Full title:

The *Blepharisma stoltei* macronuclear genome: towards the origins of whole genome reorganization

Short title:

Macronuclear genome of *Blepharisma stoltei* strain ATCC 30299

Minakshi Singh¹, Kwee Boon Brandon Seah¹, Christiane Emmerich¹, Aditi Singh¹, Christian Woehle², Bruno Huettel², Adam Byerly³, Naomi Alexandra Stover⁴, Mayumi Sugiura⁵, Terue Harumoto⁵, Estienne Carl Swart^{1,*}

1. Max Planck Institute for Biology, Tuebingen, Germany

2. Max Planck Genome Center Cologne, Max Planck Institute for Plant Breeding, 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. Nara Women's University, Nara, Japan

* Corresponding author

Abstract

The germ-soma distinction is a defining feature of multicellular eukaryotes. Analogous to this, ciliates, a ubiquitous microbial eukaryote lineage, have morphologically and functionally distinct nuclei, but within single cells: the germline micronucleus (MIC) and somatic macronucleus (MAC). The origins and mechanisms of the MIC to MAC transformation, especially the extensive elimination of abundant internally eliminated sequences (IESs) and transposons during genome reorganization, are great biological mysteries. *Blepharisma* represents one of the two earliest diverging ciliate classes, and has unique, dual pathways of MAC development, making it ideal for investigating the functioning, origins and evolution of these processes. Here, we report the MAC genome assembly of *Blepharisma stoltei* strain ATCC 30299 (41 Mb), arranged as numerous alternative telomere-capped minichromosomes, tens to hundreds of kilobases long. The *B. stoltei* MAC genome encodes eight PiggyBac transposase homologs liberated from transposons. All are subject to purifying selection, but just one, the putative *Blepharisma* IES excisase, has a complete catalytic amino acid triad. Numerous genes encoding other domesticated transposases are present in *B. stoltei*, and often are comparably strongly upregulated in a similar timeframe to model ciliate genome reorganization homologs. Our phylogenetic investigations suggest the PiggyBac homologs may have been ancestral ciliate IES excisases. The *B. stoltei* MAC genome, together with the upcoming MIC genome, highlights the evolution and complex interplay between transposons, domesticated transposases, and genome reorganization in the context of germline-soma differentiation within single cells.

41 Abbreviations

- 42 • MIC - micronucleus
- 43 • MAC - macronucleus
- 44 • IES - interspersed eliminated sequence
- 45 • MDS - macronuclear-destined sequence
- 46 • PacBio - Pacific Biosciences
- 47 • CLR - continuous long read (PacBio)
- 48 • CCS - circular consensus sequence (PacBio)
- 49 • HiFi - High-fidelity read (PacBio)
- 50 • ATAS - alternative telomere addition site
- 51 • PBLE - PiggyBac-like element
- 52 • PGBD - PiggyBac element-derived
- 53 • Pgm - PiggyMac
- 54 • PgmL - PiggyMac-like

Introduction

Acquisition of new organelles defines and is responsible for the origins of eukaryotes [1]. Ciliates represent one of a couple of eukaryotic lineages with more than one kind of nucleus within individual cells [2]. Profound differences between the ciliate nuclei and the genomes they contain now exist, with some parallels to the germline-soma distinction in multicellular organisms [3]. Ciliate nuclear and genomic dimorphism provides a special opportunity to observe the consequences of evolution of a specialized organelle clearly derived from a general one and for studying the evolution of new, functionally differentiated genes that enable this unusual situation.

The processes responsible for large-scale developmental DNA elimination [4] in ciliate nuclei are under active study, but what the responsible ancestral molecules were remains to be determined. Knowledge of the molecules responsible for ciliate genome reorganization is dominated by *Tetrahymena* and *Paramecium* (class Oligohymenophorea), with additional input from *Oxytricha*, *Stylonychia* and *Euplotes* (class Spirotrichea) [5,6], whereas this remains to be investigated in the remaining nine or so classes. To gain fresh insights into genome reorganization and tackle questions about its origin we focused on the ciliate species *Blepharisma stoltei*. Together with its sister-class, Karyorelictea, the class Heterotrichea, to which this *B. stoltei* belongs, represent the earliest branching ciliate lineages, more distantly related to current model ciliates than those models are to each other [7]. Furthermore, *Blepharisma* exhibits distinctive alternative somatic developmental pathways, which have potential utility in disentangling processes involved in genome reorganization from preceding ones that may indirectly influence it.

Blepharisma is a distinctive genus of single-celled ciliates (Figure 1) known for the red, light-sensitive pigment, blepharismine, in the sub-pellicular membranes of representative species [8],

and unusual nuclear/developmental biology (Figure 2) [9]. To date molecular investigations and genomics of ciliates have predominantly focused on oligohymenophoreans and spirotrichs (Figure 3, Table S1). In recent years, publication of a draft genome for the heterotrich ciliate, *Stentor*, has facilitated revival of this genus for investigations of cellular regeneration [10–12]. However, significant hurdles still need to be overcome to investigate genome reorganization in *Stentor coeruleus* since requisite cell mating has not been observed in the reference somatic genome strain (personal communication, Mark Slabodnick), and very high lethality has been reported for other strains in which mating occurred [13]. We therefore focused on *Blepharisma* which is amenable to such investigations, with controlled induction of mating, and, critically, procedures for investigating cellular and nuclear development have been established from more than a century of meticulous cytological research [8,14–25].

The *Blepharisma stoltei* strains used in the present study were originally isolated in Germany (strain ATCC 30299) and Japan (strain HT-IV), with the former continuously cultured for over fifty years, and the latter for over a decade. The cells are comparatively straightforward to maintain, e.g., stable cultures can be established in a simple salt medium on a few grains of rice passaged every few months. Due to their distinctive pigmentation and large size several *Blepharisma* species are excellent subjects for introducing cell biology concepts to non-specialists, and are thus readily available for educational purposes from commercial suppliers like Carolina Biological Supply Company (USA). They are ideal subjects for behavioural and developmental investigations, e.g., as voracious predators of smaller ciliates and other unicellular species, and also exhibit pronounced phenotypic plasticity, including forming cysts and giant, cannibal cells under suitable conditions [8].

Like all ciliates [3], *Blepharisma* cells have two types of nuclei: a macronucleus (MAC) which is very large and transcriptionally active during vegetative growth, and a small, generally

transcriptionally inactive micronucleus (MIC), which serves as the germline. Each *Blepharisma* cell has one MAC and several MICs (Figure 1A, B). In vegetative propagation (asexual replication) of *Blepharisma*, cell fission results in about half of the MAC pinching off which is then distributed to each of the resulting daughter cells together with the mitotically divided MICs. Upon starvation, *Blepharisma* cells, like other ciliates, are also capable of conjugation and the associated sexual processes. Essential for developmental investigations, the intricate ballet of nuclear movements and morphological changes occurring during *Blepharisma* conjugation is well-documented [9] (Figure 2). During this process about half of the MICs in each of the cells undergo meiosis (meiotic MICs) and the rest do not (somatic MICs) (Figure 2B). One of the meiotic MICs eventually gives rise to two haploid gametic nuclei. One gametic MIC (the migratory nucleus) from each conjugating cell is exchanged with that of its partner. In parallel in partnered cells, subsequent fusion of the migratory and stationary haploid nuclei generates a zygotic nucleus (synkaryon), and after successive mitotic divisions gives rise to both new MICs and new MACs (known as anlagen). The new MACs continue to mature, eventually growing in size and DNA content [9].

Conveniently for investigations of development and genome reorganization, *Blepharisma* is one of only two ciliate genera, along with *Euplotes* [26–29], where conjugation has been shown to be mediated through pheromone-like substances called gamones. *Blepharisma* has two mating types, distinguished by their gamone production. Mating type I cells release gamone 1, a ~30 kDa tyrosine-rich [30] glycoprotein [31,32]; mating type II cells release gamone 2, calcium-3-(2'-formylamino-5'-hydroxybenzoyl) lactate, a small-molecule effector [33]. *Blepharisma* cells commit to conjugation when complementary mating types recognize each other's gamones, with the cells remaining paired while meiosis and fertilization occur and eventually new MACs begin to form.

As in model ciliates, during *Blepharisma* anlagen development, MIC-specific sequences are removed from a MIC genome copy to form a functional MAC genome (Seah, et al. in prep.). The new MAC genomes of model ciliate species are largely free of mobile elements and other forms of “junk” DNA contained in the MIC genome [34]. In the best studied ciliates, a form of genome reorganization assisted by small RNAs (sRNAs) is thought to occur [5]. Specific genome segments limited to the MIC, termed internally eliminated sequences (IESs), are excised by domesticated transposases, and may have originated from transposons [3,5,35,36]. Large scale genome-wide DNA amplification accompanies the genome reorganization process, producing thousands of copies in mature MACs of larger ciliate species [3,34].

We were motivated to investigate genome reorganization in *Blepharisma*, as, unlike model ciliates, these cells can produce two kinds of anlagen, and particularly because one of the two pathways of their development skips the complex series of mitoses, meioses, nuclear exchanges and fertilization [9] (Figure 2). Primary anlagen mature in the conventional manner from zygotic nuclei. Somatic MICs which have not undergone meiosis can give rise to secondary anlagen, which can develop into mature macronuclei [9]. This occurs frequently in strains with a high frequency of selfing (conjugation among cells within a clonal population), in preference to development of primary macronuclear anlagen [9]. This alternative pathway of MAC development has also been observed experimentally after removal of primary macronuclear anlagen by microsurgery [9]. As conjugation progresses, the old (maternal) MACs are progressively degraded [9]. Since the *B. stoltei* MIC genome has numerous IESs which interrupt genes (Seah et al. in prep.), in principle, to produce functional, mature MAC genomes, it is essential that reorganization of DNA occurs in both primary and secondary anlagen.

Here we provide essential macronuclear genome and developmental, transcriptomic resources for *B. stoltei* and present the first investigations of possible molecules involved in its genome

reorganization. Like *Stentor coeruleus* [10], *Blepharisma stoltei* genes have the shortest known spliceosomal introns, predominantly 15 or 16 nt long. From long-read sequencing the *B. stoltei* MAC genome appears to be organized in the form of numerous minichromosomes. Among the MAC-encoded transposase genes we identified in *Blepharisma* were PiggyBac transposase homologs, which, as far as we are aware, have not been reported in any ciliates other than *Paramecium* and *Tetrahymena*. A few *Blepharisma* PiggyBac homologs are substantially upregulated in MAC development, including one with a complete catalytic triad which is the main candidate IES excisase. Raising the possibility of additional IES excisases, transposases from a few different classes are also present in the MAC genome, including some with complete catalytic triads. Consistent with ancient origins of genome reorganization in ciliates, *Blepharisma* shares pronounced development-specific upregulation of homologs known to be involved in this process, notably Dicer-like and Piwi proteins responsible for sRNA biogenesis. *Blepharisma* therefore represents an invaluable outgroup for investigations of the evolution of genome reorganization.

Results

A compact, extensively fragmented genome

The draft *Blepharisma stoltei* ATCC 30299 MAC genome is compact, at 41 Mb, and AT rich (66%), like most sequenced ciliate MAC genomes, and relatively complete (Figure 3; Table S1, 2, Figure S1). From joint variant calling of reads from strains ATCC 30299 and HT-IV mapped to this assembly, strain ATCC 30299 appears to be virtually homozygous, with only 1277 heterozygous single-nucleotide polymorphisms (SNPs) observed compared to 193725 in strain HT-IV (i.e., individual heterozygosity of 3.08×10^{-5} vs. 4.67×10^{-3} respectively). Low SNP levels

were likely beneficial for the overall assembly contiguity, since heterozygosity poses significant algorithmic challenges for genome assembly software [37]. For brevity's sake, we refer to this draft genome as the *Blepharisma* MAC genome (and likewise "*Blepharisma*" for the associated strain). Though the final assembly comprises 64 telomere-to-telomere sequences, it is not possible to define conventional chromosome ends given the extensive natural fragmentation of the *Blepharisma* MAC genome (characterized in the next section), hence we refer to them simply as "contigs".

The basic telomere unit of *B. stoltei* is a permutation of CCCTAACA, like its heterotrich relative *Stentor coeruleus* [10] (Figure S2). However, a compelling candidate for a telomerase ncRNA (TERC) could not be found in either *Blepharisma* or *Stentor* using Infernal [38] and RFAM models (RF00025 - ciliate TERC; RF00024 - vertebrate TERC). Thus, it was not possible to delimit the ends of the repeat. Heterotrichs may use a different kind of ncRNA, or one which is very divergent from known TERCs. In contrast to the extremely short (20 bp) macronuclear telomeres of spirotrichs like *Oxytricha* with extreme MAC genome fragmentation [39], sequenced *Blepharisma* macronuclear telomeres are moderately long (Figure S2A), with a mode of 209 bp (i.e., 26 repeats of the 8 bp motif), extending to a few kilobases.

Approximately one in eight reads in the *Blepharisma* HiFi library were telomere-bearing and distributed across the entire genome (Figure 4A), using a moderately strict definition of possessing at least three consecutive telomeric repeats. In contrast, the telomere-bearing reads of the model ciliate *Tetrahymena thermophila* predominantly map to chromosome ends (Figure S3A), and only one in fifty nine *Tetrahymena* CLR reads are telomere-bearing, as identified by three consecutive telomeric subunit repeats (i.e., 3×CCCCAA).

Typically, only a minority of mapped reads are telomere-bearing at individual internal contig positions, and so we term them alternative telomere addition sites (ATASs) (Figure 4A). We identified a total of 46705 potential ATASs in *Blepharisma* by mapping PacBio HiFi reads onto the MAC reference, and searching for partially mapped reads that contained telomeric repeats in the unmapped (clipped) segments without any intervening sequence between the mapped segment and the beginning/end of the telomeric repeat. The majority of these sites (38686) were represented by only one mapped HiFi read.

The expected distance between telomeres, and hence the average MAC DNA molecule length, is about 130 kb. This is consistent with the raw input MAC DNA lengths, which were mostly longer than 10 kb and as long as 1.5 Mb (Figure S3C, D), and our observation that a small fraction (1.3%) of *Blepharisma*'s HiFi reads are bound by telomeres on both ends. Excluding the length of the telomeres, these telomere-bound reads may be as short as 4 kb (Figure S2B). Given the frequency of telomere-bearing reads, we expect many additional two-telomere DNA molecules longer than 12 kb, the approximate maximum length of the HiFi reads (Figure S3C, D).

Since the lengths of the sequenced two-telomere DNA molecules on average imply that they encode multiple genes, we propose classifying them as “minichromosomes”. This places them between the length of “nanochromosomes” of ciliates like *Oxytricha* and *Stylonychia*, which typically encode single genes and are on the order of a few kilobases long [39,40], and MAC chromosomes of *Paramecium tetraurelia* and *Tetrahymena thermophila* which are hundreds of kilobases to megabases long [41–43]. Recently it was reported that the *Paramecium bursaria* MAC genome is considerably more fragmented than those of other previously examined *Paramecium* species [44]. The DNA molecules from *P. bursaria* have thus also been classified as minichromosomes [44].

