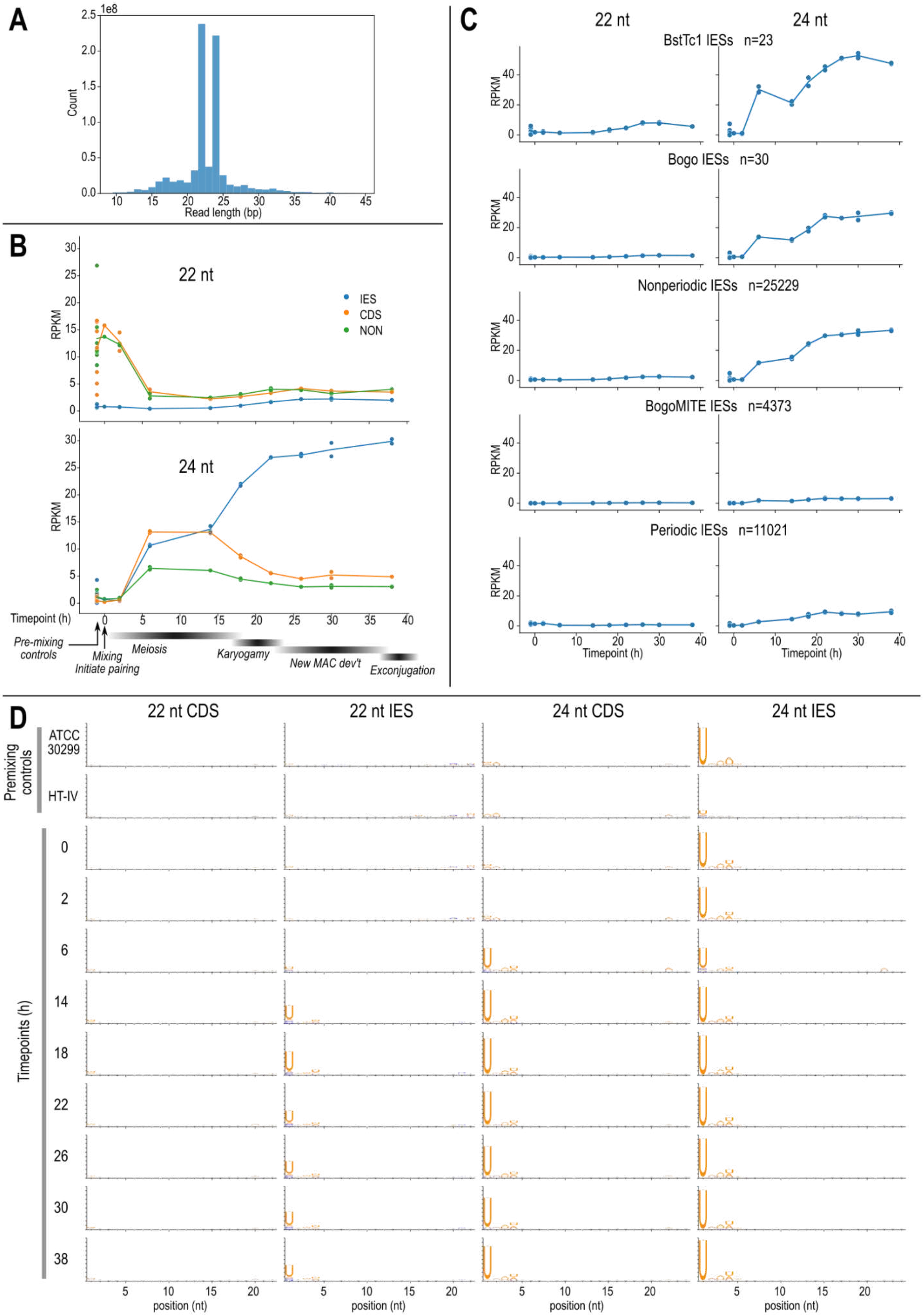

A
1e8
2.5
2.0
1.5
1.0
0.5
0.0
Count
10
15
20
25
30
35
40
45
Read length (bp)
B
22 nt
IES
CDS
NON
RPKM
24 nt
Timepoint (h)
0
5
10
15
20
25
30
35
40
Pre-mixing controls
Mixing
Initiate pairing
Meiosis
Karyogamy
New MAC dev't
Exconjugation
C
22 nt
24 nt
BstTc1 IESs n=23
Bogo IESs n=30
Nonperiodic IESs n=25229
BogoMITE IESs n=4373
Periodic IESs n=11021
RPKM
Timepoint (h)
D
22 nt CDS
22 nt IES
24 nt CDS
24 nt IES
Premixing controls
ATCC 30299
HT-IV
Timepoints (h)
0
2
6
14
18
22
26
30
38
position (nt)

Figure 7.

