Comparative genomics reveals insight into the evolutionary origin of massively scrambled genomes.

Yi Feng[1], Rafik Neme[1,2], Leslie Y. Beh[1], Xiao Chen[3], Jasper Braun[4,5], Michael Lu[1], Laura F. Landweber[1*]

1. Departments of Biochemistry and Molecular Biophysics and Biological Sciences, Columbia University, New York, NY 10032, USA

2. Current affiliation: Department of Chemistry and Biology, Universidad del Norte, Barranquilla, Colombia

3. Pacific Biosciences, Menlo Park, CA 94025, USA

4. Department of Mathematics and Statistics, University of South Florida, Tampa, FL 33620, USA

5. Current affiliation: Division of Clinical Pathology, Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

*Author for Correspondence: Laura F. Landweber, Departments of Biochemistry and Molecular Biophysics and Biological Sciences, Columbia University, New York, NY, USA, 212-305-3898, Laura.Landweber@columbia.edu

Keywords: genome rearrangement, *de novo* genome assembly, transposable elements (TEs), ciliate, scrambled gene, comparative genomics, evolution

Abstract

Ciliates are microbial eukaryotes that undergo extensive programmed genome rearrangement, a natural genome editing process that converts long germline chromosomes into smaller gene-rich somatic chromosomes. Three well-studied ciliates include *Oxytricha trifallax*, *Tetrahymena thermophila* and *Paramecium tetraurelia*, but only the *Oxytricha* lineage has a massively scrambled genome, whose assembly during development requires hundreds of thousands of precise programmed DNA joining events, representing the most complex genome dynamics of any known organism. Here we study the emergence of such complex genomes by examining the origin and evolution of discontinuous and scrambled genes in the *Oxytricha* lineage.

This study compares six genomes from three species, the germline and somatic genomes for *Euplotes woodruffi*, *Tetmemena sp.*, and the model ciliate *Oxytricha trifallax*. To complement existing data, we sequenced, assembled and annotated the germline and somatic genomes of *Euplotes woodruffi,* which provides an outgroup, and the germline genome of *Tetmemena sp.*. We find that the germline genome of *Tetmemena* is as massively scrambled and interrupted as *Oxytricha*'s: 13.6% of its gene loci require programmed translocations and/or inversions, with some genes requiring hundreds of precise gene editing events during development. This study revealed that the earlier-diverged spirotrich, *E. woodruffi*, also has a scrambled genome, but only roughly half as many loci (7.3%) are scrambled. Furthermore, its scrambled genes are less complex, together supporting the position of *Euplotes* as a possible evolutionary intermediate in this lineage, in the process of accumulating complex evolutionary genome rearrangements, all of which require extensive repair to assemble functional coding regions. Comparative analysis also reveals that scrambled loci are often associated with local duplications, supporting a gradual model for the origin of complex, scrambled genomes via many small events of DNA duplication and decay.

Introduction

Organisms do not always contain a single, static genome. Programmed genome editing is a naturally occurring and essential part of development in many organisms, including ciliates (1), nematodes (2), lampreys (3) and zebra finches (4). Most of these events involve precise removal and rejoining of large regions of DNA during postzygotic differentiation of a somatic genome from a germline genome. Ciliates are microbial eukaryotes with two types of nuclei: a somatic macronucleus (MAC) that differentiates from a germline micronucleus (MIC). In the model ciliate *Oxytricha*, the MAC is entirely active chromatin (5), and the hub of transcription. The three species that we compare are all spirotrichs, which have gene-sized "nanochromosomes" in the MAC, present at high copy number (6, 7, 8, 9, 10, 11). The diploid MIC participates in sexual reproduction, but its megabase-sized chromosomes are mostly transcriptionally silent.

Gene loci are often arranged discontinuously in the MIC, with short genic segments called Macronuclear Destined Sequences (MDSs), interrupted by stretches of non-coding DNA called Internally Eliminated Sequences (IESs) (Figure 1A). During sexual development, a new MAC genome rearranges from a copy of the zygotic MIC genome. MDSs join in the correct order and orientation, whereas MIC-limited genomic regions undergo programmed deletion, including repetitive elements, intergenic regions and IESs (Figure 1A). Though analogous to intron splicing, these events occur on DNA. The MDSs for some MAC chromosomes are *scrambled* if they require translocation or inversion during MAC development (Figure 1A). Pairs of short repeats, called *pointers*, are present at MDS-IES junctions in both scrambled and nonscrambled loci (12, 13). Pointer sequences are present twice in the MIC, at the end of MDS *n* and the beginning of MDS *n+1*. One copy of the repeat is present at each MDS-MDS junction in a mature MAC chromosome (Figure 1A). These microhomologous regions help guide MDS

recombination, but most are non-unique and the shortest pointers are just 2 bp. Thousands of long, noncoding template RNAs collectively program MDS joining (14, 15, 16).

Numerous studies have inferred the possible scope of genome rearrangement in different ciliate species using partial genome surveys. In *Paramecium*, PiggyMac-depleted cells fail to remove MIC-limited regions properly, which provided a resource to annotate ~45,000 IESs prior to assembly of a draft MIC genome (17). The use of single-cell sequencing has allowed pilot studies to sample partial MIC genomes of diverse species (9, 18, 19, 20). Alignment of tentative MIC reads to either assembled MAC genomes or single-cell transcriptome data predicts over 20 candidate scrambled loci in two basal ciliates, *Loxodes sp*. and *Blepharisma americanum* (19) and hundreds of candidate loci in the tintinnid *Schmidingerella arcuata* (20). Nearly one third (31%) of approximately 5,000 surveyed transcripts may be scrambled in *Chilodonella uncinata* (18, Figure 1B), which has four confirmed cases of scrambled genes (21, 22). Transcriptome-based surveys offer less precise estimates, and cannot distinguish RNA splicing. Several computational pipelines have been developed to facilitate the inference of genome rearrangement features by split-read mapping in the absence of complete MIC or MAC reference genomes (23, 24, 25, 26). By surveying lighter genome coverage prior to full sequencing, these tools provide partial insight into germline architecture. This helps guide selection of species for full genome sequencing and subsequent construction of complete rearrangement maps between the MIC and MAC genomes. High-quality MIC genome reference assemblies are only currently available for three ciliate genera: *Oxytricha* (1), *Tetrahymena* (27) and *Paramecium* (28, 29).