Alternative telomere addition sites in the MAC genome tend to be intergenic in model ciliates like *Oxytricha trifallax* [39]. In *Blepharisma*, we found more intergenic ATASs (28309) than intragenic ones (18396). As intergenic regions only make up 10.1 Mb of the assembly, the intergenic frequency of ATASs is about five-fold higher (2.81 per 1 kb) than intragenic frequency (0.562 per 1 kb). The presence of intragenic ATASs raises the question how the cell tolerates or deals with mRNAs encoding partial proteins transcribed from 3' truncated genes. Since the sequence data was from a clonal population, it is not possible to tell how much ATAS variability there is within individual cells. However, it is conceivable that their positional variation in single cells reflects that of the population. In this case, together with redundancy from massive DNA amplification there would likely be sufficient intact copies of every gene.

Beyond the first 2-5 bp corresponding to the junction sequences, the average base composition on the chromosome flanking ATAS junctions shows an asymmetrical bias (Figure 4B). From position +6 onwards there is an enrichment of T to about 40% and A to 35-39%, compared to the genome-wide frequencies of 33% each. At position +19 to +23, there is a slight decrease in T to 37-39%. AT values gradually decline back to about 35% each by position +150. Correspondingly, G and C are depleted downstream of ATAS junctions, dropping to a minimum of 8.6% and 11% respectively around position +37, compared to the genome-wide average of 17% each. AT enrichment and GC depletion upstream of ATAS junctions are less pronounced.

If breakage and chromosome healing were random, we would not expect such an asymmetry. This suggests that there is a nucleotide bias, whether in the initiation of breaks, telomere addition, or in the processing of breaks before telomere addition. However, we have not yet identified any conserved motif like the 15 bp chromosome breakage site (CBS) in *Tetrahymena* [45] nor a short 10-bp sequence periodicity in base composition like in *Oxytricha trifallax* [46].

Therefore, telomere addition in *B. stoltei* appears to involve base-pairing of short segments of about 2 bp between the telomere and chromosome, with a bias centered on the “CT” in the telomere unit, and an asymmetrical preference for AT-rich sequences on the chromosomal side of the junction.

Tiny spliceosomal introns

The *Blepharisma stoltei* ATCC 30299 macronuclear genome is gene-dense (25,711 predicted genes), with short intergenic regions, tiny introns and untranslated regions (UTRs) (Figure 5A). In contrast to *Stentor* which, unusually for ciliates, uses the standard genetic code, *B. stoltei* uses an alternative genetic code with UGA codons reassigned from stops to tryptophan (Figure S5).

Like *Stentor* [10], most (82%) *Blepharisma* genes have no introns. In line with genome compactness, during our inspections we also observed numerous overlapping poly(A)-tailed RNA-seq reads on opposite strands derived from convergently transcribed gene pairs. The correlation of the lengths of different noncoding region classes (intergenic regions, introns and UTRs) can be explained by them being subject to common, neutral evolutionary processes [47].

Blepharisma introns are mostly (97%) 15 or 16 nucleotides (nt) long, like those of *Stentor* (Figure 5B). Though intron reduction (7389 introns predicted in the reference *B. stoltei* MAC genome, i.e., 0.29 introns per gene) is not as extreme as some other microbial eukaryotes, like *Giardia lamblia* [48], where almost all have been lost, both *Blepharisma* and *Stentor* have much fewer introns relative to other ciliates (e.g., intron densities of 1.6, 2.3 and 4.8 introns per gene in *Paramecium*, *Oxytricha* and *Tetrahymena*, respectively [49]) and to the putative, relatively intron-rich eukaryotic common ancestor [50], along with their extreme length reduction.

282

283 *Blepharisma* 15 nt introns possess a characteristic branch-point “A”, as would be expected in

284 classical models of lariat formation during mRNA splicing (Figure 5C). 16 nt introns almost

285 invariably have an “A” at either 10 or 11 nt downstream of the donor site (i.e., only one of 499

286 does not, but has “A” at 9 nt), although this is not obvious in the consensus sequence logo

287 because the position is variable (Figure S6D). Similarly, 17 nt introns all possess “A” at 10-12 nt

288 downstream of the donor site. Only a few intron bases, 5-8 and 12, of *Blepharisma*’s 15 nt

289 introns are relatively unconstrained (Figure 5C). This leaves little room for the presence of any

290 additional regulatory elements in the mRNA or underlying DNA.

291 Extensive duplications of transmembrane protein genes

292 A notable extended ~220 kb region encoding 53 genes belonging to a single orthologous group

293 (orthogroup), OG0000085 is present on Contig_1 (Figure 5A). Four additional OG0000085

294 genes are present at the opposite end of Contig_1, and 24 copies are found on other contigs,

295 often clustered together (Figure 5D). The DNA coverage across this region is lower (74×) than

296 the rest of Contig_1 (185×). Though there is uncertainty in the exact extent, given the sheer

297 volume of reads involved, the assembled sequences certainly correspond to highly repetitive

298 regions of the MAC genome. At the junction between the lower and higher coverage regions

299 more than 30 HiFi reads link the two regions of coverage, and a similar number of telomere-

300 bearing reads are in close proximity. At the junction we also observe at least two potential

301 locations of IESs, corresponding to regions that may be partially IES/partially MDS.

302

303 Large clusters of genes from particular orthogroups can be found on additional contigs (Figure

304 5D). In total 551 (2%) of predicted *B. stoltei* genes belong to the orthogroups with the largest

305 clusters per contig. Some of the largest contiguous clusters of genes from these orthogroups are

situated at the ends of contigs, suggesting they may have caused assembly breaks beyond them. One contig, split off from other connected components in the assembly graph, predominantly encodes genes from a single orthogroup (contig_64, 43× coverage; Figure 5D; Figure S7). Further increases in read length and accuracy may allow assemblers to fully resolve these in future. Curiously, all the orthogroups corresponding to the largest contiguous clusters of genes appear to be transmembrane proteins, or decayed remnants thereof. The nature of these proteins is described in Supplemental Text ("Properties of proteins encoded by extensive duplications").

Features of gene expression during new MAC development

To gain an overview of the molecular processes during genome reorganization in *Blepharisma* we examined gene expression trends across a developmental time series. Complementary mating strains of *B. stoltei* were treated with gamones of the opposite mating type, before mixing to initiate conjugation [9,51]. Samples for morphological staging and RNA-seq were taken at intervals from the time of mixing ("0 hour" time point) up to 38 hours (Figure S8). The progression of nuclear morphological development and the proportion of cells in each stage at various time points is illustrated in Figure S9.

During conjugation in *Blepharisma*, meiosis begins around 2 h after conjugating cell pairs form and continues up to around 18 h, by when gametic nuclei generated by meiosis have been exchanged (Figure 2A, B, Figure S9). This is followed by karyogamy and mitotic multiplication of the zygotic nucleus (22 hours). At the 26 h time point, new, developing primary MACs can be observed in the conjugating pairs, in the form of large, irregular bodies (Figure 2A, Figure S9). These macronuclear anlagen mature into the new MACs of the exconjugant cell by 38 h, after which cell division generates two daughter cells. Smaller secondary MACs, derived directly from

MICs without all the intermediate nuclear stages, can also be seen from 22 hours to 30 hours, but eventually disappear, giving sole way to the primary MACs (Figure 2A, Figure S9).

Examining gene expression at 26 hours, when the majority of cells are forming a new MAC (Figure S9), we observe two broad trends: relatively stable constitutive gene expression (Table S5; Data S3), e.g., an actin homolog (ENA accession: BSTOLATCC_MAC19444) and a bacteria-like globin protein (BSTOLATCC_MAC21846), versus pronounced development-specific upregulation (Table S6; Data S3), e.g., a histone (BSTOLATCC_MAC21995) an HMG box protein (BSTOLATCC_MAC14030), and a translation initiation factor (eIF4E, BSTOLATCC_MAC5291).

We eschewed a crude, large-scale Gene Ontology (GO) enrichment analysis in favour of close scrutiny of a smaller subset of genes strongly upregulated during new MAC formation which we expect are more relevant to molecular biologists. For this, computational gene annotations in combination with BLASTP searches and examination of literature associated with homologs was used. Ranking the relative gene expression at 26 hours vs. the average expression of starved, gamone treated, and 0 hour cells, in descending order, reveals numerous genes of interest in the context of genome reorganization, including homologs of proteins known to be involved in genome reorganization in model ciliate species (Table S6).

Among the top 100 genes ranked this way (69× to 825× upregulation) nine contain PFAM transposase domains: DDE_Tnp_1_7, DDE_3, MULE and DDE_Tnp_IS1595 (e.g., BSTOLATCC_MAC2188, BSTOLATCC_MAC14490, BSTOLATCC_MAC18054, BSTOLATCC_MAC18052, respectively). We also observe small RNA (sRNA) biogenesis and transport proteins, i.e., a Piwi protein (BSTOLATCC_MAC5406) and a Dicer-like protein (BSTOLATCC_MAC1138; see also “Supplemental text”, “Homologs of small RNA-related

proteins involved in ciliate genome reorganization” and Figure S10), and a POT1 telomere-binding protein homolog (POT1.4; BSTOLATCC_MAC1496; Supplemental Text “Telomere-binding protein paralogs”). Numerous homologs of genes involved in DNA repair and chromatin are also present among these highly developmentally upregulated genes (“Supplemental text”, “Development-specific upregulation of proteins associated with DNA repair and chromatin” and “Supplemental text”, “Development-specific histone variant upregulation”). The presence of proteins involved in either transcription initiation or translation initiation among these highly upregulated genes suggests a possible manner to coordinate the regulation of development-specific gene expression (“Supplemental text”, “Development-specific upregulation of proteins associated with initiation of transcription and translation”).

PiggyBac homologs in *Blepharisma* are candidate IES excisases

In *Paramecium tetraurelia* and *Tetrahymena thermophila*, PiggyBac transposases are responsible for IES excision during genome reorganization occurring in the developing new MAC [52,53]. These transposases appear to have been domesticated, i.e., their genes are no longer contained in transposons but are encoded in the somatic genome where they provide a necessary function in genome development [52,53]. PiggyBac belongs to the DDE/D-superfamily of transposases due to the presence of a protein domain containing the catalytic triad of 3 aspartic acid residues (DDD), where the third residue can also be a glutamate (E). PiggyBac homologs typically have a DDD catalytic triad, rather than the more common DDE triad of other DDE/D transposases [54]. The DDD catalytic motif is present in the PiggyMac (Pgm) of *Paramecium* and the *Tetrahymena* PiggyBac homologs Tpb1 and Tpb2 [53,55]. Among ciliates, domesticated PiggyBac transposases have so far only been reported in the model oligohymenophorean genera *Paramecium* and *Tetrahymena*, and have not been detected in either the MAC or MIC genomes of the spirotrich *Oxytricha trifallax* [39,56].

We detected more transposase domains (9 distinct PFAM identifiers) in *B. stoltei* than any of the other ciliate species we examined (Figure 6A). Seven of these domains are also found in at least one other ciliate species. Using HMMER searches with the PFAM domain characteristic of PiggyBac homologs, DDE_Tnp_1_7 (PF13843), we found eight homologs in *B. stoltei* ATCC, none of which were flanked by terminal repeats identified by RepeatModeler. We also found PiggyBac homologs in *B. stoltei* HT-IV and *B. japonicum* R1072. Reminiscent of *Paramecium tetraurelia*, which, among ten PiggyMac homologs, has just one homolog with a complete catalytic triad [55], the DDD triad is preserved in just a single *B. stoltei* ATCC 30299 PiggyBac homolog (Figure 6B; Contig_49.g1063, BSTOLATCC_MAC17466). This gene is strongly upregulated during development from 22 to 38 hours after formation of heterotypic pairs, when new MACs develop and IES excision is required (Figure 6B). Furthermore, there are significant similarities in the basic properties of *Blepharisma* and *Paramecium* IESs, which will be reported in detail in the subsequent *B. stoltei* MIC genome paper (Seah et al., in prep.). Consequently, adopting the *Paramecium* nomenclature, we refer to our primary candidate IES excisase as *Blepharisma* PiggyMac (BPgm) and the other homologs as BPgm-Likes (BPgmLs).

PiggyMac homologs are also present in other heterotrich ciliates but have not yet been described because of genome assembly or annotation challenges. Using BPgm as a query sequence, we found convincing homologs containing the conserved catalytic DDD-motif in a genome assembly of the heterotrichous ciliate *Condyllostoma magnum* (TBLASTN e-value 2e-24 to 2e-37). While we failed to detect the DDE_Tnp_1_7 domain in predicted genes of the heterotrich *Stentor coeruleus*, we were able to detect a relatively weak adjacent TBLASTN matches split across two frames in its draft MAC genome (e-value 7e-15; SteCoe_contig_741 positions 6558-5475). After joining ORFs corresponding to this region and translating them, we obtained a more convincing DDE_Tnp_1_7 match with HMMER3 (e-value 2e-24). This either

corresponds to a pseudogene or a poorly assembled genomic region. It is also possible that additional PiggyBac homologs were missed in unassembled *Stentor* MAC genome regions.

In addition, we searched for PiggyMac homologs in the MAC genome of the pathogenic oligohymenophorean ciliate *Ichthyophthirius multifiliis* [57]. TBLASTN searches using the *T. thermophila* Tpb2 as a query returned no hits. A HMMER search using hmmscan with a six-frame translation of the *I. multifiliis* MAC genome against the PFAM-A database also did not return any matches with independent E-values (i-E-value) less than 1. We note that based on BUSCO analyses (Figure S1) the *I. multifiliis* genome appears to be less complete than other ciliates we examined. So, a better genome assembly will be needed to investigate the possibility that PiggyBac homologs are encoded elsewhere in this MAC genome.

Other than the PFAM DDE_Tnp_1_7 domain, three of the *Blepharisma* PiggyBac homologs also possess a short, characteristic cysteine-rich domain (CRD) (Figure 6C). This domain is essential for Pgm activity and *Paramecium* IES excision [58]. PiggyBac CRDs have been classified into three different groups [58]. In *Blepharisma*, the CRD consists of five cysteine residues arranged as CxxC-CxxCxxxxH-Cxxx(Y)H (where C, H, Y and x respectively denote cysteine, histidine, tyrosine and any other residue). Two *Blepharisma* homologs possess this CRD without the penultimate tyrosine residue, while the third contains a tyrosine residue before the final histidine. This -YH feature towards the end of the CxxC-CxxCxxxxH-Cxxx(Y)H CRD is shared by all the PiggyBac homologs we found in *Condyllostoma*, as well as the PiggyBac-like element (PBLE) from the bat *Myotis lucifugus* (Piggybat) and the human PiggyBac element-derived (PGBD) proteins PGBD2 and PGBD3. In contrast, PiggyBac homologs from *Paramecium* and *Tetrahymena* have a CRD with six cysteine residues arranged in the variants of the motif CxxC-CxxC-Cx{2-7}Cx{3,4}H, and group together with human PGDB4 and *Spodoptera frugiperda* PBLE (Figure 6C).