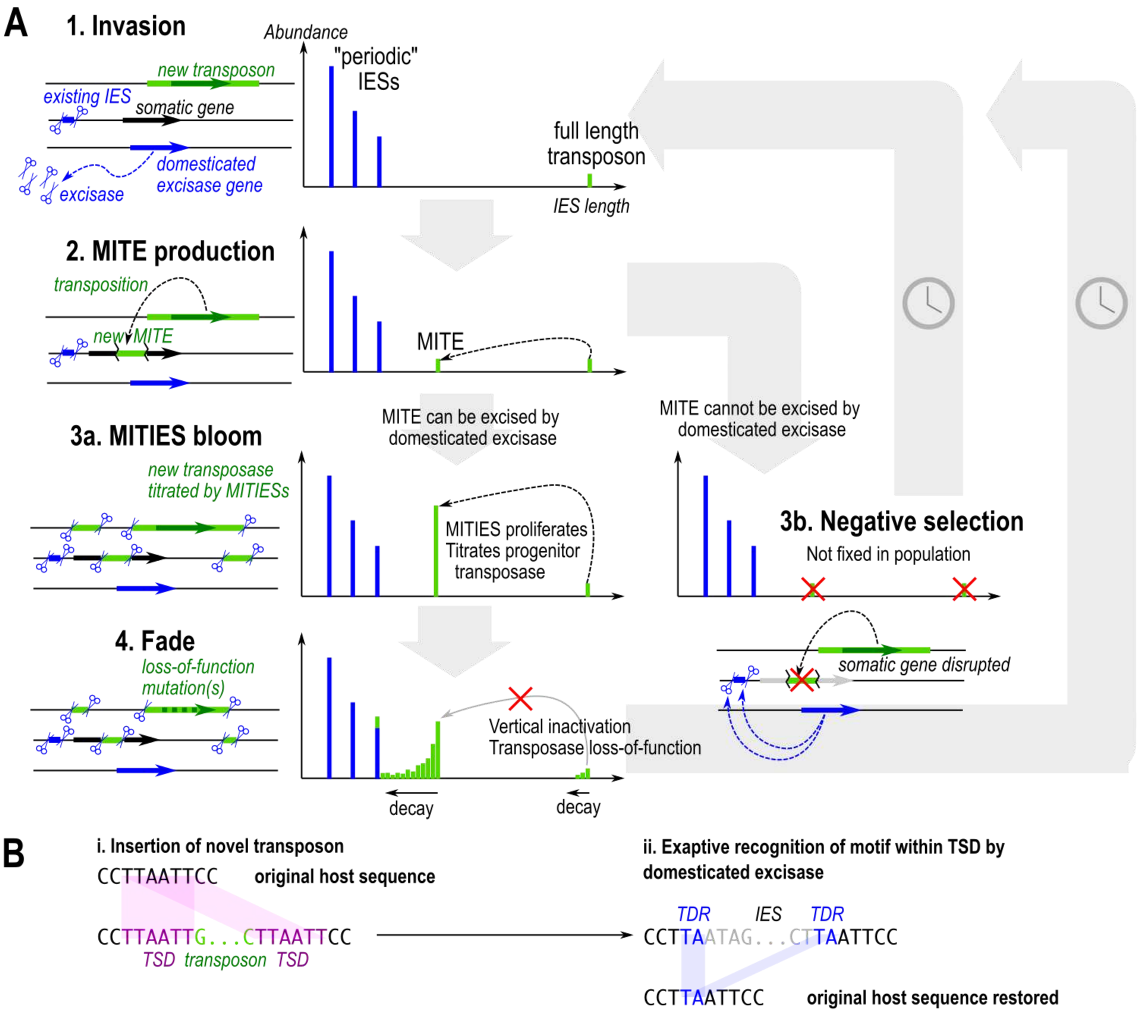

A
1. Invasion
new transposon
existing IES
somatic gene
domesticated
excisase gene
excisase
Abundance
"periodic"
IESs
full length
transposon
IES length
2. MITE production
transposition
new MITE
MITE
3a. MITIES bloom
new transposase
titrated by MITIESs
MITE can be excised by
domesticated excisase
MITIES proliferates
Titrates progenitor
transposase
MITE cannot be excised by
domesticated excisase
3b. Negative selection
Not fixed in population
somatic gene disrupted
4. Fade
loss-of-function
mutation(s)
Vertical inactivation
Transposase loss-of-function
decay
decay
B
i. Insertion of novel transposon
CCTTAATTCC
original host sequence
CCTTAATTG...CTTAATTCC
TSD transposon TSD
ii. Exaptive recognition of motif within TSD by
domesticated excisase
TDR
IES
TDR
CCTTAATAG...CTTAATTCC
CCTTAATTCC
original host sequence restored

Figure S1.