Programmed genome rearrangements in *Oxytricha* exhibit the highest accuracy and largest scale of any known natural gene-editing system, with exquisite control over hundreds of thousands of precise DNA cleavage/joining events. Accordingly, its germline genome structure

94 is arguably the most complex of any model organism (1), requiring programmed deletion of over
95 90% of the germline DNA during development and massive descrambling of the resulting
96 fragments to construct a new MAC genome of over 18,000 chromosomes (10). This differs from
97 the distantly related *Tetrahymena* and *Paramecium* that both eliminate ~30% of the germline
98 genome (27, 28). *Paramecium* uses exclusively 2 bp pointers and lacks evidence of any
99 scrambled loci. A small number of scrambled loci (4 confirmed out of 2711 candidates) have
100 been reported in *Tetrahymena* (30, Figure 1B). *Tetrahymena and Paramecium* diverged from
101 *Oxytricha* over one billion years ago (31, 32), which leaves a large gap in our understanding of
102 the emergence of complex DNA rearrangements in the *Oxytricha* lineage.
103 Open questions include how did the *Oxytricha* germline genome acquire its high number
104 of IES insertions and how do scrambled loci arise and evolve. Three previous studies tackled
105 these questions at the level of single genes and orthologs, including DNA polymerase α, actin I
106 and TEBPα (33, 34, 35, 36). Here, we provide the first comparative genomic analysis of
107 *Oxytricha trifallax* and two other spirotrichous ciliates, *Tetmemena sp.* and *Euplotes woodruffi*.
108 *Tetmemena sp.* is a hypotrich similar to *Tetmemena pustulata*, formerly *Stylonychia pustulata*
109 (7), in the same family as *Oxytricha trifallax* (Figure 1B, 1, 7). Hypotrichs are noted for the
110 presence of scrambled genes, based on previous ortholog comparisons (7, 33, 34, 35, 36 Figure
111 1B). *E. woodruffi,* together with the hypotrichous ciliates, belong to the class Spirotrichea
112 (Figure 1B). Like hypotrichs, *Euplotes* also has gene-sized nanochromosomes in the MAC
113 genome (8, 9, 37), but this outgroup uses a different genetic code (UGA is reassigned to cysteine,
114 ref. 38) and little is known about its MIC genome. A partial MIC genome of *Euplotes vannus*
115 was previously assembled, and it contains highly conserved TA pointers (9), consistent with
116 previous observations in *Euplotes crassus* (39). This differs from *Oxytricha trifallax*, which uses

longer pointers of varying lengths, with scrambled pointers typically longer than nonscrambled ones (1, Figure 1B). This observation suggests that longer pointers may supply more information to facilitate MDS descrambling, sometimes over great distances. Therefore, the preponderance of 2 bp pointers in the other *Euplotes* species could indicate limited capacity to support scrambled genes, and a partial genome survey of *E. vannus* concluded that at least 97% of loci are nonscrambled (9). Early studies of *Euplotes octocarinatus*, on the other hand, demonstrated its use of longer pointers (that usually contain TA) (40, 41), suggesting that some members of the *Euplotes* genus may have the capacity to support complex genome reorganization. To investigate the origin of scrambled genomes, we choose *E. woodruffi* as an outgroup, because it is closely related to *E. octocarinatus* (42) and feasible to culture in the lab.

This study includes the *de novo* assemblies of the micronuclear genome of *Tetmemena sp.* and both genomes of *E. woodruffi*. The availability of MIC and MAC genomes for both species allows us to annotate and compare their genome rearrangement maps and other key features to each other and to *O. trifallax*. The MIC genome of *Tetmemena* is extremely interrupted, like *Oxytricha*. While the *E. woodruffi* MIC genome is much more IES-sparse, it contains thousands of scrambled genes, whose architecture we compare to orthologous loci in the other species. We infer that the evolutionary origin of scrambled genes is associated with local duplications, providing strong support for a previously proposed simple evolutionary model requiring only duplication and decay (43) that allows for the evolutionary expansion of extremely rearranged chromosome architectures.

Results

Germline genome expansion via repetitive elements

Tetmemena sp. and *E. woodruffi* were both propagated in laboratory culture from single cells. The *E. woodruffi* MAC genome was sequenced and assembled from paired-end Illumina reads from whole cell DNA, which is mostly MAC-derived. For comparative analysis, the MAC genome of *E. woodruffi* was assembled using the same pipeline previously used for *Tetmemena sp.* (7). Because MIC DNA is significantly more sparse than MAC DNA in individual cells (13) MIC DNA was enriched before sequencing (see Methods); however, this leads to much lower sequence coverage of the MIC than the MAC. Third-generation long reads (Pacific Biosciences and Oxford Nanopore Technologies) were combined with Illumina paired-end reads (Methods, see genome coverage in Supplementary File 1) to construct hybrid genome assemblies for *Tetmemena sp.* and *E. woodruffi*. Though the final genome assemblies are still fragmented, often due to transposon or other repetitive insertions at boundaries (Figure 2 - figure supplement 1), the current draft assemblies cover most (>90%) MDSs for 89.1% of MAC nanochromosomes in *Tetmemena*, and for 90.0% of MAC nanochromosomes in *E. woodruffi*. This allowed us to establish near-complete rearrangement maps for the newly assembled genomes of *Tetmemena* and *E. woodruffi,* at a level comparable to the published reference for *O. trifallax* (1), which is appropriate for comparative analysis.

Table 1 shows a comparison of genome features for the three species. The three MAC genomes are similar in size, with most nanochromosomes bearing only one gene. The size distributions of MAC chromosomes are similar for the three species, though slightly shorter for *E. woodruffi*, consistent with prior observation via gel electrophoresis (13, Figure 2 - figure supplement 2). Like *O. trifallax* (6), the maximum number of genes encoded on one

chromosome is 7-8 (Table 1). Surprisingly, the MIC genome sizes differ substantially: the *Tetmemena* MIC genome assembly is 237 Mbp, nearly half that of *Oxytricha*. The *E. woodruffi* MIC genome assembly is even smaller, approximately 172 Mbp (Table 1).

The expansion of repetitive elements in the *Oxytricha* lineage may contribute to the difference in MIC genomes sizes (Figure 2A-C). *Oxytricha* has a variety of tranposable elements (TEs) in the MIC, with telomere-bearing elements (TBEs) of the Tc1/*mariner* family the most abundant (1, 44, Supplementary File 2). A complete TBE transposon contains three open reading frames (ORFs). ORF1 encodes a 42kD transposase with a DDE-catalytic motif. Though present only in the germline, TBEs are so abundant in hypotrichs that some were partially recovered and assembled from whole cell DNA (44). The *Oxytricha* MIC genome contains ~10,000 complete TBEs and ~24,000 partial TBEs, which occupy approximately 15.20% (75 Mbp) of the genome (Figure 2A, Supplementary File 3, 1, 44). *Tetmemena*, on the other hand, has many fewer TBE ORFs and only 48 complete TBEs (Supplementary File 3), comprising 1.83% (4.3 Mbp) of its MIC genome (Figure 2B). *Euplotes crassus* has also been reported to have an abundant transposon family called Tec elements (Transposon of *Euplotes crassus*). Like TBEs, each Tec consists of three ORFs, and ORF1 also encodes a transposase from the Tc1/*mariner* family (45, 46, 47, 48, 49). The ~57kD ORF2 encodes a tyrosine-type recombinase (50) and the 20kD ORF3 has unknown function (47). Using the three ORFs of Tec1 and Tec2 as queries for search, we identified 74 complete Tec elements in *E. woodruffi*. Collectively, Tec ORFs occupy 3.6 Mbp, corresponding to only 2.1% of the MIC genome (Figure 2C). Notably, the transposase-encoding ORF1 is more abundant than the other two TBE/Tec ORFs in all three ciliates (Supplementary File 3), consistent with its proposed role in DNA cleavage during genome rearrangement in *Oxytricha* (51).