Previous experiments involving individual or paired gene knockdowns of most of the ten *Paramecium tetraurelia* PiggyMac(-like) paralogs lead to substantial IES retention, even though only one PiggyMac gene (PGM) has the complete catalytic triad, indicating that all these proteins are functional [55]. To examine whether PiggyMac homologs in *Paramecium* are functionally constrained we examined non-synonymous (d_N) to synonymous substitution rate (d_S) ratios ($\omega = d_N/d_S$) for pairwise codon sequence alignments within and between species, with the latter using two closely related *Paramecium* species (*P. tetraurelia* and *P. octaurelia*). All the d_N/d_S values for pairwise comparisons of each of the catalytically incomplete *P. tetraurelia* PgmLs versus the complete Pgm, were less than 1, ranging from 0.01 to 0.25 (Table S7). All d_N/d_S values for pairwise comparisons between *P. tetraurelia* and *P. octaurelia* PiggyBac orthologs were also substantially less than 1, ranging from 0.02 to 0.11 (Table S8). Since $d_N/d_S = 1$ indicates genes evolving neutrally [59], none of these genes are likely pseudogenes, and all are subject to similar strong purifying selection.

Since only one of the eight homologs has the complete DDD catalytic triad characteristic of functional PiggyBac transposases, this is the only possible catalytically active PiggyMac in *B. stoltei*. The *in vitro* catalytic activity of the *T. ni* PiggyBac transposase is abolished when any of the catalytic residues are substituted by alanine [60]. Conversely, we expect all *Blepharisma* PiggyMac homologs with an incomplete catalytic triad to be catalytically inactive. In pairwise comparisons of each of the catalytically incomplete homologs versus the complete one d_N/d_S ranges from 0.0076 to 0.1351 (Table S9). As these estimates fall in a similar range to those obtained for *Paramecium* PiggyMac/PiggyMac-likes, this indicates that the *B. stoltei* PiggyMac homologs are predominantly subject to comparable purifying selection and unlikely to be pseudogenes, but may have differentiated and acquired new functional roles.

A common origin for most ciliate PiggyBac homologs

To determine whether the *Blepharisma* PiggyBac homologs share a common ciliate ancestor, or whether they arose from independent acquisitions in major ciliate groups, we explored a range of phylogenetic methods. We created a phylogeny of 155 sequences of DDE_Tnp_1_7 domains of PiggyBac homologs which are representatives of putative domesticated transposases from the genus *Blepharisma* (*B. stoltei* ATCC 30299, *B. stoltei* HT-IV, *B. japonicum*), *Condyllostoma magnum*, *Paramecium* spp., *Tetrahymena thermophila*, together with PiggyBac-like element (PBLE [61]) transposases from a stramenopile species, an archaeplastid species, several opisthokont species and an amoebozoan, homologs of the human PiggyBac-derived 1 gene (PGBD1), PGBD3, PGBD4 and PGBD5 homologs (Figure 7; Data S1). The three BPgms from the three *Blepharisma* strains whose MAC genomes we sequenced form a monophyletic clade, with the BPgms 1 and 2 as outgroups. The BPgms 3,4,5,6 and 7 are grouped together with the *Condyllostoma* Pgms (each of which appears to have a complete catalytic triad). These two clades form a heterotrich-specific group in the phylogeny. We observe that the ciliate Pgms and PgmLs largely cluster to form a single clade. *T. thermophila* Tpb7 is an exception which appears to group together with homologs of PGBD5 from *H. sapiens*, *G. gallus* and *D. rerio* but has an aBayes support value of 0.84, much lower than the mode (1.0) and average (0.91) for this phylogeny. The other exceptions are two fungal PBLEs and two red algal PBLEs with low aBayes support within the predominantly monophyletic ciliate clade.

We applied tree topology tests to ascertain the likelihood of a monophyletic clade of ciliate Pgms and PgmLs. Constrained trees were compared by different methods including bootstrap proportion using the REL method [62], a one-sided Kishino-Hasegawa test, a Shimodaira-Hasegawa test [63,64], estimating confidence using expected likelihood weights [65], and an approximately unbiased test [66] (Table S10). A tree with the constraint of monophyly for all

ciliate Pgms and PgmLs lacks support in comparison to an unconstrained tree with all the ciliate Pgms and PgmLs and eukaryotic PBLEs. After excluding *T. thermophila* Tpb7, a tree constrained with a clade consisting of the remaining ciliate Pgms and PgmLs is supported by all tests. In other words, the ciliate PiggyBac homologs except Tpb7 likely originated in a ciliate common ancestor.

Additional domesticated transposases

Three *B. stoltei* ATCC30299 MAC genome-encoded proteins possess PFAM domain DDE_1 (PF03184; Figure 8A). In PFAM version 35 the most common domain combinations (architectures) with DDE_1, aside from proteins with just this domain detected (5898 sequences), are with an N-terminal PFAM domain HTH_Tnp_Tc5 (PF03221) alone (2240 sequences), and both an N-terminal CENP-B_N domain (PF04218) and central HTH_Tnp_Tc5 domain (1255 sequences). All the *B. stoltei* proteins possess the HTH_Tnp_Tc5 domain. Though pairwise sequence identity is low amongst the *Blepharisma* proteins (avg. 28.3%) in their multiple sequence alignment, the CENP-B_N domain in one of them appears to align reasonably well to corresponding regions in the two proteins lacking this domain, suggesting that it deteriorated in this pair, beyond the recognition capabilities of HMMER3 with the given PFAM domain model and search parameters.

The N-terminal domains of *Homo sapiens* CENP-B proteins (corresponding to PFAM: CENP-B_N, PF04218, and HTH_Tnp_Tc5) are able to bind 17 bp centromeric DNA repeats and are involved in centromeric binding during chromosome segregation [67]. Potential convergent domestication of transposases with these domains has been proposed for mammalian and fission yeast lineages [68]. The PDC2 protein in budding yeast *Saccharomyces cerevisiae*, which regulates pyruvate decarboxylase transcription [69], is considered structurally

homologous to human CENP-B and shares the same pair of domains, but has not been reported to play a role in centromere binding [68,70]. The CENP-B_N domain abbreviation is somewhat unfortunate, since the domain is also characteristic of numerous transposases with no known or expected role in centromeric function, notably the Tigger and PogoR families, and so presumably originated as a transposase domain [71]. As judged from searches of the UniProt database, the characteristic N-terminal CENP-B DNA-binding domain (PF04218) is detectable in a few protist clades (e.g., Rhodophyta and Cryptophyceae), but, among alveolates, only in *Stentor coeruleus*. BLASTp matches for all three proteins in GenBank are annotated either as Jerky or Tigger homologs (Jerky transposases belong to the Tigger transposase family [71]). Given that none of the *B. stoltei* DDE_1 domain proteins appears to have a complete catalytic triad, we think it is unlikely they are involved in transposition or IES excision.

Six MAC-encoded transposases containing the DDE_3 (PF13358) domain are present in *B. stoltei*, all of which are substantially upregulated in MAC development (Figure 8B). Five of these possess the complete DDE catalytic triad. The DDE_3 domain is characteristic of DDE transposases encoded by the Telomere-Bearing Element transposons (TBEs) of *Oxytricha trifallax* [72,73], which, despite being MIC genome-limited, are thought to be involved in IES excision, rather than a domesticated PiggyMac [74]. Other DDE_3-containing transposons, called Tec elements, are found in another spirotrichous ciliate, *Euplotes crassus*, but no role in genome reorganization has been established for these [75]. TBEs and Tec elements do not share obvious features with one another, other than both possessing an encoded protein belonging to the IS630-Tc1 transposase (super)-family [76]. All six DDE_3 genes in *B. stoltei* ATCC 30299 have at least 150× HiFi read coverage, consistent with the regions encoding these being *bona fide* macronuclear DNA rather than MIC DNA contaminants.

The *B. stoltei* DDE_3 domain transposases appear to be more closely related to the IS630 family than to *Oxytricha* TBE transposases and *Euplotes* Tec transposases, and thus acquired independently of them. BLASTP searches of GenBank NR with these proteins as queries, returned bacterial or fungal proteins for most of the hundred top hits, and among these the most common classifications are “IS630 family” transposases. One of the top hits is a MIC genome-encoded protein in *Oxytricha trifallax* with a DDE_3 domain which is not a TBE transposase (GenBank accession: KEJ83017.1). IS630 transposases diverge considerably from Tc1-Mariner transposases, and hence are considered an outgroup to them [77]. On the other hand, IS630-related transposases encoded by Anchois transposons have been detected in the *Paramecium tetraurelia* MIC genome [35]. Since all but one of the *B. stoltei* paralogs appear to possess complete catalytic triads, the possibility that they may excise a subset of IESs needs to be considered.

Among other ciliates with draft MAC genomes we examined, the PFAM IS1595- and MULE transposase-like domains (PF12762 and PF10551) have so far only been observed in the spirotrichs *Oxytricha* and *Stylonychia* [39,40]. DDE_Tnp_IS1595 domains are characteristic of the Merlin transposon superfamily and MULE is part of the Mutator transposon superfamily [54]. Currently no particular functions have been ascribed to these proteins in these ciliates, but they are substantially upregulated during the development of these ciliates [39,56]. Both transposase-like domains are found in MAC-encoded proteins in *B. stoltei* ATCC 30299 and the underlying genes are also upregulated during *Blepharisma* MAC development (Figure 9A, B). The genes encoding DDE_Tnp_IS1595 and MULE proteins, like the genes encoding the DDE_1 and DDE_3 proteins, appear to lack the flanking terminal inverted repeats characteristic of a functional transposon, as identified by RepeatModeler, consistent with transposase domestication. A number of members of both IS1595 and MULE transposases, including those that are upregulated during new MAC formation, also appear to have complete catalytic triads.

557
558 In addition to the transposases, we also detected a family (> 30 copies) of APE-type non-LTR
559 retrotransposase genes with the two domains characteristic of such retrotransposases, i.e., an
560 APE endonuclease domain (PFAM “exo_endo_phos_2”; PF14529) and a reverse transcriptase
561 domain (PFAM “RVT_1”; PF00078) present on adjacent genes. Unlike the conventional
562 transposase-derived genes in *B. stoltei*, the expression of all these genes throughout the
563 conditions we examined is very low to negligible, and some of them also appear to be truncated
564 pseudogenes (Data S3; workbook “RVT1 + exo_endo_phos_2”). Since it is necessary to
565 understand the impact of the presence of IESs in some of these genes, their detailed analyses
566 will be reported in the context of the *Blepharisma stoltei* MIC genome, where their nature and
567 origin will be more evident.

568 Discussion

569 The genus *Blepharisma* represents one of the earliest diverging ciliate lineages, the
570 heterotrichs, forming an outgroup to the best-studied and deeply divergent oligohymenophorean
571 and spirotrich ciliates [7]. *Blepharisma* species thus provide an excellent vantage point to
572 compare unique processes that have accompanied the evolution of nuclear and genomic
573 dimorphism in ciliates, particularly the extensive genomic reorganization that occurs during the
574 new macronucleus (MAC) development. Comparisons between *Blepharisma* and other ciliates
575 also enable deeper consideration of the potential state of the ciliate common ancestor. The
576 annotated draft *B. stoltei* ATCC 30299 MAC genome and associated transcriptomic data provide
577 the basis for comparative studies of key characteristics of genome reorganization processes.

A minichromosomal architecture with pronounced telomere

addition site variation

Ciliate MAC genome architectures span from longer DNA molecules in the oligohymenophoreans *Paramecium tetraurelia* and *Tetrahymena thermophila*, which are more similar to conventional eukaryotic chromosomes, to numerous, predominantly single-gene nanochromosomes of spirotrichs like *Oxytricha trifallax* and *Euplotes crassus* [3]. The *B. stoltei* ATCC 30299 MAC genome organization (and like that of *Paramecium bursaria* [44]) lies between these two extremes with many minichromosomes representing numerous alternative telomere addition sites. This form of natural sequence heterogeneity creates significant new challenges for generating single representative consensus sequences, because genome assembly methods assume chromosome boundaries are generally consistent, as they are in multicellular eukaryote models, down to the terminology used to describe them [78]. Fortunately, the reference *Blepharisma* MAC genome is largely homozygous, promoting contiguity in the final, 41 Mb assembly of 64 two-telomere-capped contigs. Nevertheless, plenty of room for refinement remains, including development of new ways to better assemble, describe and organize the constituents of genomes with such extreme sequence variability.

Despite the abundance of *Blepharisma* MAC genome telomeres, we did not detect a typical ncRNA gene corresponding to the telomerase RNA component (TERC) of the ribozyme responsible for telomere synthesis in the MAC genome. We suspect this is due to ncRNAs presenting a far greater challenge to detect than protein-coding genes and the presence of highly divergent ncRNA with insufficient similarity to the handful of taxonomically-restricted TERCs identified in oligohymenophorean and spirotrich ciliates and other eukaryotes so far. On the other hand, five homologs of POT1, the canonical telomere-binding protein, are present

(Figure S4). One of these is highly upregulated when the new MAC genome is forming, and presumably like model ciliates the DNA is fragmented and replicated, requiring further telomere synthesis and telomere-binding proteins.

A potentially catalytically active PiggyMac is present in

Blepharisma

In current models of IES excision, MIC-limited sequences are demarcated by deposition of methylation marks on histones in an sRNA-dependent process [5]. These sequences are then recognized by domesticated transposases whose excision is supported by additional proteins, including those that recognize these marks [5]. Together with MIC sequencing and from the same time course we employed in the present study for RNA-seq we have observed abundant, development-specific sRNA production in *Blepharisma* like that in other ciliates (Seah et al. in prep), and so we examined the expression of potential genes responsible for their biogenesis, along with other genes highly upregulated during new MAC formation. Homologs of proteins implicated in ciliate genome reorganization were present among the genes most highly differentially upregulated during new MAC development, notably including Dicer-like and Piwi proteins, and domesticated transposases.

Following this study, we will report detailed analyses of a draft *B. stoltei* ATCC 30299 MIC genome, which possesses abundant and typically TA-bound IESs (Seah et al. in prep). In oligohymenophorean ciliates *Tetrahymena* and *Paramecium* there is a considerable body of evidence that PiggyBac homologs are responsible for IES excision [35,52,53,55,79]. The responsible IES excisases in the less-studied spirotrich ciliates, *Oxytricha*, *Stylonychia* and *Euplotes*, are not as evident. In *Oxytricha* the TBE transposases are considered to be involved in IES excision, but, unlike the MAC genome-encoded primary IES excisase (Tpb2) in

Tetrahymena and all the *Paramecium* PiggyMacs and PiggyMac-likes, are encoded by full-length transposons present in the MIC genome, but absent from the MAC [74]. The considerable heterogeneity in *Oxytricha* IES boundaries and “unscrambling” of a large subset of them [56], together with the observation of pronounced developmental upregulation of numerous additional MAC- and MIC-encoded transposases raises the possibility that additional transposases other than those of TBEs could also be involved in IES excision [39,56]. Knowledge of IESs in other ciliates is sparse (primarily confined to the phyllopharyngean *Chilodonella uncinata* [80,81]), and, as far as we are aware, no specific IES excisases have been proposed for them.