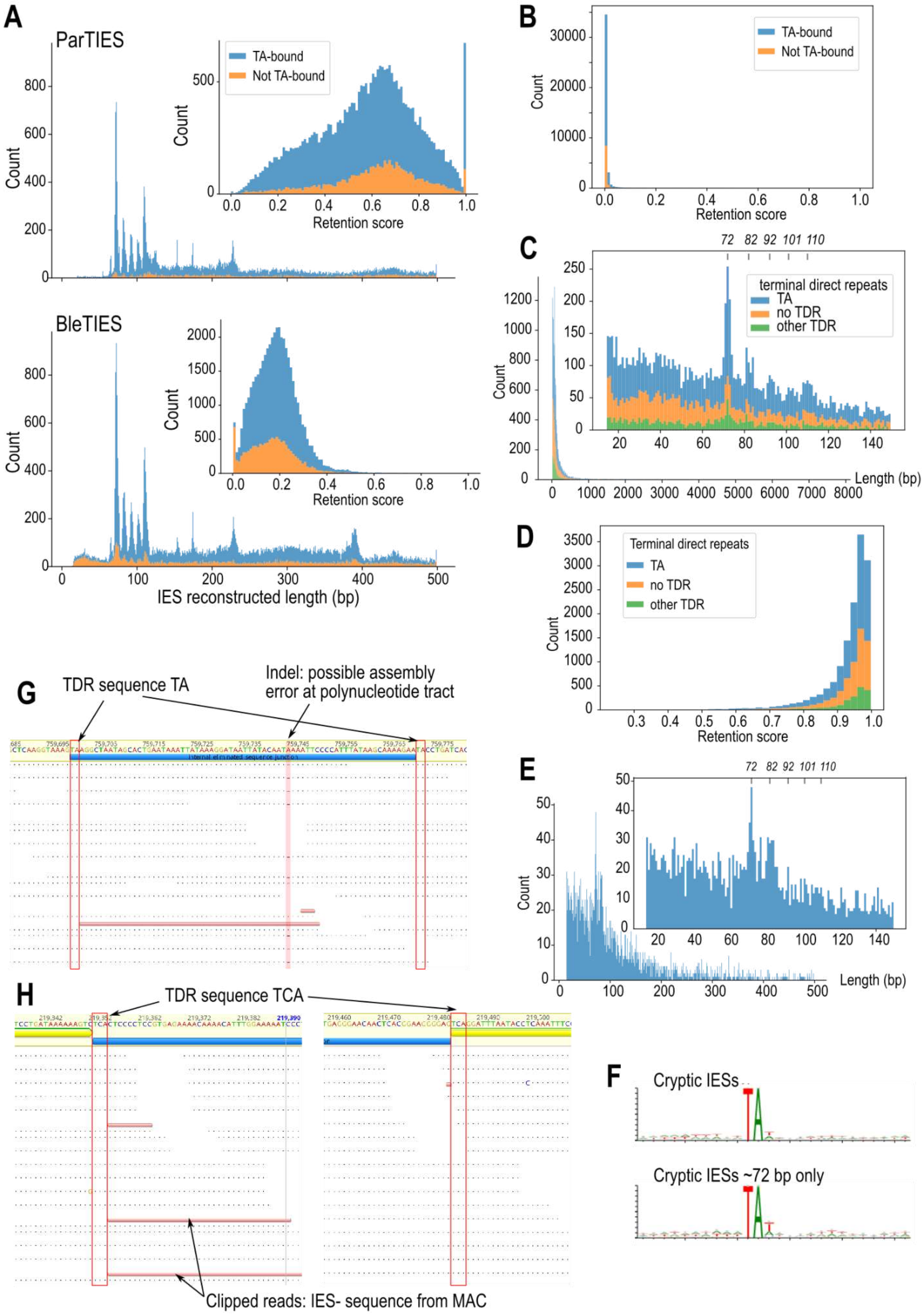

A
ParTIES
Count
800
600
400
200
0
TA-bound
Not TA-bound
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Retention score
BleTIES
2000
1500
1000
500
0
100
200
300
400
500
IES reconstructed length (bp)
B
30000
20000
10000
C
72 82 92 101 110
terminal direct repeats
TA
no TDR
other TDR
1200
1000
800
600
400
250
150
50
20
40
60
80
100
120
140
2000
3000
4000
5000
6000
7000
8000
Length (bp)
D
Terminal direct repeats
3500
3000
2500
1500
0.3
0.5
0.7
0.9
E
30
10
F
Cryptic IESs
Cryptic IESs ~72 bp only
G
TDR sequence TA
Indel: possible assembly error at polynucleotide tract
H
TDR sequence TCA
Clipped reads: IES- sequence from MAC

Figure S2.