184 *Oxytricha* contains three families of TBEs. TBE3 appears to be the most ancient among
185 hypotrichs, based on previous analysis of limited MIC genome data (44). We constructed
186 phylogenetic trees using randomly subsampled TBE sequences for all three ORFs from
187 *Oxytricha* and *Tetmemena* (Figure 2D-F). This confirmed that only TBE3 is present in the
188 *Tetmemena* MIC genome, as proposed in (44). This also suggests that TBE1 and TBE2 expanded
189 in *Oxytricha* after its divergence from other hypotrichous ciliates. As illustrated in Figure 2 -
190 figure supplement 1, the MIC genome contexts of TBEs in *Oxytricha* and *Tetmemena* are
191 similar, with many TE insertions within IESs, consistent with either IESs as hotspots for TE
192 insertion or with the model (49) that some TE insertions may have generated IESs, as
193 demonstrated in *Paramecium* (29, 52). Subsequent sequence evolution at the edges of IES/MDS
194 pointers (36) can give rise to boundaries that no longer correspond precisely to TBE ends. For
195 further discussion of the conservation of TBE locations, see the section, "*Oxytricha* and
196 *Tetmemena* share conserved rearrangement junctions" below.

197 Additionally, Repeatmodeler/Repeatmasker identified that *Oxytricha* has more MIC
198 repeats in the "Other" category than *Tetmemena* or *E. woodruffi* (Figure 2, subcategories of
199 repeat content in Supplementary File 2). 214 Mbp of the *Oxytricha* MIC genome (43%, which is
200 greater than 35.9% reported in ref. 1 that used earlier versions of the software) is considered
201 repetitive (including TBEs, tandem repeats and other repeats in Figure 2), versus 31.7 Mbp for
202 *Tetmemena* (13.4%) and 28.5 Mbp (16.8%) for *E. woodruffi*. *Oxytricha*'s additional ~180 Mbp
203 in repeat content partially explains the significantly larger MIC genome size of *Oxytricha* versus
204 the other spirotrich ciliates.

The *E. woodruffi* genome has fewer IESs

We used the genome rearrangement annotation tool SDRAP (53) to annotate the MIC genomes of *Oxytricha, Tetmemena* and *E. woodruffi* (Methods). Consistent with their close genetic distance, the genomes of *O. trifallax* and *Tetmemena* have similarly high levels of discontinuity (Figure 3A): We annotated over 215,299 MDSs in *Oxytricha* and over 215,624 in *Tetmemena* MDSs with similar MDS length distributions (Figure 3A). By contrast, *E. woodruffi* MDSs are typically longer, which indicates a less interrupted genome (Figure 3A). We compared the number of MDSs between single-copy orthologs for single-gene MAC chromosomes across the three species and found that the orthologs have similar CDS lengths (Figure 3 - figure supplement 1A-B). There is a strong positive correlation between number of MDSs for orthologous genes in *Oxytricha* and *Tetmemena* ($R^2 = 0.75$, Figure 3B). There is no correlation among number of MDSs between orthologs of *E. woodruffi* and *Oxytricha* ($R^2 = 0.003$, Figure 3C), since *E. woodruffi* orthologs typically contain fewer MDSs.

The *E. woodruffi* genome is generally much less interrupted than that of *Oxytricha* or *Tetmemena*. 39.9% of MAC nanochromosomes in *E. woodruffi* lack IESs (IES-less nanochromosomes) compared to only 4.1% and 4.4% in *Oxytricha* and *Tetmemena*, respectively. The sparse IES distribution (as measured by plotting pointer distributions) in *E. woodruffi* displays a curious 5' end bias on single-gene MAC chromosomes, oriented in gene direction (Figure 3E). A weak 5' bias is also present in *Oxytricha* (Figure 3D) and *Tetmemena* (Figure 3 - figure supplement 1C). In addition, *E. woodruffi* IESs preferentially accumulate in the 5' UTR, a short distance upstream of start codons (Figure 3F). Notably, the median distance between the 5' telomere and start codon in *E. woodruffi* is just 54 bp for single-gene chromosomes.

E. woodruffi has an intermediate level of genome scrambling

Scrambled genome rearrangements exist in all three species, which we report here for the first time in *Tetmemena* and the early-diverged *E. woodruffi*. Previous studies have described scrambled genes with confirmed MIC-MAC rearrangement maps for a limited species of hypotrichs (1, 7, 33, 34, 35, 36) and *Chilodonella* (21, 22), but not in *Euplotes*. Consistent with the phylogenetic placement of *Euplotes* as an earlier-diverged outgroup to hypotrichs (54, 55), the *E. woodruffi* genome is scrambled, but it contains approximately half as many scrambled genes (2429 genes encoded on 1913 chromosomes, or 7.3% of genes), versus 15.6% scrambled in *Oxytricha trifallax* (3613 genes encoded on 2852 chromosomes) and 13.6% in *Tetmemena* (3371 genes encoded on 2556 chromosomes). The *E. woodruffi* lineage may therefore reflect an evolutionary intermediate stage between ancestral genomes with only modest levels of genome scrambling versus the more massively scrambled genomes of hypotrichs.

We infer that many genes were likely scrambled in the last common ancestor of *Oxytricha* and *Tetmemena*, because these two species share approximately half of their scrambled genes (Supplementary File 4). Furthermore, most scrambled genes are not new genes, since they possess at least one ortholog in other ciliate species (Supplementary File 4, Supplementary File 5).

Scrambled genes are associated with local paralogy

Notably, scrambled genes in all three species generally have more paralogs (Figure 4). We identified orthogroups containing genes derived from the same gene in the last common ancestor of the three species (Methods). For each species, orthogroups with at least one scrambled gene are significantly larger than those containing no scrambled genes (p-value <1e-5,

Mann-Whitney U test, Figure 4A-C). This association suggests a possible role of gene duplication in the origin of scrambled genes.