Since the oligohymenophorean ciliate PiggyBac homologs are clear IES excisases, we sought and found eight homologs of these genes in the *B. stoltei* ATCC 30299 MAC genome. *Blepharisma* is the first ciliate genus aside from *Tetrahymena* and *Paramecium* in which such proteins have been reported. Additional searches revealed clear PiggyBac homologs in *Condyllostoma magnum*, and a weaker pair of matches in *Stentor coeruleus*, suggesting that these are a common feature of heterotrich ciliates. Reminiscent of *Paramecium tetraurelia*, in which just one of the nine PiggyBac homologs, PiggyMac, has a complete DDD catalytic triad [55], a single *B. stoltei* PiggyBac homolog has a complete DDD catalytic triad. The gene encoding this protein in *B. stoltei* ATCC 30299 is highly upregulated during MAC development, together with two other homologs that lack a complete catalytic triad. As is characteristic of PiggyBac homologs, each of these three homologs also has a C-terminal, cysteine-rich, zinc finger domain, which most closely resemble those of *Condyllostoma magnum* homologs. The organization of the heterotrich PiggyBac homolog zinc finger domains appears to be more similar to that of comparable domains of *Homo sapiens* PGBD2 and PGBD3 homologs than that of the zinc finger domains in *Paramecium* and *Tetrahymena* PiggyBac homologs (which are more similar to the *Homo sapiens* PGBD4 zinc fingers [58]).

In *Paramecium*, since the discovery of multiple PiggyBac homologs (PiggyMac-likes), there have been questions about their role, since, aside from PiggyMac, they all have incomplete catalytic triads, and are thus likely catalytically inactive, but nevertheless their gene knockdowns lead to pronounced IES retention [55]. It has therefore been proposed that the PiggyMac-likes may function as heteromeric multi-subunit complexes in conjunction with PiggyMac during DNA excision [55]. On the other hand cryo-EM structures available for moth PiggyBac transposase support a model in which these proteins function as a homodimeric complex *in vitro* [82]. Furthermore, the primary *Tetrahymena* PiggyBac, Tpb2, is able to perform cleavage *in vitro*, without the assistance of other PiggyBac homologs [53]. In other eukaryotes, domesticated PiggyBacs without complete catalytic triads are thought to be retained by virtue of useful DNA-binding roles [83]. One possibility for such purely DNA-binding transposases in ciliates could be in competitively regulating (taming) the excision of DNA by the catalytically active transposases. Future experimental analyses of the BPgm and the BPgm-likes could aid in resolving the conundrums and understanding of possible interactions between catalytically active and inactive transposases.

Domesticated PiggyBac homologs are the main candidate IES excisases in *Blepharisma*

In addition to the PiggyBac homologs, we also found MAC genome-encoded transposases with the PFAM domains “DDE_1”, “DDE_3”, “DDE_Tnp_IS1595” and “MULE” in *Blepharisma*. All the genes encoding proteins with these domains lack flanking terminal repeats characteristic of active transposons, suggesting these genes are further classes of domesticated transposases. In *Blepharisma* and numerous other organisms, the DDE_1 domains co-occur with CENPB domains. Two such proteins represent totally different proposed exaptations in mammals (centromere-binding protein) and fission yeast (regulatory protein) [68–70]. Given the great

evolutionary distances involved, there is no specific reason to expect that the *Blepharisma* homologs have either function. None of the three proteins with co-occurring DDE_1 and CENPB domains have a complete catalytic triad, making it unlikely that these are active transposases or IES excisases, though all three are noticeably upregulated during MAC development. Six proteins with the PFAM domain DDE_3 are also encoded by *Blepharisma* MAC genes, of which five possess a complete catalytic triad. DDE_3 domains are also characteristic of TBE transposases in *Oxytricha* and Tec transposases in *Euplotes*. All the “DDE_3” protein genes are upregulated during conjugation in *B. stoltei*, and are expressed at particularly high levels during development of the new MAC. A number of DDE_Tnp_IS1595 and MULE domain-containing proteins have complete catalytic triads and also show pronounced upregulation during *B. stoltei* MAC development.

Upon excision, classical cut-and-paste transposases in eukaryotes typically leave behind additional bases, notably including those of the target-site duplication that arose when they were inserted, forming a “footprint” [84]. PiggyBac homologs are unique in performing precise, “seamless” excision in eukaryotes [85], conserving the number of bases at the site of transposon insertion after excision, a property that makes them popular for genetic engineering [82]. *Tetrahymena* Tpb2 is the one exception among PiggyBac homologs associated with imprecise excision in this eukaryote [53]. Since intragenic IESs are abundant in *Blepharisma*, like *Paramecium* and unlike *Tetrahymena*, it is essential that these are excised precisely without addition or removal of nucleotides, which, at best, could result in translation of additional amino acids and, at worst, translation frameshifts.

Though there are clearly numerous additional domesticated transposases with complete catalytic triads and whose genes are substantially upregulated during *Blepharisma* development, whether they are capable of excision, and if this is precise, needs to be

established. *Tetrahymena* has distinct domesticated transposases that excise different subsets of IESs, namely those that are predominant, imprecisely excised and intergenic (by Tpb2) [53], versus those that are rare, precisely excised and intragenic (by Tpb1 and Tpb6) [79,86]. We could envisage if the additional *Blepharisma* domesticated transposases are still capable of excision, but not a precise form, an involvement in excision of a subset of the numerous intergenic IESs. Thus, while we cannot rule out the involvement of other possible domesticated transposases and PiggyMac-likes, currently *B. stoltei* PiggyMac is the primary candidate IES excisase.

A possible PiggyBac homolog ciliate common ancestor

In light of the discovery of PiggyMacs in heterotrichous ciliates, we investigated whether the last common ancestor of ciliates also possessed a PiggyMac. Phylogenetic analyses, together with tree topology tests, indicate that the oligohymenophorean and heterotrichous PiggyMacs and PiggyMac-likes form a monophyletic clade. The lack of PiggyBac homologs in other ciliate classes such as the Spirotrichea, and potentially the oligohymenophorean *Ichthyophthirius multifiliis* (with the caveat of apparent lower genome completeness), raises the question whether the PiggyMacs were lost in these lineages or were gained independently from the same source by both heterotrichs and a subset of oligohymenophoreans. We think the former is more likely. However, the alternative cannot be completely dismissed, because non-model ciliates, where the genome assembly quality allows reliable gene and domain annotations, have only been sparsely sampled.

Future directions

The *B. stoltei* ATCC 30299 MAC genome together with a corresponding MIC genome (Seah et al., in prep.) pave the way for future investigations of a peculiar, direct pathway to new MAC

725 genome development which skips the upstream complexity of meiotic and fertilization-
 726 associated nuclear developmental processes of the standard pathway [9]. The pair of *B. stoltei*
 727 strains we have are both now low frequency selfers, in which the conventional, indirect MAC
 728 development pathway dominates. Comparisons with fresh, high frequency *Blepharisma* selfers
 729 collected from the wild will facilitate comparative gene expression analyses with the direct MAC
 730 development pathway, which will assist in distinguishing expression upregulation due to meiotic
 731 and mitotic processes preceding indirect new MAC development. As a representative of one of
 732 the two deepest branching ciliate lineages, the assembled somatic genome will facilitate
 733 investigations beyond those reported here into the enigmatic origins of nuclear and genomic
 734 dualism within single cells.

Materials and Methods

Strains and localities

The strains used and their original isolation localities were: *Blepharisma stoltei* ATCC 30299, Lake Federsee, Germany [24]; *Blepharisma stoltei* HT-IV, Aichi prefecture, Japan; *Blepharisma japonicum* R1072, from an isolate from Bangalore, India [22].

Cell cultivation, harvesting and cleanup

For genomic DNA isolation *B. stoltei* ATCC 30299 and HT-IV cells were cultured in Synthetic Medium for *Blepharisma* (SMB) [87] at 27°C. *Blepharisma*s were fed *Chlorogonium elongatum* grown in Tris-acetate phosphate (TAP) medium [88] at room temperature. *Chlorogonium* cells were pelleted at 1500 g at room temperature for 3 minutes to remove most of the TAP medium, and resuspended in 50 mL SMB. 50 ml of dense *Chlorogonium* was used to feed 1 litre of *Blepharisma* culture once every three days.

Blepharisma stoltei ATCC 30299 and HT-IV cells used for RNA extraction were cultured in Lettuce medium inoculated with *Enterbacter aerogenes* and maintained at 25°C [89].

Blepharisma cultures were concentrated by centrifugation in pear-shaped flasks at 100 g for 2 minutes using a Hettich Rotanta 460 centrifuge with swing out buckets. Pelleted cells were washed with SMB and centrifuged again at 100 g for 2 minutes. The washed pellet was then transferred to a cylindrical tube capped with a 100 µm-pore nylon membrane at the base and immersed in SMB to filter residual algal debris from the washed cells. The cells were allowed to diffuse through the membrane overnight into the surrounding medium. The next day, the

cylinder with the membrane was carefully removed while attempting to minimize dislodging any debris collected on the membrane. Cell density after harvesting was determined by cell counting under the microscope.

DNA isolation, library preparation and sequencing

B. stoltei macronuclei were isolated by sucrose gradient centrifugation [4]. DNA was isolated with a Qiagen 20/G genomic-tip kit according to the manufacturer's instructions. Purified DNA from the isolated MACs was fragmented, size selected and used to prepare libraries according to standard PacBio HiFi SMRTbell protocols. The libraries were sequenced in circular consensus mode to generate HiFi reads.

Total genomic DNA from *B. stoltei* HT-IV and *B. stoltei* ATCC 30299 was isolated with the SigmaAldrich GenElute Mammalian genomic DNA kit. A sequencing library was prepared with a NEBnext FS DNA Library Prep Kit for Illumina and sequenced on an Illumina HiSeq 3000 sequencer, generating 150 bp paired-end reads.

Total genomic DNA from *B. japonicum* was isolated with the Qiagen MagAttract HMW DNA kit. A long-read PacBio sequencing library was prepared using the SMRTbell® Express Template Preparation Kit 2.0 according to the manufacturers' instructions and sequenced on an PacBio Sequel platform with 1 SMRT cell. Independently, total genomic DNA from *B. japonicum* was isolated with the SigmaAldrich GenElute Mammalian genomic DNA kit and an sequencing library was prepared with the TruSeq Nano DNA Library Prep Kit (Illumina) and sequenced on an Illumina NovaSeq6000 to generate 150 bp paired-end reads.

Gamone 1/ Cell-Free Fluid (CFF) isolation and conjugation activity

assay

B. stoltei ATCC 30299 cells were cultured and harvested and concentrated to a density of 2000 cells/mL according to the procedure described in “Cell cultivation, Harvesting and Cleanup”. This concentrated cell culture was incubated overnight at 27°C. The next day, the cells were harvested, and the supernatant collected and preserved at 4°C at all times after extraction. The supernatant was then filtered through a 0.22 µm-pore filter. BSA (10 mg/mL) was added to produce the final CFF at a final BSA concentration of 0.01%.

To assess the activity of the CFF, serial dilutions of the CFF were made to obtain the gamone activity in terms of units (U) [90]. The activity of the isolated CFF was 2^{10} U.

Conjugation time course and RNA isolation for high-throughput

sequencing

B. stoltei cells for the complementary strains, ATCC 30299 and HT-IV, were cultivated and harvested by gentle centrifugation to achieve a final cell concentration of 2000 cells/ml for each strain. Non-gamone treated ATCC 30299 (A1) and HT-IV cells (H1) were collected (time point: - 3 hours). Strain ATCC 30299 cells were then treated with synthetic gamone 2 (final concentration 1.5 µg/mL) and strain HT-IV cells were treated with cell-free fluid with a gamone 1 activity of $\sim 2^{10}$ U/ml for three hours (Figure S8).

Homotypic pair formation in both cultures was checked after three hours. More than 75% of the cells in both cultures formed homotypic pairs. At this point the samples A2 (ATCC 30299) and H2 (HT-IV) were independently isolated for RNA extraction as gamone-treated control cells just

before mixing. For the rest of the culture, homotypic pairs in both cultures were separated by pipetting them gently with a wide-bore pipette tip. Once all pairs had been separated, the two cultures were mixed together. This constitutes the experiment's 0-h time point. The conjugating culture was observed and samples collected for RNA isolation or cell fixation at 2 h, 6 h, 14 h, 18 h, 22 h, 26 h, 30 h and 38 h (Figure S8). Further details of the sample staging approach are described in [9] and [51]. At each time point including samples A1, H1, A2 and H2, 7 mL of culture was harvested for RNA-extraction using Trizol. The total RNA obtained was then separated into a small RNA fraction < 200 nt and a fraction with RNA fragments > 200 nt using the Zymo RNA Clean and Concentrator-5 kit according to the manufacturer's instructions. RNA-seq libraries were prepared by BGI according to their standard protocols and sequenced on a BGISEq 500 instrument.

Separate 2 mL aliquots of cells at each time point for which RNA was extracted were concentrated by centrifuging gently at 100 rcf. 50 µL of the concentrated cells were fixed with Carnoy's fixative (ethanol:acetic acid, 6:1), stained with DAPI and imaged to determine the state of nuclear development [9].

Cell fixation and imaging

B. stoltei cells were harvested as above ("Cell cultivation"), and fixed with an equal volume of "ZFAE" fixative, containing zinc sulfate (0.25 M, Sigma Aldrich), formalin, glacial acetic acid and ethanol (Carl Roth), freshly prepared by mixing in a ratio of 10:2:2:5. Fixed cells were pelleted (1000 g; 1 min), resuspended in 1% TritonX-100 in PHEM buffer to permeabilize (5 min; room temperature), pelleted and resuspended in 2% (w/v) formaldehyde in PHEM buffer to fix further (10 min; room temp.), then pelleted and washed twice with 3% (w/v) BSA in TBSTEM buffer

(~10 min; room temp.). For indirect immunofluorescence, washed cells were incubated with primary antibody rat anti-alpha tubulin (Abcam, ab6161; 1:100 dilution in 3% w/v BSA/TBSTEM; 60 min; room temp.) then secondary antibody goat anti-rat IgG H&L labeled with AlexaFluor 488 (Abcam, ab150157, 1:500 dilution in 3% w/v BSA/TBSTEM; 20 min; room temp.). Nuclei were counterstained with DAPI (1 µg/mL) in 3% (w/v) BSA/TBSTEM. A z-stack of images was acquired using a confocal laser scanning microscope (Leica TCS SP8), equipped with a HC PL APO 40× 1.30 Oil CS2 objective and a 1 photomultiplier tube and 3 HyD detectors, for DAPI (405 nm excitation, 420-470 nm emission) and Alexa Fluor 488 (488 nm excitation, 510-530 nm emission). Scanning was performed in sequential exposure mode. Spatial sampling was achieved according to Nyquist criteria. ImageJ (Fiji) [91] was used to adjust image contrast and brightness and overlay the DAPI and AlexaFluor 488 channels. The z-stack was temporally color-coded.

For a nuclear 3D reconstruction (Figure 1B), cells were fixed in 1% (w/v) formaldehyde and 0.25% (w/v) glutaraldehyde. Nuclei were stained with Hoechst 33342 (Invitrogen) (5 µM in the culture media), and imaged with a confocal laser scanning microscope (Zeiss, LSM780) equipped with an LD C-Apochromat 40x/1,1 W Korr objective and a 32 channel GaAsP array detector, with 405 nm excitation and 420-470 nm emission. Spatial sampling was achieved according to Nyquist criteria. The IMARIS (Bitplane) software v8.0.2 was used for three-dimensional reconstructions and contrast adjustments.

For scanning electron microscopy (SEM) cells were fixed in 0.5% osmium tetroxide/2% formaldehyde/2.5% glutaraldehyde in 0.06x PHEM for 5-10 minutes at room temperature. Cells were post-fixed with 2% formaldehyde/2.5% glutaraldehyde in 0.06x PHEM for 4-6 hours at room temperature. Subsequently, samples were dehydrated in a graded ethanol series with 1 day per step followed by critical point drying (Polaron) with CO₂. Finally, the cells were sputter-

coated with a 4 nm thick layer of platinum (CCU-010, Safematic) and examined with a field emission scanning electron microscope (Regulus 8230, Hitachi High Technologies) at an accelerating voltage of 3 kV.