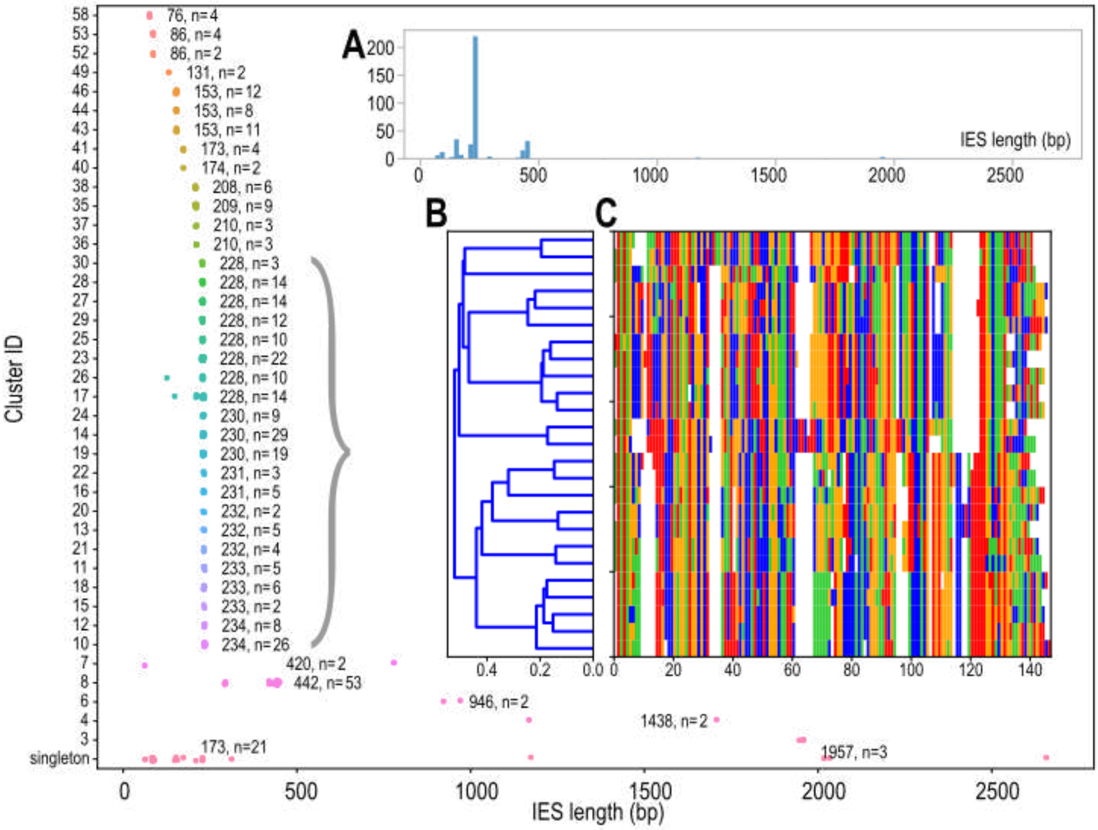

A
200
150
100
50
0
IES length (bp)
0
500
1000
1500
2000
2500
B
0.4
0.2
0.0
C
0
20
40
60
80
100
120
140
Cluster ID
58
53
52
49
46
44
43
41
40
38
35
37
36
30
28
27
29
25
23
26
17
24
14
19
22
16
20
13
21
11
18
15
12
10
7
8
6
4
3
singleton
76, n=4
86, n=4
86, n=2
131, n=2
153, n=12
153, n=8
153, n=11
173, n=4
174, n=2
208, n=6
209, n=9
210, n=3
210, n=3
228, n=3
228, n=14
228, n=14
228, n=12
228, n=10
228, n=22
228, n=10
228, n=14
230, n=9
230, n=29
230, n=19
231, n=3
231, n=5
232, n=2
232, n=5
232, n=4
233, n=5
233, n=6
233, n=2
234, n=8
234, n=26
420, n=2
442, n=53
946, n=2
1438, n=2
173, n=21
1957, n=3
0
500
1000
1500
2000
2500
IES length (bp)

Figure S3.