Scrambled pointers are generally longer than nonscrambled ones in all three species (Figure 3 - figure supplement 2), consistent with prior observations (1) and the possibility that longer pointers participate in more complex rearrangements, including recombination between MDSs separated by greater distances (56). Scrambled and nonscrambled IESs also differ in their length distribution (Figure 3 - figure supplement 2). Notably, scrambled "pointers" in *E. woodruffi* can be as long as several hundred base pairs (median 48 bp, average 212 bp) unlike the more typical 2-20 bp canonical pointers. These long "pointers" in *E. woodruffi* are more likely partial MDS duplications (Figure 4 - figure supplement 1A). We also identified MDSs that map to two or more paralogous regions within the same MIC contig (Supplementary File 6), therefore representing MDS duplications and not alleles. Such paralogous regions could be alternatively incorporated into the rearranged MAC product. Notably, we find that, for all three species, there are significantly more scrambled than nonscrambled MAC chromosomes that contain at least one paralogous MDS (chi-square test, p-value <1e-10; Supplementary File 6). An example is shown in Figure 4 - figure supplement 1A (MDS 7 and 7').

The presence of paralogous MDSs can contribute to the origin of scrambled rearrangements, as proposed in an elegant model by Gao et al. (ref.43; illustrated in Figure 4 - figure supplement 1B). The model proposes that initial MDS duplications permit alternative use of either MDS copy into the mature MAC chromosome. As mutations accumulate in redundant paralogs, cells that incorporate the least decayed MDS regions into the MAC gene would have both a fitness advantage and a better match to the template RNA (14) that guides rearrangement, thus increasing the likelihood of incorporation into the MAC chromosome. The paralogous

regions containing more mutations would gradually decay into IESs and scrambled pointers eventually reduced to a shorter length. The extended length "pointers" that we identified in *E. woodruffi* may reflect an intermediate stage in the origin of scrambled genes (Figure 4 - figure supplement 1B).

This model may generally explain the abundance and expansion of "odd-even" patterns in ciliate scrambled genes (56, 57). As illustrated in Figure 4 - figure supplement 1A, the even- and odd-numbered MDSs for many scrambled genes derive from different MIC genome clusters. The model predicts that the IES between MDS *n-1* and *n+1* often derives from ancestral duplication of a region containing MDS *n* (Figure 4 - figure supplement 2A). To test this hypothesis explicitly, we extracted from all odd-even scrambled loci in the three species all sets of corresponding MDS/IES pairs that are flanked by identical pointers on both sides, i.e. all pairs of scrambled MDSs and IESs, where the IES between MDS *n-1* and *n+1* is directly exchanged for MDS *n* during DNA rearrangement (S1 and S2 in Figure 4 - figure supplement 2A). To exclude the possibility of alleles confounding this analysis, MDS and IES pairs were only considered if they map to the same MIC contig. In *E. woodruffi*, the lengths of these MDS/IES pairs strongly correlate (Spearman correlation ρ=0.755, p<1e-5, Figure 4 - figure supplement 2B). Moreover, many MDS and IES sequence pairs also share sequence similarity, consistent with paralogy: For 248 MDS-IES pairs of similar length, 90.3% share a core sequence with ~97.5% identity across 8-100% of both the IES and MDS length. The lowest end of these observations is also compatible with an alternative model (34) in which direct recombination between IESs and MDSs at short repeats can lead to expansion of odd-even patterns. For *Oxytricha* and *Tetmemena*, the MDS and IES lengths for such MDS/IES pairs also display a weakly-positive correlation (*p-values* and Spearman correlation ρ shown in Figure 4 - figure

supplement 2D-E). Remarkably, the odd-even-containing loci that are species-specific, and therefore became scrambled more recently, have the strongest length correlation (Figure 4 - figure supplement 2C-E) and more pairs that display sequence similarity (Supplementary File 7) relative to older loci (scrambled in two or more species). This result is consistent with an evolutionary process in which mutations accumulate in one copy of the MDS, gradually obscuring its sequence homology and ability to be incorporated as a functional MDS, and eventually its ability to be recognized by the template RNAs that guide DNA rearrangement. This analysis also suggests that most of the odd-even scrambled loci in *E. woodruffi* arose recently, because there is greater sequence similarity between MDSs and the corresponding IESs that they replace. Conversely, we infer that most loci that are scrambled in both *Oxytricha* and *Tetmemena* became scrambled earlier in evolution, since they display weaker sequence similarity between exchanged MDS and IES regions.

Scrambled and nonscrambled genes display nearly identical expression support (the presence of at least one read in all 3 replicates) in both *Oxytricha* (Supplementary File 8) and *Tetmemena*. *E. woodruffi* has slightly more expression support for nonscrambled vs. scrambled genes (Figure 4 - figure supplement 3), which could be explained by more recent acquisition of thousands of scrambled loci in *E. woodruffi*. In some of those cases the nonscrambled paralogs may still contribute the major function. The distribution of expression levels is similar for scrambled vs. nonscrambled genes in all three species, supporting their authenticity (Figure 4 - figure supplement 3), although in a Mann-Whitney U test, the average expression level of three replicates is significantly higher in nonscrambled genes for *Oxytricha* and *E. woodruffi*, but not significant for *Tetmemena*.

Oxytricha and *Tetmemena* share conserved DNA rearrangement junctions

To understand the conservation of genome rearrangement patterns, we developed a pipeline guided by protein sequence alignment to compare pointer positions for orthologous genes between any two species (Methods, Figure 5A). We compared pointers for 2503 three-species single-copy orthologs. 4448 pointer locations are conserved between *Oxytricha* and *Tetmemena* on 1345 ortholog pairs (Supplementary File 9), representing 38.3% of pointers in these orthologs in *Oxytricha* and 30.9% in *Tetmemena*. For *Oxytricha*/*E. woodruffi* and *Tetmemena*/*E. woodruffi* comparisons, 56 and 58 pointer pairs are conserved, respectively. We also identified 23 pointer locations shared among all three species (Supplementary File 9, Figure 5B, Figure 5 - source data 1).

To test if these pointer locations are genuinely conserved versus coincidental matching by chance, we performed a Monte Carlo simulation, as also used to study intron conservation (58). We randomly shuffled pointer positions on CDSs 1000 times, and counted the number of conserved pointer pairs expected for each simulation (Methods). Of the 1000 simulations, none exceeded the observed number of conserved pointer pairs between *Oxytricha* and *Tetmemena* (p-value < 0.001), suggesting evolutionary conservation of pointer positions (Supplementary File 9). A similar result was obtained for pointers conserved in all three species (Supplementary File 9). However, the numbers of pointer pairs conserved between *Oxytricha*/*E. woodruffi* and *Tetmemena*/*E. woodruffi* are similar to the expectations by chance (Supplementary File 9). The low level of pointer conservation of either hypotrich with *E. woodruffi* may reflect the smaller number of IESs in *E. woodruffi*; hence most pointers would have arisen in the hypotrich lineage. Furthermore, *E. woodruffi* is genetically more distant from the two hypotrichs; hence the accumulation of substitutions would obscure protein sequence homology, which we used to

compare pointer locations. For ortholog pairs between *Oxytricha* and *Tetmemena*, scrambled pointers are significantly more conserved than nonscrambled ones (chi-square test, *p*-value <1e-10, Supplementary File 10). We also find that most pointer sequences differ even if the positions are conserved (Figure 5B, Figure 5 - source data 1, Supplementary File 11), suggesting that substitutions may accumulate in pointers without substantially altering rearrangement boundaries.