Genome assembly

Two MAC genome assemblies for *B. stoltei* ATCC 30299 (70× and 76× coverage) were produced with Flye (version 2.7-b1585) [92] for the two separate PacBio Sequel II libraries (independent replicates) using default parameters and the switches: --pacbio-hifi -g 45m. The approximate genome assembly size was chosen based on preliminary Illumina genome assemblies of approximately 40 Mb. Additional assemblies using the combined coverage (145×) of the two libraries were produced using either Flye version 2.7-b1585 or 2.8.1-b1676, and the same parameters. Two rounds of extension and merging were then used, first comparing the 70× and 76× assemblies to each other, then comparing the 145× assembly to the former merged assembly. Assembly graphs were all relatively simple, with few tangles to be resolved (Figure S7). Minimap2 [93] was used for pairwise comparison of the assemblies using the parameters: -x asm5 --frag=yes --secondary=no, and the resultant aligned sequences were visually inspected and manually merged or extended where possible using Geneious (version 2020.1.2) [94].

Visual inspection of read mapping to the combined assembly was then used to trim off contig ends where there was little correspondence between the assembly consensus and the mapped reads - which we classify as "cruft". Read mapping to cruft regions was often lower or uneven, suggestive of repeats. Alternatively, these features could be due to trace MIC sequences, or sites of alternative chromosome breakage during development which lead to sequences that are

neither purely MAC nor MIC. A few contigs with similar dubious mapping of reads at internal locations, which were also clear sites of chromosome fragmentation (evident by abundant telomere-bearing reads in the vicinity) were split apart and trimmed back as for the contig ends. Telomere-bearing reads mapped to the non-trimmed region nearest to the trimmed site were then used to define contig ends, adding representative telomeric repeats from one of the underlying sequences mapped to each of the ends. The main genome assembly with gene predictions can be obtained from the European Nucleotide Archive (ENA) (PRJEB40285; accession GCA_905310155). “Cruft” sequences are also available from the same project accession.

Two separate assemblies were generated for *Blepharisma japonicum*. A genome assembly for *Blepharisma japonicum* strain R1072 was generated from Illumina reads, using SPAdes genome assembler (v3.14.0) [95]. An assembly with PacBio Sequel long reads was produced with Ra (v0.2.1) [96], which uses the Overlap-Layout-Consensus paradigm. The assembly produced with Ra was more contiguous, with 268 contigs, in comparison to 1510 contigs in the SPAdes assembly, and was chosen as the reference assembly for *Blepharisma japonicum* (ENA accession: ERR6474383).

Condyllostoma magnum genomic reads (study accession PRJEB9019) from a previous study [97] were reassembled to improve contiguity and remove bacterial contamination. Reads were trimmed with bbduk.sh from the BBmap package v38.22 (<https://sourceforge.net/projects/bbmap/>), using minimum PHRED quality score 2 (both ends) and k-mer trimming for Illumina adapters and Phi-X phage sequence (right end), retaining only reads ≥ 25 bp. Trimmed reads were error-corrected and reassembled with SPAdes v3.13.0 [95] using k-mer values 21, 33, 55, 77, 99. To identify potential contaminants, the unassembled reads were screened with phyloFlash v3.3b1 [98] against SILVA v132 [99]; the coding density

under the standard genetic code and prokaryotic gene model were also estimated using Prodigal v2.6.3 [100]. Plotting the coverage vs. GC% of the initial assembly showed that most of the likely bacterial contigs (high prokaryotic coding density, lower coverage, presence of bacterial SSU rRNA sequences) had $\geq 40\%$ GC, so we retained only contigs with $< 40\%$ GC as the final *C. magnum* genome bin. The final assembly is available from the ENA bioproject PRJEB48875 (accession GCA_920105805).

All assemblies were inspected with the quality assessment tool QUAST [101].

Variant calling

Illumina total genomic DNA-seq libraries for *B. stoltei* strains ATCC 30299 (ENA accession: ERR6061285) and HT-IV (ERR6064674) were mapped to the ATCC 30299 reference assembly with bowtie2 v2.4.2 [102]. Alignments were tagged with the MC tag (CIGAR string for mate/next segment) using samtools [103] fixmate. The BAM file was sorted and indexed, read groups were added with bamaddrg (commit 9baba65, <https://github.com/ekg/bamaddrg>), and duplicate reads were removed with Picard MarkDuplicates v2.25.1 (<http://broadinstitute.github.io/picard/>). Variants were called from the combined BAM file with freebayes v1.3.2 [104] in diploid mode, with maximum coverage 1000 (option -g). The resultant VCF file was combined and indexed with bcftools v1.12 [103], then filtered to retain only SNPs with quality score > 20 , and at least one alternate allele.

Annotation of alternative telomere addition sites

Alternative telomere addition sites (ATASs) were annotated by mapping PacBio HiFi reads to the curated reference MAC assembly described above, using minimap2 and the following flags: -x asm20 --secondary=no --MD. We expect reads representing alternative telomere additions to

have one portion mapping to the assembly (excluding telomeric regions), with the other portion containing telomeric repeats being soft-clipped in the BAM record. For each mapped read with a soft-clipped segment, we extracted the clipped sequence, and the coordinates and orientation of the clip relative to the reference. We searched for ≥ 24 bp tandem direct repeats of the telomere unit (i.e., ≥ 3 repeats of the 8 bp unit) in the clipped segment with NCRF v1.01.02 [105], which can detect tandem repeats in the presence of noise, e.g., from sequencing error. The orientation of the telomere sequence, the distance from the end of the telomeric repeat to the clip junction ('gap'), and the number of telomere-bearing reads vs. total mapped reads at each junction were also recorded. Junctions with zero gap between telomere repeat and clip junction were annotated as ATASs. The above procedure was implemented in the MILTEL module of the software package BleTIES v0.1.3 [106].

MILTEL output was processed with Python scripts depending on Biopython [107], pybedtools [108], Bedtools [109], and Matplotlib [110], to summarize statistics of junction sequences and telomere permutations at ATAS junctions, and to extract genomic sequences flanking ATASs for sequence logos. Logos were drawn with Weblogo v3.7.5 [111], with sequences oriented such that the telomere would be added on the 5' end of the ATAS junctions.

To calculate the expected minichromosome length, we assumed that ATASs were independent and identically distributed in the genome following a Poisson distribution. About 47×10^3 ATASs were annotated, supported on average by a single read. Given a genome of 42 Mbp at $145 \times$ coverage, the expected rate of encountering an ATAS is $47 \times 10^3 / (145 \times 42 \text{ Mbp})$, so the distance between ATASs (i.e., the minichromosome length) is exponentially distributed with expectation $(145 \times 42 \text{ Mbp}) / 47 \times 10^3 = 130 \text{ kbp}$.

RNA-seq read mapping

To permit correct mapping of tiny introns RNA-seq data was mapped to the *B. stoltei* ATCC 30299 MAC genome using a version of HISAT2 [112] with modified source code, with the static variable minIntronLen in hisat2.cpp lowered to 9 from 20 (change available in the HISAT2 github fork: <https://github.com/Swart-lab/hisat2/>; commit hash 86527b9). HISAT2 was run with default parameters and parameters --min-intronlen 9 --max-intronlen 500. It should be noted that RNA-seq from timepoints in which *B. stoltei* ATCC 30299 and *B. stoltei* HT-IV cells were mixed together were only mapped to the former genome assembly, and so reads for up to three alleles may map to each of the genes in this assembly.

Genetic code prediction

We used the program PORC (Prediction Of Reassigned Codons; available from <https://github.com/Swart-lab/PORC>) previously written to predict genetic codes in protist transcriptomes [97] to predict the *B. stoltei* genetic code. This program was used to translate the draft *B. stoltei* ATCC 30299 genome assembly in all six frames (with the standard genetic code). Like the program FACIL [113] that inspired PORC, the frequencies of amino acids in PFAM (version 34.0) protein domain profiles aligned to the six frame translation by HMMER 3.1b2 [114] (default search parameters; domains used for prediction with conditional E-values < 1e-20), and correspondingly also to the underlying codon, are used to infer the most likely amino acid encoded by each codon (Figure S5).

Gene prediction

We created a wrapper program, Intronarrator, to predict genes in *Blepharisma* and other heterotrichs, accommodating their tiny introns. Intronarrator can be downloaded and installed

together with dependencies via Conda from GitHub (<https://github.com/Swart-lab/Intronarrator>). Intronarrator directly infers introns from spliced RNA-seq reads mapped by HISAT2 from the entire developmental time course we generated. RNA-seq reads densely cover almost the entire *Blepharisma* MAC genome, aside from intergenic regions, and most potential protein-coding genes (Figure 5A). After predicting the introns and removing them to create an intron-minus genome, Intronarrator runs AUGUSTUS (version 3.3.3) using its intronless model. It then adds back the introns to the intronless gene predictions to produce the final gene predictions.

Introns are inferred from “CIGAR” string annotations in mapped RNA-seq BAM files, using the regular expression “[0-9]+M([0-9][0-9])N[0-9]+M” to select spliced reads. For intron inference we only used primary alignments with: MAPQ >= 10; just a single “N”, indicating one potential intron, per read; and at least 6 mapped bases flanking both the 5’ and 3’ intron boundaries (to limit spurious chance matches of a few bases that might otherwise lead to incorrect intron prediction). The most important parameters for Intronarrator are a cut-off of 0.2 for the fraction of spliced reads covering a potential intron, and a minimum of 10 or more spliced reads to call an intron. The splicing fraction cut-off was chosen based on the overall distribution of splicing (Figure S6A-C). From our visual examination of mapped RNA-seq reads and gene predictions, values less than this were typically “cryptic” excision events [115] which remove potentially essential protein-coding sequences, rather than genuine introns. Intronarrator classifies an intron as sense (7389 in total, excluding alternative splicing), when the majority of reads (irrespective of splicing) mapping to the intron are the same strand, and antisense (554 in total) when they are not. The most frequently spliced intron was chosen in rare cases of overlapping alternative intron splicing.

To eliminate spurious prediction of protein-coding genes overlapping ncRNA genes, we also incorporated ncRNA prediction in Intronarrator. Infernal [38] (default parameters; e-value < 1e-6)

was used to predict a restricted set of conserved ncRNAs models (i.e., tRNAs, rRNAs, SRP, and spliceosomal RNAs) from RFAM 14.0 [116]. These ncRNAs were hard-masked (with “N” characters) before AUGUSTUS gene prediction. Both Infernal ncRNA predictions (excluding tRNAs) and tRNA-scan SE 2.0 [117] (default parameters) tRNA predictions are annotated in the *B. stoltei* ATCC 30299 assembly deposited in the European Nucleotide Archive.

Since we found that *Blepharisma stoltei*, like *Blepharisma japonicum* [97], uses a non-standard genetic code, with UGA codon translated as tryptophan, gene predictions use the “The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (transl_table=4)” from the NCBI genetic codes. The default AUGUSTUS gene prediction parameters override alternative (mitochondrial) start codons permitted by NCBI genetic code 4, other than ATG. So, all predicted *B. stoltei* gene coding sequences begin with ATG.

RNA-seq read mapping relative to gene predictions of Contig_1 of *B. stoltei* ATCC30299 was visualized with PyGenomeTracks [118].

Assessment of genome completeness

A BUSCO (version 4.0.2) [119] analysis of the assembled MAC genomes of *B. stoltei* and *B. japonicum* was performed on the set of predicted proteins (BUSCO mode -prot) using the BUSCO Alveolata database. The completeness of the *Blepharisma* genomes was compared to the protein-level BUSCO analysis of the published genome assemblies of ciliates *T. thermophila*, *P. tetraurelia*, *S. coeruleus* and *I. multifiliis* (Figure S1).

Gene annotation

Pannzer2 [120] (default parameters) and EggNog (version 2.0.1) [121] were used for gene annotation. Annotations were combined and are available from the Max Planck Society's Open Research Repository, Edmond (<https://dx.doi.org/10.17617/3.8c>). Protein domain annotations were performed using hmmscan from HMMER3 (version 3.3, Nov 2019) [114] vs. the PFAM database (Pfam-A.full, 33.0, retrieved on June 23, 2020) with default parameters.

Gene expression analysis

Features from RNA-seq reads mapped to the *B. stoltei* ATCC 30299 genome over the developmental time-course were extracted using featureCounts from the Subread package [122]. Further analysis was performed using the R software environment. Genes with a total read count of less than 50, across all timepoints, were filtered out of the dataset. The remaining genes were passed as a DGEList object to edgeR [123]. Each time point, representing one library, was normalized for library size using the edgeR function calcNormFactors. The normalized read counts were transformed into TPM (transcripts per million) values [124,125]. The TPM-values for different genes were compared across timepoints to examine changes in gene expression. Heatmaps showing log₂(TPM) changes across timepoints were plotted using the tidyverse collection of R packages (<https://www.tidyverse.org/>) and RColorBrewer (<https://rdrr.io/cran/RColorBrewer/>). Tabulated gene expression estimates together with protein annotations are available from Edmond (<https://dx.doi.org/10.17617/3.8c>).

Sequence visualization and analysis

Nucleotide and amino acid sequences were visualized using Geneious Prime (Biomatters Ltd.) [94]. Multiple sequence alignments were performed with MAFFT version 7.450 [126,127]. Phylogenetic trees were constructed with PhyML version 3.3.20180621 [128].

Orthogroup inference and analysis of orthogroup clusters

OrthoFinder version 2.5.2 with default parameters (i.e., using Diamond for searching, MAFFT for multiple alignment and FastTree for phylogenies) was used to define orthogroups, i.e., sets of genes descended from the last common ancestor of the chosen species. Proteomes for the following ciliate species were used: *Tetrahymena thermophila*, *Oxytricha trifallax*, *Stentor coeruleus* (data from ciliate.org [129]); *Euplotes octocarinatus* (EOGD [130]); *Paramecium tetraurelia*, *Paramecium caudatum* (data from ParameciumDB [131]); plus *Perkinsus marinus* ATCC 50983 (GenBank accession: AAXJ000000000) as a non-ciliate outgroup. Orthogroup clusters are available as Data S2, or from Edmond (<https://dx.doi.org/10.17617/3.8c>).

d_N/d_S estimation

We generated pairwise coding sequence alignments of PiggyMac paralog nucleotide sequences from *P. tetraurelia* and *P. octaurelia* using MAFFT version 7.450 [127] [126] (algorithm: “auto”, scoring matrix: 200PAM/k=2, gap open penalty 1.53, offset value 0.123) using the “translation align” panel of Geneious Prime (version 2020.1.2) [94]. PAML version 4.9 [132] was used to estimate d_N/d_S values in pairwise mode (runmode = -2, seqtype = 1, CodonFreq = 2). For *Blepharisma stoltei*, we generated pairwise coding sequence alignments of the *Blepharisma* PiggyMac homolog, BPgm (Contig_49.g1063; BSTOLATCC_MAC17466), with the *Blepharisma* Pgm-likes (BPgmLs) using ClustalW 2.1 [133] (Genetic code: *Blepharisma*, Protein alignment

options: ClustalW alignment, Cost matrix: BLOSUM, Gap open cost: 10, Gap open penalty: 0.1) using “translation align” panel of Geneious Prime (version 2020.1.2). PAML version 4.9 was used to estimate dN/dS values in pairwise mode (runmode = -2, seqtype = 1, CodonFreq = 2).