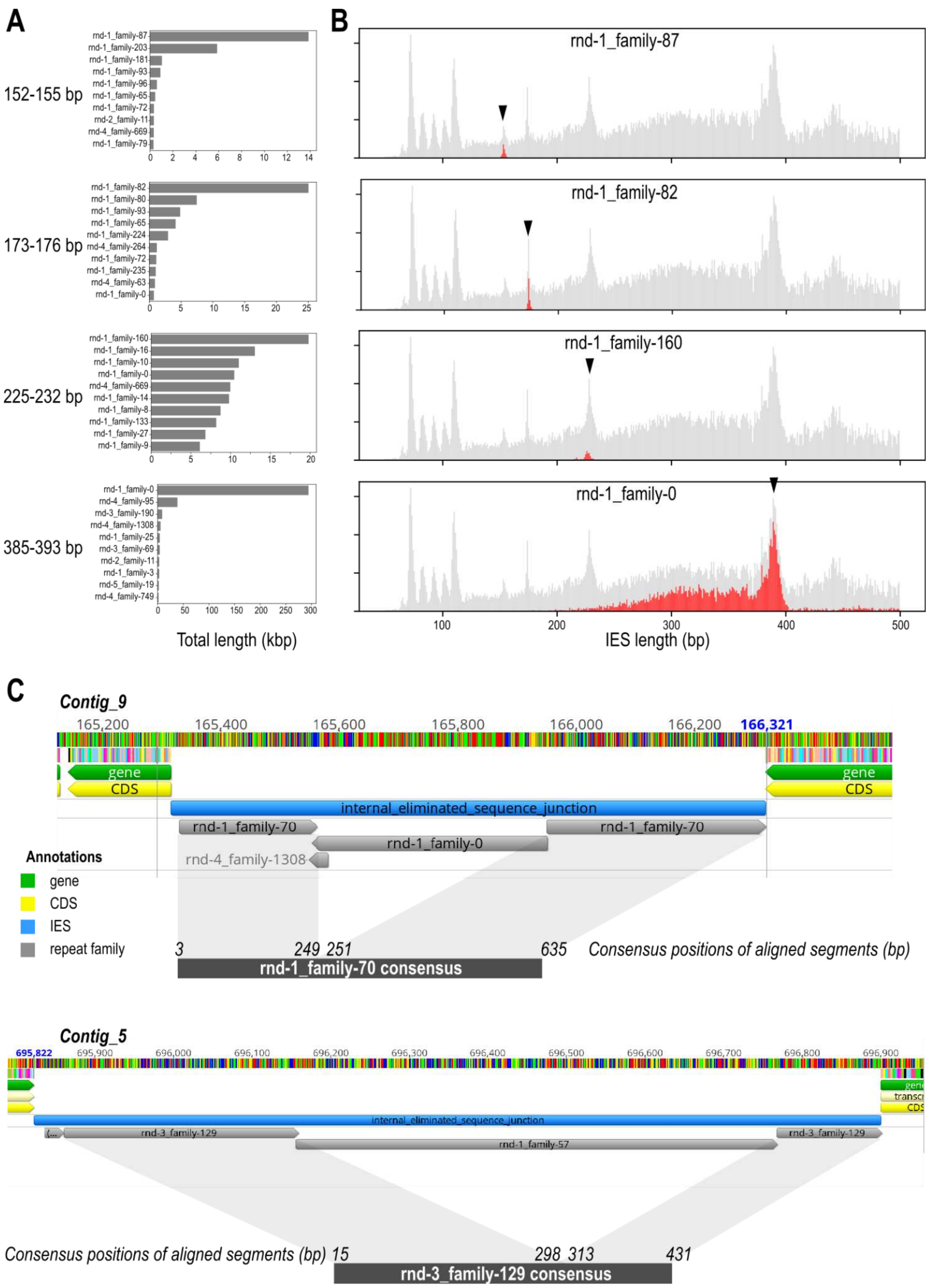

A
152-155 bp
rnd-1_family-87
rnd-1_family-203
rnd-1_family-181
rnd-1_family-93
rnd-1_family-96
rnd-1_family-65
rnd-1_family-72
rnd-2_family-11
rnd-4_family-669
rnd-1_family-79
0 2 4 6 8 10 12 14
173-176 bp
rnd-1_family-82
rnd-1_family-80
rnd-1_family-93
rnd-1_family-65
rnd-1_family-224
rnd-4_family-264
rnd-1_family-72
rnd-1_family-235
rnd-4_family-63
rnd-1_family-0
0 5 10 15 20 25
225-232 bp
rnd-1_family-160
rnd-1_family-16
rnd-1_family-10
rnd-1_family-0
rnd-4_family-669
rnd-1_family-14
rnd-1_family-8
rnd-1_family-133
rnd-1_family-27
rnd-1_family-9
0 5 10 15 20
385-393 bp
rnd-1_family-0
rnd-4_family-95
rnd-3_family-190
rnd-4_family-1308
rnd-1_family-25
rnd-3_family-69
rnd-2_family-11
rnd-1_family-3
rnd-5_family-19
rnd-4_family-749
0 50 100 150 200 250 300
Total length (kbp)
B
rnd-1_family-87
rnd-1_family-82
rnd-1_family-160
rnd-1_family-0
100 200 300 400 500
IES length (bp)
C
Contig_9
165,200 165,400 165,600 165,800 166,000 166,200 166,321
gene
CDS
gene
CDS
internal_eliminated_sequence_junction
rnd-1_family-70
rnd-1_family-70
rnd-1_family-0
rnd-4_family-1308
Annotations
gene
CDS
IES
repeat family
3 249 251 635 Consensus positions of aligned segments (bp)
rnd-1_family-70 consensus
Contig_5
695,822 695,900 696,000 696,100 696,200 696,300 696,400 696,500 696,600 696,700 696,800 696,900
gene
transcr
CDS
internal_eliminated_sequence_junction
(...
rnd-3_family-129
rnd-3_family-129
rnd-1_family-57
Consensus positions of aligned segments (bp) 15 298 313 431
rnd-3_family-129 consensus

Figure S4.