Oxytricha and *Tetmemena* both contain a high copy number of TBE transposons (1, 44, Supplementary File 3). We investigated the level of TBE conservation between these two species. To identify orthologous insertions, we focus on TBE insertions in nonscrambled IESs on single-copy orthologs, which include 1706 *Oxytricha* TBEs inserted in 1296 nonscrambled IESs (multiple TBEs can be inserted into an IES) and 180 *Tetmemena* TBEs inserted into 170 nonscrambled IESs. We refer to the pointer flanking a TBE-containing IES as a *TBE pointer*. No TBE pointer locations are conserved between two species. This suggests that TBEs might invade the genomes of *Oxytricha* and *Tetmemena* independently, or still be actively mobile in the genome. Only 27 *Oxytricha* TBE pointers (containing 36 TBEs) are conserved with non-TBE pointers in *Tetmemena* (Figure 5 - source data 2, Figure 5C). No *Tetmemena* TBE pointer is conserved with an *Oxytricha* non-TBE pointer. This suggests that TBE insertions may preferentially produce new rearrangement junctions instead of inserting into an existing IES.

Intron locations sometimes coincide with DNA rearrangement junctions

Ciliate genomes are generally intron-poor. *Oxytricha* averages 1.7 introns/gene, *Tetmemena* has 1.1, and *E. woodruffi* has 2.2. Among three-species orthologs, intron locations sometimes map near pointer positions (within a 20 bp window, Figure 5B, Figure 5 - figure

supplement 1). IESs and introns are both non-coding sequences removed from mature transcripts, though at different stages. A previous single-gene study observed that an IES in *Paraurostyla* overlaps the position of an intron in *Uroleptus*, *Urostyla* and also the human homolog (34). This observation suggested an intron-IES conversion model in which the ability to eliminate non-coding sequences as either DNA or RNA provides a potential backup mechanism. Such interconversion has also been observed between two strains of *Stylonychia* (59). In the present study, we identified 174 potential cases of intron-IES conversion in the three species (Figure 5 - figure supplement 1, Supplementary File 12): 103 (59.2%) *E. woodruffi* introns map near *Oxytricha*/*Tetmemena* pointers. We used a 20 bp window for this analysis, since one would only expect the boundaries of introns and IESs to coincide precisely if they were recent evolutionary conversions. A Monte Carlo simulation for these intron-IES comparisons (Supplementary File 12) revealed that $p < 0.001$ for most three-species comparisons. For two-species comparisons, we identify 306 cases where an intron boundary in one species precisely coincides with a pointer sequence in another species, with strongest statistical support for the comparison between *Oxytricha* intron positions vs. *Tetmemena* IES junctions ($p = 0.008$) (Supplementary File 13). Notably, *Tetmemena* intron locations rarely coincide with *Oxytricha* IESs (Supplementary File 13), suggesting a possible bias in the direction of intron-IES conversion during evolution.

The observation that *E. woodruffi* has the most introns but the smallest number of IESs per gene (Figure 3) is consistent with removal of intragenic non-coding sequences as either DNA or RNA. The intron-sparseness of ciliates is compatible with a hypothesis that it is advantageous to eliminate non-coding sequences earlier at the DNA level, with intron deletion sometimes providing an opportunity for repair if they fail to be excised as IESs (34).

Evolution of complex genome rearrangements: Russian doll genes

Genome rearrangements in the *Oxytricha* lineage can include overlapping and nested loci, with MDSs for different MAC loci embedded in each other (1, 60). When multiple gene loci are nested in each other, these have been called Russian doll loci (60). *Oxytricha* contains two loci with five or more layers of nested genes (60). *Oxytricha* and *Tetmemena* display a high degree of synteny and conservation in both Russian doll loci. In the first Russian doll gene cluster, one nested gene (green) is present in *Oxytricha* but absent in *Tetmemena* (Figure 6A, Figure 6 - figure supplement 1, Figure 6 - figure supplement 2), confirmed by PCR (Methods). *Oxytricha* also has a complete TBE3 insertion in the green gene (Figure 6A, Figure 6 - figure supplement 1A), hinting at a possible link between transposon and new gene insertion. In addition, a two-gene chromosome in *Oxytricha* (orange) is present as two single-gene chromosomes in *Tetmemena* (Figure 6A, Figure 6 - figure supplement 1). In *Oxytricha*, seven orange MDSs ligate across two loci via an 18 bp pointer (TATATCTATACTAAACTT) to form a 2-gene nanochromosome. However, in *Tetmemena*, telomeres are added to the ends of both gene loci instead, forming two independent MAC chromosomes (Figure 6A, Figure 6 - figure supplement 1). The second Russian doll locus has an example of a long, conserved pointer (orange dotted line) that bridges three other loci (the green and blue scrambled loci and one nonscrambled locus, Figure 6B). Close to this region is a decayed TBE insertion (769bp) in *Oxytricha.* None of the *E. woodruffi* orthologs of both Russian doll loci map to the same MIC contig, which suggests that the Russian doll cluster arose after the divergence of *Euplotes* from the common ancestor of *Oxytricha* and *Tetmemena.*

Discussion

The highly diverse ciliate clade provides a valuable resource for evolutionary studies of genome rearrangement. However, full assembly and annotation of germline MIC genomes has concentrated on the model ciliates *Tetrahymena, Paramecium* and *Oxytricha*. To provide insight into genome evolution in this lineage, we assembled and compared germline and somatic genomes of *Tetmemena sp*. and an outgroup, *Euplotes woodruffi*, to that of *Oxytricha trifallax*. This expands our knowledge of the diversity of ciliate genome structures and the evolutionary origin of complex genome rearrangements.

Dramatic variation in transposon copy number (TBE and Tec elements) from the Tc1/*mariner* family appears to explain most of the variation in MIC genome size. In many eukaryotic taxa, genome size can differ dramatically even for closely related species, a phenomenon known as the "C-value paradox" (61). Our present observations are compatible with previous reports that the repeat content of the genome, especially transposon content, positively correlates with genome size (62).

Oxytricha has three TBE families in the MIC genome, but only TBE3 is present in *Tetmemena*, consistent with our previous conclusion that TBE3 is ancestral to the base of the transposon lineage in hypotrichous ciliates (44). Tens of thousands of TBE1/2 transposons then expanded specifically in *Oxytricha*. Despite a high copy number of TBEs in both *Oxytricha* and *Tetmemena*, we find no identical TBE locations in nonscrambled IESs, even among syntenic Russian doll regions. These observations suggest that TBEs may be active in these genomes and contribute to the evolution of genome structure.