Phylogenetic analysis

Protein sequences of PBLEs and PGDBs were obtained from Bouallègue et al [61]. Protein sequences of *Paramecium* and *Tetrahymena* Pgms and PgmLs were obtained from ParameciumDB [131] (PGM, PGMLs1-5) and ciliate.org [129] (Tpb1, Tpb2, Tpb7, LIA5), respectively. *Condyllostoma* and *Blepharisma* Pgms and PgmLs were obtained from genome assemblies (accessions GCA_920105805 and GCA_905310155, respectively). Other PiggyBac protein sequences (*T. ni*, *D. purpureum*) were obtained from Uniprot [134]. Sequence manipulation was done using Geneious (Biomatters Ltd.). The Geneious plug-in for InterProScan [135] was used to identify DDE_Tnp_1_7 domains using the PFAM-A database [136]. The DDE_Tnp_1_7 domains were extracted and aligned using the MAFFT plug-in (v7.450) for Geneious [127] (Algorithm: Auto, Scoring matrix: BLOSUM62, Gap open penalty: 1.53, Offset value: 0.123). The best-fit model of evolution for each alignment was identified using ModelTest-NG [137] (modeltest-ng -i <input_file>.phy -d aa -p 4 -T phyml -o <output_file>.out). The best model according to Bayesian criteria (BIC) were used to create phylogenetic trees using PhyML [128] [phyml -i <input_file>.phy -d aa -m LG -f m -v 0 -a e -c 4 -o tlr]. Phylogenetic trees thus generated were visualized with FigTree (v1.4.4) (Andrew Rambaut, <http://tree.bio.ed.ac.uk/>). Tree topology tests were performed using IQtree (v2.1.4-beta) [138] (iqtree -s <input file> -m LG+G4 -z <unconstrained and constrained treefiles> -n 0 -zb 10000 -au -nt AUTO).

Repeat annotation

Interspersed repeat element families were predicted with RepeatModeler v2.0.1 (default settings, random number seed 12345) with the following dependencies: rmbblast 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 with RepeatMasker, using rmbblast as the search engine.

Terminal inverted repeats (TIRs) 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.

Data availability

The draft *Blepharisma stoltei* ATCC 30299 macronuclear genome assembly is accessible from bleph.ciliate.org and from the European Nucleotide Archive (ENA) bioproject PRJEB40285 under the accession GCA_905310155. PacBio CCS reads (ERR5873783 and ERR5873334)

and subreads (ERR5962314) used to assemble the genome are also available from ENA. Illumina DNA-seq data for the *B. stoltei* ATCC 30299 and HT-IV strains is available from accessions ERR6061285 and ERR6064674, respectively. The RNA-seq developmental time course is available from the bioproject PRJEB45374 (accessions ERR6049461-ERR6049485). Illumina and PacBio Sequel sequencing data for *Blepharisma japonicum* strain R1702 is available from the ENA bioproject PRJEB46921 (Illumina accessions: ERR6473251, ERR6474356; PacBio accession: ERR6474383).

Acknowledgements

This paper is dedicated in memory of Akio Miyake and his decades of inspirational *Blepharisma* research. We thank Federico Buonanno for the provision of *B. stoltei* ATCC 30299 cells and culturing advice, Christa Lanz and the MPI for Biology's genome center, Jürgen Berger for SEM imaging, and Sebastien Colin and the MPI for Biology's Light Microscopy Facility for the 3D nuclear reconstruction.

Bibliography

1. Martin WF, Garg S, Zimorski V. Endosymbiotic theories for eukaryote origin. *Philos Trans R Soc Lond B Biol Sci.* 2015;370: 20140330. doi:10.1098/rstb.2014.0330
2. Orias E. On the evolution of the karyorelict ciliate life cycle: heterophasic ciliates and the origin of ciliate binary fission. *BioSystems.* 1991;25: 67–73. doi:10.1016/0303-2647(91)90013-B

- 1126 3. Prescott DM. The DNA of ciliated protozoa. Microbiol Rev. 1994;58: 233–267.
1127 doi:10.1128/mr.58.2.233-267.1994
- 1128 4. Lauth MR, Spear BB, Heumann J, Prescott DM. DNA of ciliated protozoa: DNA
1129 sequence diminution during macronuclear development of Oxytricha. Cell. 1976;7: 67–
1130 74. doi:10.1016/0092-8674(76)90256-7
- 1131 5. Chalker DL, Meyer E, Mochizuki K. Epigenetics of ciliates. Cold Spring Harb Perspect
1132 Biol. 2013;5: a017764. doi:10.1101/cshperspect.a017764
- 1133 6. Vogt A, Goldman AD, Mochizuki K, Landweber LF. Transposon domestication versus
1134 mutualism in ciliate genome rearrangements. PLoS Genet. 2013;9: e1003659.
1135 doi:10.1371/journal.pgen.1003659
- 1136 7. Lynn DH, editor. The Ciliated Protozoa. Dordrecht: Springer Netherlands; 2010.
1137 doi:10.1007/978-1-4020-8239-9
- 1138 8. Giese AC. Blepharisma: The Biology of a Light-sensitive Protozoan. illustrated ed.
1139 Stanford University Press; 1973.
- 1140 9. Miyake A, Rivola V, Harumoto T. Double paths of macronucleus differentiation at
1141 conjugation in Blepharisma japonicum. Eur J Protistol. 1991;27: 178–200.
1142 doi:10.1016/S0932-4739(11)80340-8
- 1143 10. Slabodnick MM, Ruby JG, Reiff SB, Swart EC, Gosai S, Prabakaran S, et al. The
1144 Macronuclear Genome of Stentor coeruleus Reveals Tiny Introns in a Giant Cell. Curr
1145 Biol. 2017;27: 569–575. doi:10.1016/j.cub.2016.12.057
- 1146 11. Slabodnick MM, Marshall WF. Stentor coeruleus. Curr Biol. 2014;24: R783-4.
1147 doi:10.1016/j.cub.2014.06.044
- 1148 12. Zhang KS, Blauch LR, Huang W, Marshall WF, Tang SKY. Microfluidic guillotine reveals
1149 multiple timescales and mechanical modes of wound response in Stentor coeruleus.

1150 BMC Biol. 2021;19: 63. doi:10.1186/s12915-021-00970-0

1151 13. Rapport EW, Rapport DJ, Berger J, Kupers V. Induction of conjugation in *Stentor*
1152 *coeruleus*. Trans Am Microsc Soc. 1976;95: 220–224. doi:10.2307/3225068

1153 14. Miyake A, Harumoto T. Asymmetrical cell division in *Blepharisma japonicum*: difference
1154 between daughter cells in mating-type expression. Exp Cell Res. 1990;190: 65–68.
1155 doi:10.1016/0014-4827(90)90144-y

1156 15. Terazima MN, Harumoto T. Defense function of pigment granules in the ciliate
1157 *Blepharisma japonicum* against two predatory protists, *Amoeba proteus* (Rhizopodea)
1158 and *Climacostomum virens* (Ciliata). Zool Sci. 2004;21: 823–828. doi:10.2108/zsj.21.823

1159 16. Kumazawa H. Homopolar Grafting in *Blepharisma japonicum*. Journal of Experimental
1160 Zoology. 1979;207: 1–16.

1161 17. Sugiura M, Shiotani H, Suzaki T, Harumoto T. Behavioural changes induced by the
1162 conjugation-inducing pheromones, gamone 1 and 2, in the ciliate *Blepharisma*
1163 *japonicum*. Eur J Protistol. 2010;46: 143–149. doi:10.1016/j.ejop.2010.01.002

1164 18. Young D. Macronuclear reorganization in *Blepharisma undulans*. Doctoral dissertation.
1165 1937.

1166 19. Friedl E, Miyake A, Heckmann K. Requirement of successive protein syntheses for the
1167 progress of meiosis in *Blepharisma*. Exp Cell Res. 1983;145: 105–113.
1168 doi:10.1016/s0014-4827(83)80013-5

1169 20. Miyake A, Tulli M, Nobili R. Requirement of protein synthesis in the initiation of meiosis
1170 and other nuclear changes in conjugation of *Blepharisma*. Exp Cell Res. 1979;120: 87–
1171 93. doi:10.1016/0014-4827(79)90539-1

1172 21. Inaba F. Conjugation between two strains of *Blepharisma*. J Protozool. 1965;12: 146–
1173 151. doi:10.1111/j.1550-7408.1965.tb01828.x

- 1174 22. Harumoto T, Miyake A, Ishikawa N, Sugibayashi R, Zenfuku K, Iio H. Chemical defense
1175 by means of pigmented extrusomes in the ciliate *Blepharisma japonicum*. *Eur J Protistol*.
1176 1998;34: 458–470. doi:10.1016/S0932-4739(98)80014-X
- 1177 23. Kobayashi M, Miura M, Takusagawa M, Sugiura M, Harumoto T. Two possible barriers
1178 blocking conjugation between different megakaryotypes of *Blepharisma*. *Zool Sci*.
1179 2015;32: 53–61. doi:10.2108/zs140151
- 1180 24. Repak AJ. Encystment and Excystment of the Heterotrichous Ciliate *Blepharisma stoltei*
1181 Isquith. *Journal of Protozoology*. 1968;5: 407–412.
- 1182 25. Salvini M, Durante M, Nobili R. Characterization of macronuclear DNA in *Blepharisma*
1183 *japonicum*. *Protoplasma*. 1983;117: 82–88. doi:10.1007/BF01281787
- 1184 26. Kimball RF. The nature and inheritance of mating types in *Euplotes patella*. *Genetics*.
1185 1942;27: 269–285.
- 1186 27. Katashima RYO. Mating Types in *Euplotes eurystomus*. *J Protozool*. 1959;6: 75–83.
1187 doi:10.1111/j.1550-7408.1959.tb03930.x
- 1188 28. Luporini P, Miceli C, Ortenzi C. Evidence that the ciliate *Euplotes raikovi* releases
1189 mating-inducing factors (gamones). *J Exp Zool*. 1983;226: 1–9.
1190 doi:10.1002/jez.1402260102
- 1191 29. Vallesi A, Giuli G, Bradshaw RA, Luporini P. Autocrine mitogenic activity of pheromones
1192 produced by the protozoan ciliate *Euplotes raikovi*. *Nature*. 1995;376: 522–524.
1193 doi:10.1038/376522a0
- 1194 30. Braun V, Miyake A. Composition of blepharmone, a conjugation-inducing glycoprotein of
1195 the ciliate *Blepharisma*. *FEBS Lett*. 1975;53: 131–134. doi:10.1016/0014-
1196 5793(75)80003-2
- 1197 31. Miyake A, Beyer J. Blepharmone: a conjugation-inducing glycoprotein in the ciliate

1198 blepharisma. Science. 1974;185: 621–623. doi:10.1126/science.185.4151.621

1199 32. Sugiura M, Harumoto T. Identification, characterization, and complete amino acid
1200 sequence of the conjugation-inducing glycoprotein (blepharmone) in the ciliate
1201 *Blepharisma japonicum*. Proc Natl Acad Sci USA. 2001;98: 14446–14451.
1202 doi:10.1073/pnas.221457698

1203 33. Kubota T, Tokoroyama T, Tsukuda Y, Koyama H, Miyake A. Isolation and structure
1204 determination of blepharismine, a conjugation initiating gamone in the ciliate blepharisma.
1205 Science. 1973;179: 400–402. doi:10.1126/science.179.4071.400

1206 34. Klobutcher LA, Herrick G. Developmental genome reorganization in ciliated protozoa:
1207 the transposon link. Prog Nucleic Acid Res Mol Biol. 1997;56: 1–62. doi:10.1016/S0079-
1208 6603(08)61001-6

1209 35. Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury J-M, Denby Wilkes C, et al. The
1210 *Paramecium* germline genome provides a niche for intragenic parasitic DNA:
1211 evolutionary dynamics of internal eliminated sequences. PLoS Genet. 2012;8:
1212 e1002984. doi:10.1371/journal.pgen.1002984

1213 36. Klobutcher LA, Herrick G. Consensus inverted terminal repeat sequence of *Paramecium*
1214 IESs: resemblance to termini of Tc1-related and *Euplotes* Tec transposons. Nucleic
1215 Acids Res. 1995;23: 2006–2013. doi:10.1093/nar/23.11.2006

1216 37. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased
1217 diploid genome assembly with single-molecule real-time sequencing. Nat Methods.
1218 2016;13: 1050–1054. doi:10.1038/nmeth.4035

1219 38. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments.
1220 Bioinformatics. 2009;25: 1335–1337. doi:10.1093/bioinformatics/btp157

1221 39. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, et al. The *Oxytricha trifallax*

1222 macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes.
 1223 PLoS Biol. 2013;11: e1001473. doi:10.1371/journal.pbio.1001473

1224 40. Aeschlimann SH, Jönsson F, Postberg J, Stover NA, Petera RL, Lipps H-J, et al. The
 1225 draft assembly of the radically organized *Stylonychia lemnae* macronuclear genome.
 1226 Genome Biol Evol. 2014;6: 1707–1723. doi:10.1093/gbe/evu139

1227 41. Zagulski M, Nowak JK, Le Mouël A, Nowacki M, Migdalski A, Gromadka R, et al. High
 1228 coding density on the largest *Paramecium tetraurelia* somatic chromosome. Curr Biol.
 1229 2004;14: 1397–1404. doi:10.1016/j.cub.2004.07.029

1230 42. Sheng Y, Duan L, Cheng T, Qiao Y, Stover NA, Gao S. The completed macronuclear
 1231 genome of a model ciliate *Tetrahymena thermophila* and its application in genome
 1232 scrambling and copy number analyses. Sci China Life Sci. 2020;63: 1534–1542.
 1233 doi:10.1007/s11427-020-1689-4

1234 43. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-
 1235 genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 2006;444:
 1236 171–178. doi:10.1038/nature05230

1237 44. Cheng Y-H, Liu C-FJ, Yu Y-H, Jhou Y-T, Fujishima M, Tsai IJ, et al. Genome plasticity in
 1238 *Paramecium bursaria* revealed by population genomics. BMC Biol. 2020;18: 180.
 1239 doi:10.1186/s12915-020-00912-2

1240 45. Yao MC, Yao CH, Monks B. The controlling sequence for site-specific chromosome
 1241 breakage in *Tetrahymena*. Cell. 1990;63: 763–772. doi:10.1016/0092-8674(90)90142-2

1242 46. Cavalcanti ARO, Dunn DM, Weiss R, Herrick G, Landweber LF, Doak TG. Sequence
 1243 features of *Oxytricha trifallax* (class Spirotrichea) macronuclear telomeric and
 1244 subtelomeric sequences. Protist. 2004;155: 311–322. doi:10.1078/1434461041844196

1245 47. Lynch M. The origins of eukaryotic gene structure. Mol Biol Evol. 2006;23: 450–468.