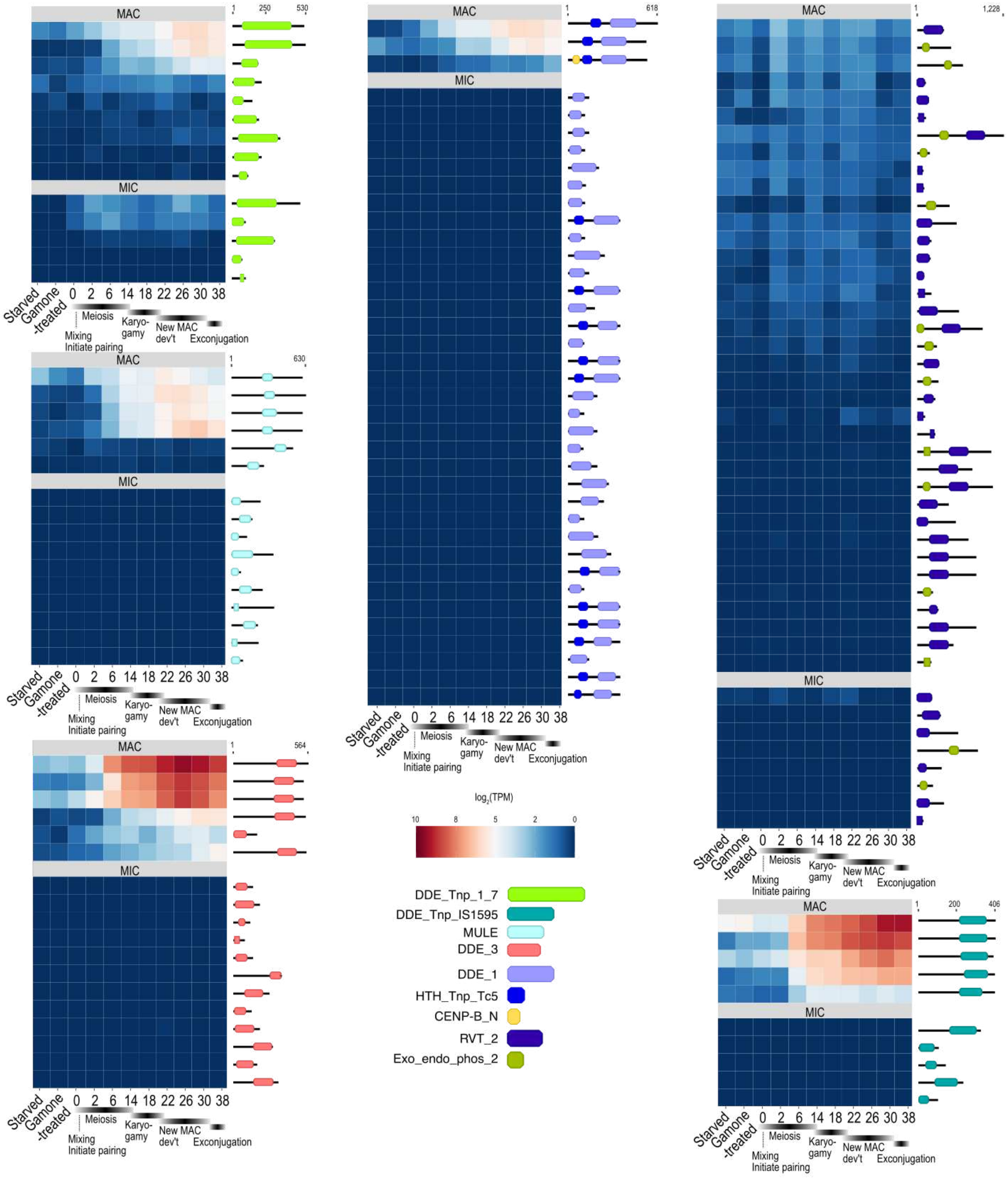

MAC
MIC
1
250
530
630
564
618
1,228
200
406
Starved
Gamone
-treated
0 2 6 14 18 22 26 30 38
Meiosis
Karyo-
gamy
New MAC
dev't
Exconjugation
Mixing
Initiate pairing
log2(TPM)
10
8
5
2
0
DDE_Tnp_1_7
DDE_Tnp_IS1595
MULE
DDE_3
DDE_1
HTH_Tnp_Tc5
CENP-B_N
RVT_2
Exo_endo_phos_2

Figure S5.