In the relatively IES-poor genome of *E. woodruffi*, IESs accumulate upstream of start codons, similar to the 5' bias of introns in intron-poor organisms (63). The simplest model to explain 5' intron bias is homologous recombination between a reverse transcript of an intron-lacking mRNA and the original DNA locus to erase introns in the coding region (63). A similar mechanism could simultaneously erase IESs in coding regions via germline recombination between the MIC chromosome and a reverse transcript; however, they are usually in different subcellular locations. More plausibly, a source for DNA recombination could be a MAC nanochromosome, since they are already abundant at high copy number, but another source could be by capture of a reverse transcript of a long non-coding template RNA that guides DNA rearrangement (14, 15). Either recombination event in the germline would lead to loss of IESs, while retaining introns, but neither would necessarily provide a bias for IES-loss in coding regions. Any of these infrequent events would be meaningful on an evolutionary time scale, even if developmentally rare. The 5' bias of IESs could also reflect an evolutionary bias for continuous coding regions. Alternatively, upstream IESs might regulate gene expression or cell growth (29), like some introns (64, 65).

This study investigated the evolution of scrambled genes by comparing *Oxytricha* and *Tetmemena* to *E. woodruffi*, as an earlier-diverged representative of the spirotrich lineage. While *E. woodruffi* has approximately half as many scrambled genes as *Tetmemena* and *Oxytricha*, its genes are also much more continuous. For example, the most scrambled gene in *E. woodruffi,* encoding a DNA replication licensing factor (EUPWOO_MAC_28518, 3kb), has only 20 scrambled junctions. The most scrambled gene in *Tetmemena* (LASU02015934.1, 14.7kb, encoding a hydrocephalus-inducing-like protein) has 204 scrambled pointers, and the most scrambled gene in *Oxytricha* (Contig17454.0, 13.7kb, encoding a dynein heavy chain family

451 protein, ref. 1) is similarly complex, with 195 scrambled junctions. Together, these observations
452 are consistent with our interpretation that *E. woodruffi* reflects an evolutionary intermediate
453 stage, as it contains both fewer scrambled loci and fewer scrambled junctions within its
454 scrambled loci. The observation that the most scrambled locus differs in each species is also
455 consistent with the conclusion that complex gene architectures may continue to elaborate
456 independently.

457 We observed that scrambled genes in each species tend to have more paralogs than
458 nonscrambled genes. Similarly, in *Chilodonella uncinata* (18), a distantly-related ciliate in the
459 class *Phyllopharyngea* that also has scrambled genes, scrambled gene families (orthogroups)
460 contain more genes (~2.9) than nonscrambled gene families (~1.3) (18). Apart from duplications
461 at the gene level, *E. woodruffi* often contains partial MDS duplications at scrambled junctions,
462 annotated as unusually long "pointers" (Figure 4 - figure supplement 1). We also demonstrate
463 that odd-even scrambled patterns could readily arise from local duplications (Figure 4 - figure
464 supplement 2). These observations are most consistent with a simple model (43) in which local
465 duplications permit combinatorial DNA recombination between paralogous germline regions,
466 and mutation accumulation in either paralog establishes an odd-even scrambled pattern that can
467 propagate by weaving together segments from paralogous sources. Other proposed models
468 include Hoffman and Prescott's (66) IES-invasion model that suggested that pairs of IESs could
469 invade an MDS, and then subsequently recombine with another IES to yield odd-even scrambled
470 regions; however, a previous examination did not find support for this model (34). Prescott *et al.*
471 (67) also proposed that some odd-even scrambled loci could arise suddenly via reciprocal
472 recombination with loops of A/T-rich DNA, but this does not exploit paralogy, only the high A/T
473 content in the MIC. We previously proposed a gradual model (34, 56) in which MDS/IES

recombination at short AT-rich repeats (precursors to pointers) could generate and propagate odd-even scrambled patterns . While limited comparisons of orthologs favored the stepwise recombination models (33, 34, 35, 36), none of the earlier models accounted for the widespread existence of partial paralogy, revealed by genome assemblies.

Local duplications provide a buffer against mutations, allowing paralogous MDSs to repair the MAC locus during assembly of odd/even scrambled genes. Therefore, once an odd/even scrambled locus is established, a consequence is that evolution can only proceed in the direction of accumulating more scrambled junctions, as each new mutation in one paralog necessitates repair via incorporation of the other paralog (Figure 4 - figure supplement 1B). This shortens the length of the respective MDSs and increases the number of recombination junctions, creating an evolutionary ratchet that drives the increase in scrambling. The lack of the presence of an error-free, continuous version of this locus in the germline reduces the possibility of losing the scrambled pattern from the MIC genome, relative to the trend towards decreasing MDS lengths as more mutations accumulate in either paralog, with a resulting increase in the levels of scrambling and fragmentation (68, 69). The only opportunity to repair a scrambled locus in the MIC would be a rare event that replaces the locus via recombination with a continuous version from the parental MAC, with the source being either parental MAC DNA or a reverse transcript of a template RNA (14, 15), as discussed above.

Recent exciting reports have also described scrambled genomes in metazoa, including cephalopods (70, 71), but those events entail primarily evolutionary shuffling of gene order, without accompanying genome editing or repair. The ciliate lineage is remarkable in having evolved a sophisticated mechanism of RNA-guided genome editing that allows accurate and

precise DNA repair of translocations and inversions. The future opportunity to harness this system to develop novel tools for genome editing outside of *Oxytricha* offers exciting directions.

Methods

DNA collection and sequencing of *Tetmemena sp.*

Tetmemena sp. (strain SeJ-2015; ref. 7) was isolated as a single cell from a stock culture and propagated as a clonal strain via vegetative (asexual) cell culture. Cells were cultured in Pringsheim media (0.11mM Na_2HPO_4, 0.08mM $MgSO_4$, 0.85mM $Ca(NO_3)_2$, 0.35mM KCl, pH 7.0) and fed with *Chlamydomonas reinhardtii*, together with 0.1%(v/v) of an overnight culture of non-virulent *Klebsiella pneumoniae*. Macronuclei and micronuclei were isolated using sucrose gradient centrifugation (72). Genomic DNA was subsequently purified using the Nucleospin Tissue Kit (Takara Bio USA, Inc.). Macronuclear DNA was sequenced and assembled in Chen *et al.* (7). Micronuclear DNA was further size-selected via BluePippin (Sage Science) for PacBio sequencing, or via 0.6% (w/v) SeaKem Gold agarose electrophoresis (Lonza) for Illumina sequencing. Micronuclear DNA purification and sequencing protocols are described in (1).