- doi:10.1093/molbev/msj050
48. Roy SW, Hudson AJ, Joseph J, Yee J, Russell AG. Numerous fragmented spliceosomal introns, AT-AC splicing, and an unusual dynein gene expression pathway in *Giardia lamblia*. *Mol Biol Evol*. 2012;29: 43–49. doi:10.1093/molbev/msr063
49. Bondarenko VS, Gelfand MS. Evolution of the Exon-Intron Structure in Ciliate Genomes. *PLoS ONE*. 2016;11: e0161476. doi:10.1371/journal.pone.0161476
50. Csuros M, Rogozin IB, Koonin EV. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol*. 2011;7: e1002150. doi:10.1371/journal.pcbi.1002150
51. Sugiura M, Tanaka Y, Suzuki T, Harumoto T. Alternative gene expression in type I and type II cells may enable further nuclear changes during conjugation of *Blepharisma japonicum*. *Protist*. 2012;163: 204–216. doi:10.1016/j.protis.2011.07.007
52. Baudry C, Malinsky S, Restituto M, Kapusta A, Rosa S, Meyer E, et al. PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev*. 2009;23: 2478–2483. doi:10.1101/gad.547309
53. Cheng C-Y, Vogt A, Mochizuki K, Yao M-C. A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in *Tetrahymena thermophila*. *Mol Biol Cell*. 2010;21: 1753–1762. doi:10.1091/mbc.e09-12-1079
54. Yuan Y-W, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci USA*. 2011;108: 7884–7889. doi:10.1073/pnas.1104208108
55. Bischerour J, Bhullar S, Denby Wilkes C, Régnier V, Mathy N, Dubois E, et al. Six

1270 domesticated PiggyBac transposases together carry out programmed DNA elimination in
1271 Paramecium. *eLife*. 2018;7. doi:10.7554/eLife.37927

1272 56. Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM, Swart EC, et al. The
1273 architecture of a scrambled genome reveals massive levels of genomic rearrangement
1274 during development. *Cell*. 2014;158: 1187–1198. doi:10.1016/j.cell.2014.07.034

1275 57. Coyne RS, Hannick L, Shanmugam D, Hostetler JB, Bami D, Joardar VS, et al.
1276 Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-living
1277 relatives and a host species provide insights into adoption of a parasitic lifestyle and
1278 prospects for disease control. *Genome Biol*. 2011;12: R100. doi:10.1186/gb-2011-12-10-
1279 r100

1280 58. Guérineau M, Bessa L, Moriau S, Lescop E, Bontems F, Mathy N, et al. The unusual
1281 structure of the PiggyMac cysteine-rich domain reveals zinc finger diversity in PiggyBac-
1282 related transposases. *Mob DNA*. 2021;12: 12. doi:10.1186/s13100-021-00240-4

1283 59. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates
1284 under realistic evolutionary models. *Mol Biol Evol*. 2000;17: 32–43.
1285 doi:10.1093/oxfordjournals.molbev.a026236

1286 60. Mitra R, Fain-Thornton J, Craig NL. piggyBac can bypass DNA synthesis during cut and
1287 paste transposition. *EMBO J*. 2008;27: 1097–1109. doi:10.1038/emboj.2008.41

1288 61. Bouallègue M, Rouault J-D, Hua-Van A, Makni M, Capy P. Molecular Evolution of
1289 piggyBac Superfamily: From Selfishness to Domestication. *Genome Biol Evol*. 2017;9:
1290 323–339. doi:10.1093/gbe/evw292

1291 62. Kishino H, Miyata T, Hasegawa M. Maximum likelihood inference of protein phylogeny
1292 and the origin of chloroplasts. *J Mol Evol*. 1990;31: 151–160. doi:10.1007/BF02109483

1293 63. Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the

1294 evolutionary tree topologies from DNA sequence data, and the branching order in
1295 hominoidea. J Mol Evol. 1989;29: 170–179. doi:10.1007/BF02100115

1296 64. Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H. Appropriate likelihood ratio
1297 tests and marginal distributions for evolutionary tree models with constraints on
1298 parameters. Mol Biol Evol. 2000;17: 798–803.
1299 doi:10.1093/oxfordjournals.molbev.a026358

1300 65. Strimmer K, Rambaut A. Inferring confidence sets of possibly misspecified gene trees.
1301 Proc Biol Sci. 2002;269: 137–142. doi:10.1098/rspb.2001.1862

1302 66. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol.
1303 2002;51: 492–508. doi:10.1080/10635150290069913

1304 67. Tanaka Y, Nureki O, Kurumizaka H, Fukai S, Kawaguchi S, Ikuta M, et al. Crystal
1305 structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B
1306 induce kinks in the CENP-B box DNA. EMBO J. 2001;20: 6612–6618.
1307 doi:10.1093/emboj/20.23.6612

1308 68. Casola C, Hucks D, Feschotte C. Convergent domestication of pogo-like transposases
1309 into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol. 2008;25:
1310 29–41. doi:10.1093/molbev/msm221

1311 69. Hohmann S. Characterisation of PDC2, a gene necessary for high level expression of
1312 pyruvate decarboxylase structural genes in *Saccharomyces cerevisiae*. Mol Gen Genet.
1313 1993;241: 657–666. doi:10.1007/BF00279908

1314 70. Mojzita D, Hohmann S. Pdc2 coordinates expression of the THI regulon in the yeast
1315 *Saccharomyces cerevisiae*. Mol Genet Genomics. 2006;276: 147–161.
1316 doi:10.1007/s00438-006-0130-z

1317 71. Gao B, Wang Y, Diaby M, Zong W, Shen D, Wang S, et al. Evolution of pogo, a separate

1318 superfamily of IS630-Tc1-mariner transposons, revealing recurrent domestication events
1319 in vertebrates. *Mob DNA*. 2020;11: 25. doi:10.1186/s13100-020-00220-0

1320 72. Witherspoon DJ, Doak TG, Williams KR, Seegmiller A, Seger J, Herrick G. Selection on
1321 the protein-coding genes of the TBE1 family of transposable elements in the ciliates
1322 *Oxytricha fallax* and *O. trifallax*. *Mol Biol Evol*. 1997;14: 696–706.
1323 doi:10.1093/oxfordjournals.molbev.a025809

1324 73. Williams K, Doak TG, Herrick G. Developmental precise excision of *Oxytricha trifallax*
1325 telomere-bearing elements and formation of circles closed by a copy of the flanking
1326 target duplication. *EMBO J*. 1993;12: 4593–4601. doi:10.1002/j.1460-
1327 2075.1993.tb06148.x

1328 74. Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG, Landweber LF. A functional
1329 role for transposases in a large eukaryotic genome. *Science*. 2009;324: 935–938.
1330 doi:10.1126/science.1170023

1331 75. Jahn CL, Doktor SZ, Frels JS, Jaraczewski JW, Krikau MF. Structures of the *Euplotes*
1332 *crassus* Tec1 and Tec2 elements: identification of putative transposase coding regions.
1333 *Gene*. 1993;133: 71–78. doi:10.1016/0378-1119(93)90226-s

1334 76. Doak TG, Doerder FP, Jahn CL, Herrick G. A proposed superfamily of transposase
1335 genes: transposon-like elements in ciliated protozoa and a common “D35E” motif. *Proc*
1336 *Natl Acad Sci USA*. 1994;91: 942–946. doi:10.1073/pnas.91.3.942

1337 77. Dupeyron M, Baril T, Bass C, Hayward A. Phylogenetic analysis of the Tc1/mariner
1338 superfamily reveals the unexplored diversity of pogo-like elements. *Mob DNA*. 2020;11:
1339 21. doi:10.1186/s13100-020-00212-0

1340 78. Lewin HA, Graves JAM, Ryder OA, Graphodatsky AS, O'Brien SJ. Precision
1341 nomenclature for the new genomics. *Gigascience*. 2019;8.

- doi:10.1093/gigascience/giz086
79. Feng L, Wang G, Hamilton EP, Xiong J, Yan G, Chen K, et al. A germline-limited piggyBac transposase gene is required for precise excision in Tetrahymena genome rearrangement. *Nucleic Acids Res.* 2017;45: 9481–9502. doi:10.1093/nar/gkx652
80. Zufall RA, Sturm M, Mahon BC. Evolution of germline-limited sequences in two populations of the ciliate *Chilodonella uncinata*. *J Mol Evol.* 2012;74: 140–146. doi:10.1007/s00239-012-9493-4
81. Zufall RA, Katz LA. Micronuclear and macronuclear forms of beta-tubulin genes in the ciliate *Chilodonella uncinata* reveal insights into genome processing and protein evolution. *J Eukaryot Microbiol.* 2007;54: 275–282. doi:10.1111/j.1550-7408.2007.00267.x
82. Chen Q, Luo W, Veatch RA, Hickman AB, Wilson MH, Dyda F. Structural basis of seamless excision and specific targeting by piggyBac transposase. *Nat Commun.* 2020;11: 3446. doi:10.1038/s41467-020-17128-1
83. Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, Robertson HM, et al. Molecular evolutionary analysis of the widespread piggyBac transposon family and related “domesticated” sequences. *Mol Genet Genomics.* 2003;270: 173–180. doi:10.1007/s00438-003-0909-0
84. van Luenen HG, Colloms SD, Plasterk RH. The mechanism of transposition of Tc3 in *C. elegans*. *Cell.* 1994;79: 293–301. doi:10.1016/0092-8674(94)90198-8
85. Elick TA, Bauser CA, Fraser MJ. Excision of the piggyBac transposable element in vitro is a precise event that is enhanced by the expression of its encoded transposase. *Genetica.* 1996;98: 33–41. doi:10.1007/BF00120216
86. Cheng C-Y, Young JM, Lin C-YG, Chao J-L, Malik HS, Yao M-C. The piggyBac

1366 transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision
1367 during the developmental rearrangement of key genes in *Tetrahymena thermophila*.
1368 *Genes Dev.* 2016;30: 2724–2736. doi:10.1101/gad.290460.116

1369 87. Miyake A, Beyer J. Cell interaction by means of soluble factors (gamones) in conjugation
1370 of *Blepharisma intermedium*. *Exp Cell Res.* 1973;76: 15–24. doi:10.1016/0014-
1371 4827(73)90413-8

1372 88. Andersen RA. *Algal Culturing Techniques*. 1st Edition. 2004.

1373 89. Miyake A, Harumoto T, Salvi B, Rivola V. Defensive function of pigment granules in
1374 *Blepharisma japonicum*. *Eur J Protistol.* 1990;25: 310–315. doi:10.1016/S0932-
1375 4739(11)80122-7

1376 90. Miyake A. Cell interaction by gamones in *blepharisma*. *Sexual interactions in eukaryotic*
1377 *microbes*. Elsevier; 1981. pp. 95–129. doi:10.1016/B978-0-12-524160-1.50010-2

1378 91. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an
1379 open-source platform for biological-image analysis. *Nat Methods.* 2012;9: 676–682.
1380 doi:10.1038/nmeth.2019

1381 92. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using
1382 repeat graphs. *Nat Biotechnol.* 2019;37: 540–546. doi:10.1038/s41587-019-0072-8

1383 93. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:
1384 3094–3100. doi:10.1093/bioinformatics/bty191

1385 94. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious
1386 Basic: an integrated and extendable desktop software platform for the organization and
1387 analysis of sequence data. *Bioinformatics.* 2012;28: 1647–1649.
1388 doi:10.1093/bioinformatics/bts199

1389 95. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo

1390 assembler. Curr Protoc Bioinformatics. 2020;70: e102. doi:10.1002/cpbi.102

1391 96. Vaser R, Sikic M. Yet another de novo genome assembler. BioRxiv. 2019.

1392 doi:10.1101/656306

1393 97. Swart EC, Serra V, Petroni G, Nowacki M. Genetic Codes with No Dedicated Stop

1394 Codon: Context-Dependent Translation Termination. Cell. 2016;166: 691–702.

1395 doi:10.1016/j.cell.2016.06.020

1396 98. Gruber-Vodicka HR, Seah BKB, Priesse E. phyloFlash: Rapid Small-Subunit rRNA

1397 Profiling and Targeted Assembly from Metagenomes. mSystems. 2020;5.

1398 doi:10.1128/mSystems.00920-20

1399 99. Quast C, Priesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA

1400 ribosomal RNA gene database project: improved data processing and web-based tools.

1401 Nucleic Acids Res. 2013;41: D590-6. doi:10.1093/nar/gks1219

1402 100. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic

1403 gene recognition and translation initiation site identification. BMC Bioinformatics.

1404 2010;11: 119. doi:10.1186/1471-2105-11-119

1405 101. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for

1406 genome assemblies. Bioinformatics. 2013;29: 1072–1075.

1407 doi:10.1093/bioinformatics/btt086

1408 102. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods.

1409 2012;9: 357–359. doi:10.1038/nmeth.1923

1410 103. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of

1411 SAMtools and BCFtools. Gigascience. 2021;10. doi:10.1093/gigascience/giab008

1412 104. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. v2

1413 ed. arXiv; 2012 Jul p. arXiv:1207.3907 [q-bio.GN]. Available:

1414 <https://arxiv.org/abs/1207.3907>

1415 105. Harris RS, Cechova M, Makova KD. Noise-cancelling repeat finder: uncovering tandem
1416 repeats in error-prone long-read sequencing data. *Bioinformatics*. 2019;35: 4809–4811.
1417 doi:10.1093/bioinformatics/btz484

1418 106. Seah BKB, Swart EC. BleTIES: Annotation of natural genome editing in ciliates using
1419 long read sequencing. *BioRxiv*. 2021. doi:10.1101/2021.05.18.444610

1420 107. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely
1421 available Python tools for computational molecular biology and bioinformatics.
1422 *Bioinformatics*. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

1423 108. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for
1424 manipulating genomic datasets and annotations. *Bioinformatics*. 2011;27: 3423–3424.
1425 doi:10.1093/bioinformatics/btr539

1426 109. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
1427 features. *Bioinformatics*. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

1428 110. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng*. 2007;9: 90–95.
1429 doi:10.1109/MCSE.2007.55

1430 111. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator.
1431 *Genome Res*. 2004;14: 1188–1190. doi:10.1101/gr.849004

1432 112. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and
1433 genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;37: 907–915.
1434 doi:10.1038/s41587-019-0201-4

1435 113. Dutilh BE, Jurgelenaite R, Szklarczyk R, van Hijum SAFT, Harhangi HR, Schmid M, et
1436 al. FACIL: Fast and Accurate Genetic Code Inference and Logo. *Bioinformatics*.
1437 2011;27: 1929–1933. doi:10.1093/bioinformatics/btr316

1438 114. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7: e1002195.
1439 doi:10.1371/journal.pcbi.1002195

1440 115. Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, et al. The fitness
1441 cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biol.
1442 2017;18: 208. doi:10.1186/s13059-017-1344-6

1443 116. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam
1444 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids
1445 Res. 2018;46: D335–D342. doi:10.1093/nar/gkx1038

1446 117. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: Improved Detection and
1447 Functional Classification of Transfer RNA Genes. BioRxiv. 2019. doi:10.1101/614032

1448 118. Lopez-Delisle L, Rabbani L, Wolff J, Bhardwaj V, Backofen R, Grüning B, et al.
1449 pyGenomeTracks: reproducible plots for multivariate genomic datasets. Bioinformatics.
1450 2021;37: 422–423. doi:10.1093/bioinformatics/btaa692

1451 119. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al.
1452 BUSCO applications from quality assessments to gene prediction and phylogenomics.
1453 Mol Biol Evol. 2018;35: 543–548. doi:10.1093/molbev/msx319

1454 120. Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server.
1455 Nucleic Acids Res. 2018;46: W84–W88. doi:10.1093/nar/gky350

1456 121. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et
1457 al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology
1458 resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:
1459 D309–D314. doi:10.1093/nar/gky1085

1460 122. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for
1461 assigning sequence reads to genomic features. Bioinformatics. 2014;30: 923–930.