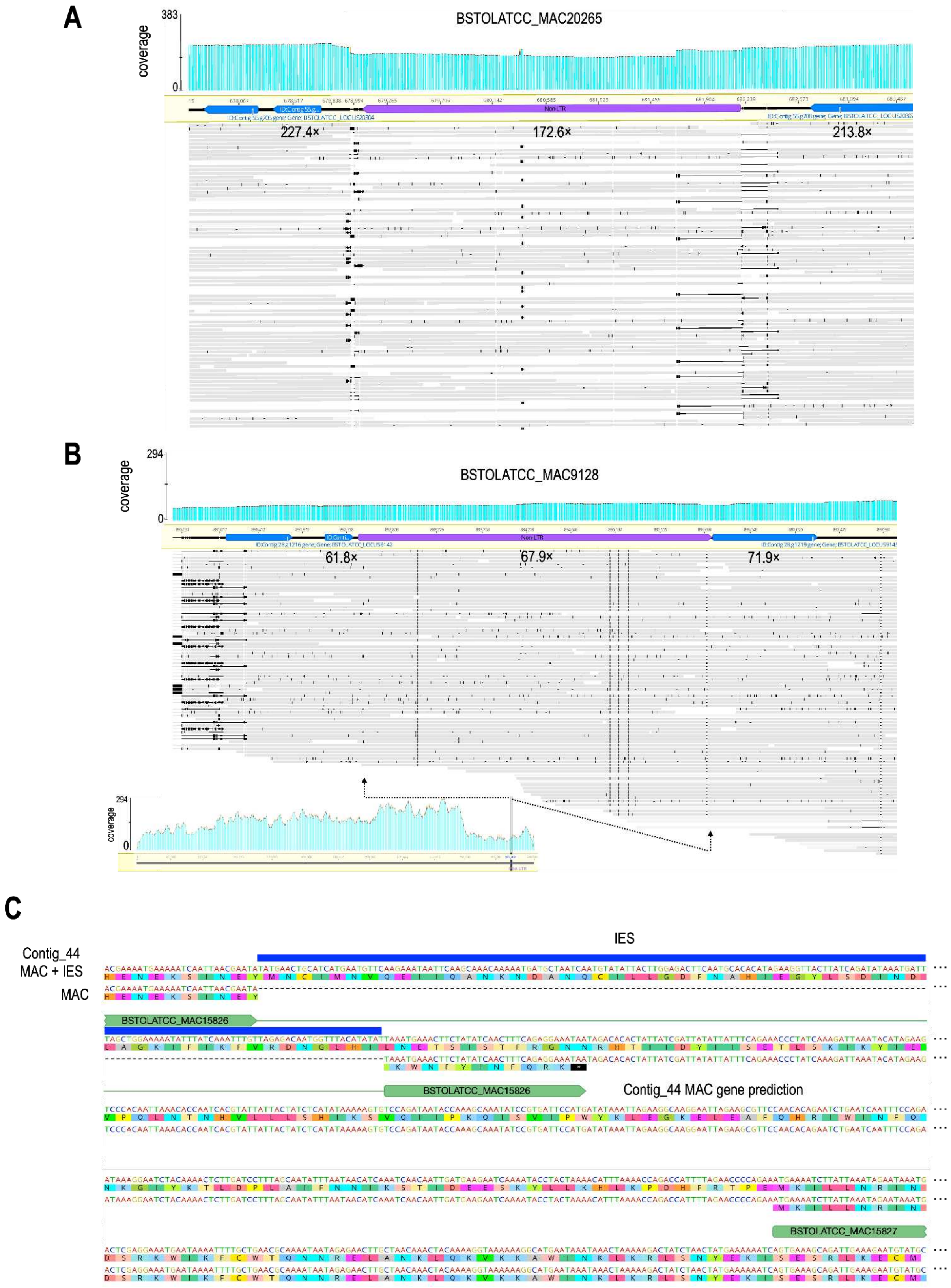
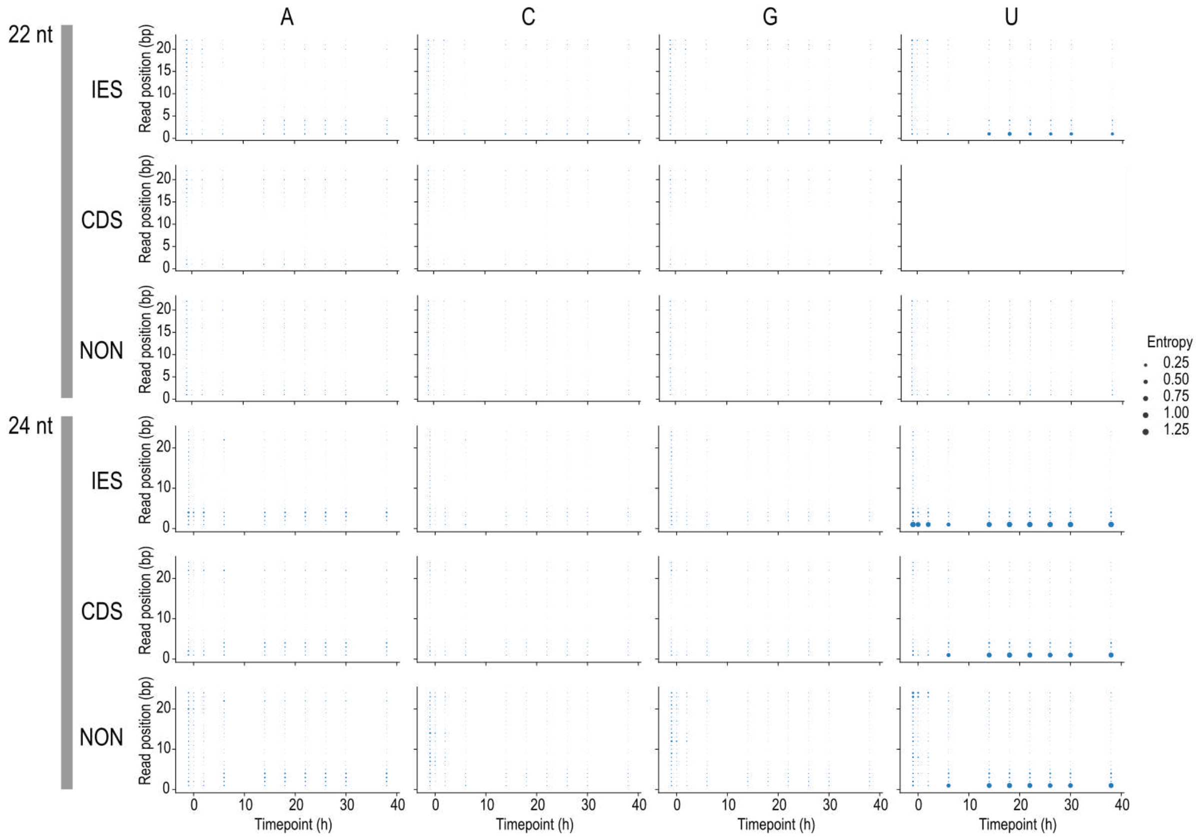

A
383
coverage
0
BSTOLATCC_MAC20265
Non-LTR
227.4×
172.6×
213.8×
B
294
coverage
0
BSTOLATCC_MAC9128
Non-LTR
61.8×
67.9×
71.9×
294
coverage
0
C
IES
Contig_44
MAC + IES
MAC
BSTOLATCC_MAC15826
BSTOLATCC_MAC15826
Contig_44 MAC gene prediction
BSTOLATCC_MAC15827

Figure S6.


A
C
G
U
22 nt
24 nt
IES
CDS
NON
Read position (bp)
0
5
10
15
20
Timepoint (h)
0
10
20
30
40
Entropy
0.25
0.50
0.75
1.00
1.25