DNA collection and sequencing for *E. woodruffi*

E. woodruffi (strain Iz01) was cultured in Volvic water at room temperature and fed with green algae every 2-3 days. We fed cells with *Chlamydomonas reinhardtii* for MAC DNA collection, and switched to *Chlorogonium capillatum* for MIC DNA collection. In order to remove algal contamination, cells were starved for at least 2-3 days before collection. Cells were washed and concentrated as in Chen et al. (1). Because MAC DNA is predominant in whole cell DNA, we used whole cell DNA (purified via NucleoSpin Tissue kit, Takara Bio USA, Inc.) for

519 MAC genome sequencing. Paired-end sequencing was performed on an Illumina Hiseq2000 at
520 the Princeton University Genomics Core Facility.
521 MIC DNA was enriched from whole cell DNA and sequenced via three sequencing
522 platforms (Illumina, Pacific Biosciences and Oxford Nanopore Technologies). We used
523 conventional and pulse-field gel electrophoresis to enrich MIC DNA:
524 1) High-molecular-weight DNA was separated from whole cell DNA by gel-
525 electrophoresis (0.25% agarose gel at 4 °C, 120 V for 4 hr). The top band was cut from the gel
526 and purified with the QIAGEN QIAquick kit. The purified high-molecular-weight DNA was
527 directly sent to the group of Dr. Robert Sebra at the Icahn School of Medicine at Mount Sinai for
528 library construction and sequencing. BluePippin (Sage Science) separation was used before
529 sequencing to select DNA >10kb. DNA was sequenced on two platforms: Illumina HiSeq2500
530 (150 bp paired-end reads) and PacBio Sequel (SMRT reads).
531 2) High-molecular-weight DNA was also enriched by pulsed-field gel electrophoresis
532 (PFGE). *E. woodruffi* cells were mixed with 1% low-melt agarose to form plugs according to
533 Akematsu et al. (73), with addition of 1 hr incubation with 50 µg/ml RNase (Invitrogen
534 AM2288) in 10 mM Tris-HCl (pH7.5) at 37 °C for RNA depletion. After three washes of 1 hr
535 with 1x TE buffer, the DNA plugs were incubated in 1mM PMSF to inactivate proteinase K,
536 followed by MspJI (New England Biolabs) digestion at ^{m}CNNR(9/13) sites to remove
537 contaminant DNA (^{m}C indicates C5-methylation or C5- hydroxymethylation). Previous reports
538 have shown that no methylcytosine is detectable in vegetative cells of *Oxytricha* (74),
539 *Tetrahymena* (75) and *Paramecium* (76), suggesting that C5-methylation and C5-
540 hydroxymethylation are rarely involved in the vegetative growth of the ciliate lineage. We also
541 validated by qPCR that the quantity of two randomly selected MIC loci is not changed after the

MspJI digestion. On the contrary, algal genomic DNA is significantly digested by MspJI. Based on these results, we conclude that MspJI digestion can be used to remove bacterial and algal DNA with C5-methylation and C5-hydroxymethylation, leaving *E. woodruffi* MIC DNA intact. The agarose plugs containing digested DNA were then inserted into wells of 1.0% Certified Megabase agarose gel (Bio-Rad) for PFGE (CHEF-DR II System, Bio-Rad). The DNA was separated at 6 V, 14°C with 0.5X TBE buffer at a 120° angle for 24 hr with switch time of 60-120 seconds. We validated by qPCR that the *E. woodruffi* MIC chromosomes were not mobilized from the well, while the MAC DNA migrated into the gel. The MIC DNA was then extracted by phenol-chloroform purification. Library preparation and sequencing were performed at Oxford Nanopore Technologies (New York, NY).

MAC genome assembly of *E. woodruffi*

We assembled the MAC genome of *E.woodruffi* using the same pipeline for *Tetmemena sp.* (7) for comparative analysis: two draft genomes were assembled by SPAdes (77) and Trinity (78), and were then merged by CAP3 (79). Trinity, which is a software developed for *de novo* transcriptome assembly (78), has been used to assemble hypotrich MAC genomes (7) because their nanochromosome genome structure is similar to transcriptomes, including properties such as variable copy number and alternative isoforms (10). Telomeric reads were mapped to contigs by BLAT (80), and contigs were further extended and capped by telomeres when at least 5 reads pile up at a position near ends by custom python scripts (https://github.com/yifeng-evo/Oxytricha_Tetmemena_Euplotes/tree/main/MAC_genome_telomere_capping). The mitochondrial DNA was removed if the contig has a TBLASTX (81) hit on the *Oxytricha* mitochondrial genome (Genbank accession JN383842.1 and JN383843.1) or two *Euplotes*

mitochondrial genomes (*Euplotes minuta* GQ903130.1, *Euplotes crassus* GQ903131.1). Algal contigs were removed by BLASTN to all *Chlamydomonas reinhardtii* nucleotide sequences downloaded from Genbank. Non-telomeric contigs were mapped to bacterial NR by BLASTX to remove bacterial contaminations. The genome was further compressed by CD-HIT (82) in two steps: 1) contigs <500bp were removed if 90% of the short contig can be aligned to a contig >=500bp with 90% similarity (-c 0.9 -aS 0.9 -uS 0.1); 2) Then the genome was compressed by 95% similarity (-c 0.95 -aS 0.9 -uS 0.1). Contigs shorter than 500bp without telomeres were removed. Nine contigs, likely Tec contaminants from the MIC genome, were also excluded (Tblastn, "-db_gencode 10 -evalue 1e-5"), and they could be assembled due to the high copy number in the MIC genome (47, 48, Genbank accessions of Tec ORFs are AAA62601.1, AAA62602.1, AAA62603.1, AAA91339.1, AAA91340.1, AAA91341.1, AAA91342.1).

RNA sequencing of *E. woodruffi and Tetmemena sp.*

Three biological replicates of total RNA was isolated from asexually growing *E. woodruffi* and *Tetmemena sp.* cells using TRIzol reagent (Thermo Fisher Scientific), and enriched for the poly(A)+ fraction using the NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England Biolabs). Stranded RNA-seq libraries were constructed using the ScriptSeq v2 RNA-seq library preparation kit (Epicentre) and sequenced on an Illumina Nextseq500 at the Columbia Genome Center. For *E.woodruffi*, the transcriptome was assembled by Trinity (78) and transcript alignments to the MAC genome were generated by PASA (83).

Gene prediction of the *E. woodruffi* MAC genome and validation of MAC genome completeness

We followed the gene prediction pipeline developed by the Broad institute (https://github.com/PASApipeline/PASApipeline/wiki) using EVidenceModeler (EVM, 84) to generate the final gene predictions. EVM produced gene structures by weighted combination of evidence from three resources: *ab initio* prediction, protein alignments and transcript alignments (the weight was 3, 3, 10 respectively). *Ab initio* prediction was generated by BRAKER2 pipeline (85). Protein alignments for EVM were generated by mapping *Oxytricha* proteins to the *E. woodruffi* MAC genome by Exonerate (86). EVM predicted 33379 genes on MAC chromosomes with at least one telomere.