1462 doi:10.1093/bioinformatics/btt656

1463 123. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential
1464 expression analysis of digital gene expression data. *Bioinformatics*. 2010;26: 139–140.
1465 doi:10.1093/bioinformatics/btp616

1466 124. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression
1467 estimation with read mapping uncertainty. *Bioinformatics*. 2010;26: 493–500.
1468 doi:10.1093/bioinformatics/btp692

1469 125. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data:
1470 RPKM measure is inconsistent among samples. *Theory Biosci*. 2012;131: 281–285.
1471 doi:10.1007/s12064-012-0162-3

1472 126. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple
1473 sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30: 3059–
1474 3066. doi:10.1093/nar/gkf436

1475 127. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
1476 improvements in performance and usability. *Mol Biol Evol*. 2013;30: 772–780.
1477 doi:10.1093/molbev/mst010

1478 128. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms
1479 and methods to estimate maximum-likelihood phylogenies: assessing the performance
1480 of PhyML 3.0. *Syst Biol*. 2010;59: 307–321. doi:10.1093/sysbio/syq010

1481 129. Stover NA, Punia RS, Bowen MS, Dolins SB, Clark TG. Tetrahymena Genome
1482 Database Wiki: a community-maintained model organism database. *Database (Oxford)*.
1483 2012;2012: bas007. doi:10.1093/database/bas007

1484 130. Wang R-L, Miao W, Wang W, Xiong J, Liang A-H. EOGD: the Euplotes octocarinatus
1485 genome database. *BMC Genomics*. 2018;19: 63. doi:10.1186/s12864-018-4445-z

131. Arnaiz O, Meyer E, Sperling L. ParameciumDB 2019: integrating genomic data across the genus for functional and evolutionary biology. *Nucleic Acids Res.* 2020;48: D599–D605. doi:10.1093/nar/gkz948
132. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;24: 1586–1591. doi:10.1093/molbev/msm088
133. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics.* 2007;23: 2947–2948. doi:10.1093/bioinformatics/btm404
134. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49: D480–D489. doi:10.1093/nar/gkaa1100
135. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30: 1236–1240. doi:10.1093/bioinformatics/btu031
136. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49: D412–D419. doi:10.1093/nar/gkaa913
137. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol Biol Evol.* 2020;37: 291–294. doi:10.1093/molbev/msz189
138. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol.* 2020;37: 1530–1534. doi:10.1093/molbev/msaa015
139. Suzuki S. Parthenogenetic conjugation in *Blepharisma undulans japonicus* Suzuki. *Bull Yamagata Univ Natural Sci.* 1957;4: 69–84.

1510 140. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB
1511 v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes
1512 for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 2019;47:
1513 D807–D811. doi:10.1093/nar/gky1053

1514 141. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo
1515 genome assemblies. *Bioinformatics.* 2015;31: 3350–3352.
1516 doi:10.1093/bioinformatics/btv383

1517

Figure captions

Figure 1. *Blepharisma* cells and properties of ciliate MAC genomes. A. *B. stoltei* ATCC 30299 cell stained with anti-alpha tubulin-Alexa488 (depth-color coded red to yellow) and DAPI (cyan). **B.** Snapshot of a 3D reconstruction (Imaris, Bitplane) from CLSM images of Hoechst 33342 (dsDNA dye, Invitrogen™) fluorescence (Ex405 nm / Em420-470 nm). **C.** Scanning electron microscope (SEM) image of a *B. stoltei* ATCC 30299 cell (false color).

Figure 2. Nuclear development during *Blepharisma stoltei* conjugation. A. Cell development cycle with inset showing the four main kinds of nuclei, highlighted with arrows. Illustration of various cell stages adapted from [139]). Detailed staging collected for an RNA-seq time course is provided in Figure S9. **B.** Schematic of nuclear processes occurring during conjugation (classified according to, and modified from [9]). Nuclear events occurring before and up to, but not including fusion of the gametic nuclei (syngamy) are classified into sixteen pre-gamic stages where the MICs undergo meiosis and the haploid products of meiotic MICs are exchanged between the conjugating cells, followed by karyogamy. After karyogamy, cells are classified into 10 stages S (synkaryon), D1 (1st mitosis), I1 (1st interphase), D2 (2nd mitosis), I2 (2nd interphase), D3 (3rd mitosis), I3 (3rd interphase), D4 (4th mitosis), E1 (1st embryonic stage), E2 (2nd embryonic stage). After E2, the exconjugants divide further and are classified into 6 stages of cell division (CD1-6) which we did not follow here.

Figure 3. Basic properties of ciliate MAC genomes. In cell diagrams MACs are green and MICs are small black dots in close proximity to MACs. Citations for genome properties are in Data S1.

Figure 4. Alternative telomere addition sites. A. Mapping of a subset telomere-containing HiFi reads to a *B. stoltei* MAC genome contig region, with alternative telomere addition sites (ATASs) shown by blue (5') or mauve (3') arrows. Pink bars at read ends indicate soft-masking by minimap2, typically of telomeric repeats. **B.** Mean base frequencies in +/- 1 kbp flanking ATAS junctions.

Figure 5. A gene-dense somatic genome with tiny introns. A. HiFi (DNA) and RNA-seq coverage across a representative *B. stoltei* ATCC30299 MAC genome contig (Contig_1). Y scale is linear for HiFi reads and logarithmic (base 10) for RNA-seq. Plus strand (relative to the contig) RNA-seq coverage is green; minus strand RNA-seq coverage is blue. Between the RNA-seq coverage graphs each arrow represents a predicted gene, colored green when transcribed from the sense strand and blue from the antisense one. Two orthogroups classified by OrthoFinder are shown in orange and pink, respectively. **B.** Distribution of intron lengths from predicted genes. **C.** Sequence logos for 15 bp introns (splicing frequency > 0.5). **D.** Clustered orthogroups (Data S2) in the *B. stoltei* MAC genome.

Figure 6. MAC genome-encoded transposases in ciliates and properties of a putative *Blepharisma* IES excisase. A. Presence/absence matrix of PFAM transposase domains detected in predicted MAC genome-encoded ciliate proteins. Ciliate classes are indicated before the binomial species names. Note that *Ichthyophthirius multifiliis* is not shown because no domains with i-E-values < 1e-4 were detected. PiggyBac homolog phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*. **B.** DDE_Tnp_1_7 domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*. **C.** Cysteine-rich domains of PiggyBac homologs. PBLE transposases: Ago (*Aphis gossypii*); Bmo (*Bombyx mori*); Cag (*Ctenoplusia agnata*); Har (*Helicoverpa armigera*); Hvi (*Heliothis virescens*); PB-Tni (*Trichoplusia ni*); Mlu (PiggyBat from *Myotis lucifugus*); PLE-wu (*Spodoptera frugiperda*).

Domesticated PGBD transposases: Oni (*Oreochromis niloticus*); Pny (*Pundamilia nyererei*); Lia5, Tpb1, Tpb2, Tpb6 and Tpb7 (*Tetrahymena thermophila*); Pgm, PgmL1, PgmL2, PgmL3a/b/c, PgmL4a/b, PgmL5a/b (*Paramecium tetraurelia*); Tru (*Takifugu rubripes*); Pgbd2, Pgbd3 and Pgbd4 (*Homo sapiens*).

Figure 7. Phylogeny of ciliate PiggyBac homologs, eukaryotic PBLEs and PGBD5

homologs. Colors indicate eukaryotic lineages: purple=Alveolata, green=Stramenopiles, cyan=Archaeplastida, yellow=Opisthokonta, blue=Amoebozoa. Clade highlighted in pink represents all heterotrich Pgms and PgmLs. aBayes values generated by PhyML are represented as line widths and branch colors. aBayes values lower than 0.9 are represented in red-grey, with lower values appearing more red. The tree is rooted at the PiggyBac-like element of *Entamoeba invadens*.

Figure 8. DDE_1 and DDE_3 domain transposases in *Blepharisma*. A. DDE_1 domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*. **B.** DDE_3 domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*.

Figure 9. DDE_Tnp_IS1595 and MULE domain transposases in *Blepharisma*. A. DDE_Tnp_IS1595 domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*. **B.** MULE domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*.

Supplementary figure captions

Figure S1. Analysis of assembly completeness by Benchmarking Universal Single-Copy

Orthologs (BUSCO). Completeness of the *B. stoltei* ATCC 30299 MAC assembly was estimated by the percentage of BUSCOs found in the assembly with reference to the OrthoDB v10 alveolate database [140]. The nature of the ortholog-matches is indicated by characters followed by counts: C (complete orthologs) - light blue, D (duplicated orthologs) - dark blue, F (fragmented orthologs) - yellow and M (missing orthologs) - red.

Figure S2. Properties of minichromosomes, telomeres, and alternative telomere addition

sites. A. Length distribution of telomeres of telomere-bearing HiFi reads. **B.** Length distribution of HiFi reads delimited by telomeres. **C.** Diagram of a telomere-bearing read mapped onto genome reference at an ATAS. Sequence which is ambiguously chromosomal or telomeric is “junction sequence”; junction coordinate which maximizes telomere repeat length on the read is the “first identifiable breakpoint”; the coordinate maximizing alignment length to reference is the “last identifiable breakpoint”. The last telomeric unit permutation at the last identifiable breakpoint is underlined (length 8 bp). **D.** Sequence logos of chromosomal sequence at ATAS junctions, sorted by which permutation of the telomeric repeat is present (plot labels). Logos are aligned to the “last identifiable breakpoint” between positions 20 and 21; telomeric repeats on telomere-bearing reads begin to the left of the breakpoint. **E.** Frequencies of 2-mers in whole genome (blue), in telomeres (green), and at ATAS junctions (chromosomal side after last identifiable breakpoint, orange). **F.** Histogram of junction sequence lengths for ATASs in *B. stoltei*. **G.** Counts of each telomere repeat permutation at ATAS junctions (last identifiable breakpoint).

Figure S3. Femto Pulse analyses of *B. stoltei* MAC DNA and mapping of *T. thermophila* telomere-bearing reads. **A.** Mapping of PacBio CLR reads with 3 consecutive telomeric repeats to a representative *T. thermophila* MAC chromosome (Chr_001 from ciliate.org). **B.** Length distribution of input MAC DNA sizes prior to fragmentation and library preparation (Femto Pulse; LM = lower maker) - replicate 1. RFU=relative fluorescent units. **C.** Length distribution of input MAC DNA sizes prior to fragmentation and library preparation (Femto Pulse; LM = lower maker) - replicate 2.

Figure S4. POT1 paralog phylogeny, PFAM domain architecture, and gene expression in *Blepharisma*. Diagram elements as described in Figure 6B.

Figure S5. *Blepharisma stoltei* genetic code. Prediction for *B. stoltei* ATCC 30299 MAC genome by PORC; codons that are stops in the standard genetic code are highlighted in orange.

Figure S6. Intron splicing. **A.** Distribution of intron splicing fraction of candidate sense introns in the *B. stoltei* MAC genome. **B.** Distribution of intron splicing fractions of introns according to intron lengths. **C.** Distribution of intron splicing fraction of candidate antisense introns. **D.** Sequence logos for all predicted 16 nt introns, and 16 nt introns with “A” at either position -7 or -6 (counting from the 3’ end). The number of introns underlying the logos are indicated to the right. **E.** Distribution of intron splicing fractions of introns according to intron lengths. **F.** Sample of RNA-seq reads mapped to a GT-GG intron from gene BSTOLATCC_MAC21551 (Contig_57.g761). Translation in alternative reading frames downstream of the predicted intron leads to premature stop codons soon after the intron.

Figure S7. *B. stoltei* ATCC30299 MAC genome assembly graph. Bandage [141] representation of Flye 2.8.1 assembly graph. Edges corresponding to contigs are colored by

coverage (brightest pink = 160×, black=0×).

Figure S8. Experimental approach for conjugation RNA-seq time series. Complementary mating type strains of *Blepharisma stoltei* were harvested and cleaned by starving overnight. The cleaned cultures were treated in a time-staggered format, with gamones of the complementary mating type, where gamone 2 was a solution of the synthetic gamone 2 calcium salt and gamone 1 was provided as the cell-free fluid (CFF) harvested from mating-type I cells. Two sets of time-staggered gamone-treated cultures were used for the time series. Set I, indicated by the solid line, was mixed and used to observe and collect samples at 0 hours, 2 hours, 6 hours, 26 hours and 30 hours after mixing. Set II, indicated by the dashed lines, was mixed and used to observe and collect samples at 14 hours, 18 hours, 22 hours and 38 hours after mixing. Test tubes indicate Trizol samples prepared for RNA-extraction which were stored at -80 °C before processing. Cells collected for imaging were obtained shortly before the remainder were transferred into Trizol.

Figure S9. Developmental staging of *B. stoltei* for RNA-seq. A. The classification of nuclear morphology into stages has been done according to previous descriptions [9]. Nuclear events occurring before and up to, but not including fusion of the gametic nuclei (syngamy) are classified into sixteen stages indicated by roman numerals. These are the pre-gamic stages of conjugation where the MICs undergo meiosis and the haploid products of meiotic MICs are exchanged between the conjugating cells. Stages after syngamy are classified into 10 stages, stage S (synkaryon), D1 (1st mitosis), I1(1st interphase), D2 (2nd mitosis), I2 (2nd interphase), D3 (3rd mitosis), I3 (3rd interphase), D4 (4th mitosis), E1 (1st embryonic stage), E2 (2nd embryonic stage). After E2, the exconjugants divide further and are classified into 6 stages of cell division (CD1-6) which we did not follow here. Illustration of various cell stages adapted from [139]). Stacked bars show the proportion of cells at each time point at different stages of

development, preceded by the number of cells inspected (n).

Figure S10. Small RNA-related proteins in *Blepharisma*. **A.** ResIII, Helicase_c and Ribonuclease_3 domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*. **B.** PIWI domain phylogeny with PFAM domain architecture and gene expression heatmap for *B. stoltei*.

Figure S11. Histones and histone-domain-containing proteins in *Blepharisma*. Gene expression heatmaps are shown as in previous figures, are clustered according to major histone type as classified using HistoneDB domain models. Domains from PFAM and HistoneDB are shown to the right.

Figure 1.

Figure 2.

A.

B.

Figure 3.

Figure 4.

A.

B.

Figure 5.

Figure 8.

A. DDE_1 domain-containing proteins

B. DDE_3 domain-containing proteins

Figure 9

A. DDE_Tnp_IS1595 domain-containing proteins

B. MULE domain-containing proteins

Figure S1.

Figure S2.

Figure S3.

A.

C.

D.

Figure S4.

Figure S5.

Figure S6.

Figure S7.

Figure S8.

Figure S9.

Figure S10.

A. Dicers, Dicer-likes and Dicer-derivatives

B. Piwi proteins

Figure S11.