We assessed MAC genome completeness using three methods: 1) 28294 (80.6%) of the 35099 *E. woodruffi* MAC contigs have at least one telomere. 2) In the *E. woodruffi* genes predicted on telomeric contigs, 88.8% of BUSCO (87, 88) genes in the lineage database alveolata_odb10 were identified as complete. Within the 171 BUSCO genes, 135 are complete and single-copy, 17 are complete and duplicated, 7 are fragmented and 12 are missing. This represents the best *Euplotes* MAC genome assembly available. 3) We identified 51 tRNA encoding all 20 amino acids by tRNAscan-SE (89) in the MAC genome, including two suppressor tRNAs of UAA and UAG.

MIC genome assembly of *Tetmemena sp.*

The MIC genome of *Tetmemena* was assembled with a hybrid approach to combine reads from different sequencing platforms. *Tetmemena* Illumina reads were first assembled by SPAdes (77, parameters "-k 21,33,55,77,99,127 –careful"). PacBio reads were error corrected by

610 FMLRC (90) using Illumina reads with default parameters. Corrected PacBio reads were aligned
611 to both the MAC genome and the Illumina MIC assembly with BLASTN. Reads were removed
612 if they start or end with telomeres or are aligned better to the MAC. The remaining reads were
613 assembled with wtdbg2 (91, parameters "-x rs"). The PacBio assembly was polished by Pilon
614 (92) with the "--diploid" option. The Illumina and PacBio assemblies were merged by
615 quickmerge (93) with the "-l 5000" option.
616
617 **MIC genome assembly of *E. woodruffi***
618 The MIC genome of *E. woodruffi* was assembled using a similar procedure as described
619 above for *Tetmemena*. *E. woodruffi* reads were filtered to remove bacterial contamination,
620 including abundant high-GC contaminants, possibly endosymbionts (94). Nanopore reads with
621 GC content >= 55% were assembled by Flye (95) with the parameter "--meta" for metagenomic
622 assembly of bacterial contigs. We used kaiju (96) to identify bacteria taxa for these contigs. 9 of
623 10 top-covered contigs derive from Proteobacteria, from which many *Euplotes* symbionts derive
624 (94). Bacterial contamination was removed from Illumina reads if perfectly mapping to these
625 metagenomic contigs by Bowtie2 (97). The cleaned Illumina reads were then assembled by
626 SPAdes with "-k 21,33,55,77,99,127" (77). Pacbio raw reads and Nanopore raw reads with GC
627 content < 55% were aligned to a concatenated database containing both the MAC genome and
628 the Illumina MIC assembly with BLASTN. Reads were removed if they start or end with
629 telomeres or align better to the MAC. Remaining PacBio/Nanopore reads were assembled by
630 Flye with "--meta" mode. The PacBio-Nanopore assembly was polished by Pilon with the "--
631 diploid" option. Illumina and PacBio-Nanopore assemblies were merged by quickmerge with the
632 "-l 10000" option. Contigs shorter than 1kb were removed.

633

634 **MIC genome decontamination**

635 The draft MIC genome of *Tetmemena* was first mapped to telomeric MAC contigs by
636 BLASTN. MIC contigs containing MDSs were included in the final assembly. The rest of the
637 MIC contigs were filtered by a decontamination pipeline: 1) contigs were aligned to the
638 *Klebsiella pneumoniae* genome, *Chlamydomonas reinhardtii* genome and the *Oxytricha*
639 mitochondrial genome by BLASTN to remove contaminants; 2) the remaining contigs were then
640 searched against the bacteria NR database and a ciliate protein database (including protein
641 sequences annotated in *Tetrahymena thermophila*: http://www.ciliate.org/system/downloads/tet-
642 latest/4-Protein%20fasta.fasta; *Paramecium tetraurelia*: http://paramecium.cgm.cnrs-gif.fr; and
643 *Oxytricha trifallax*: https://oxy.ciliate.org) by BLASTX. Contigs with higher bit score to bacteria
644 NR or G+C >45% were removed. The *E. woodruffi* MIC genome was decontaminated similarly,
645 with addition of all *Chlorogonium* sequences (the algal food source) on NCBI and the two
646 *Euplotes* mitochondrial genomes (*Euplotes minuta* GQ903130.1, *Euplotes crassus* GQ903131.1)
647 to filter contaminants.

648

649 **Repeat identification**

650 The repeat content in the MIC genomes was identified by RepeatModeler 1.0.10 (98) and
651 RepeatMasker 4.0.7 (99) with default parameters.

652

653 **TBE/Tec detection**

654 Representative *Oxytricha* TBE ORFs (Genbank accession AAB42034.1, AAB42016.1
655 and AAB42018.1) were used as queries to search TBEs in the *Oxytricha* and *Tetmemena* MIC

genomes by TBLASTN (-db_gencode 6 -evalue 1e-7 -max_target_seqs 30000). Tec ORFs were similarly detected by using *Euplotes crassus* Tec1 and Tec2 ORFs as queries (-db_gencode 10 -evalue 1e-5 -max_target_seqs 30000, Genbank accessions of Tec ORFs are AAA62601.1, AAA62602.1, AAA62603.1, AAA91339.1, AAA91340.1, AAA91341.1, AAA91342.1). Complete TBEs/Tecs were determined by custom python scripts when three ORFs are within 2000 bp from each other and in correct orientation (https://github.com/yifeng-evo/Oxytricha_Tetmemena_Euplotes/tree/main/TBE_ORFs/TBE_to_oxy_genome_tblastn_parse.py, 44). 30 TBE ORFs with >70% completeness were subsampled from each species for phylogenetic analysis (except for the 57kD ORF in *Tetmemena*, for which 21 were subsampled). The subsampled TBE ORFs were aligned using MUSCLE (100) and the alignments were trimmed by trimAl "-automated1" (101). Phylogenetic trees were constructed using PhyML 3.3 (102).

Rearrangement annotations

SDRAP (53) was used to annotate MDSs, pointers and MIC-specific regions (minimum percent identity for preliminary match annotation=95, minimum percent identity for additional match annotation=90, minimum length of pointer annotation=2). SDRAP requires MAC and MIC genomes as input. For the SDRAP annotation of *Oxytricha*, we used the MAC genome from Swart et al. (6) instead of the latest hybrid assembly that incorporated PacBio reads (10), because the former version was primarily based on Illumina reads, similar to the MAC genomes of *Tetmemena* (7, Genbank GCA_001273295.2) and *E. woodruffi* which are also Illumina assemblies. *Oxytricha* and *Tetmemena* MAC genomes were preprocessed by removing MAC contigs with TBE ORFs, considered MIC contaminants (44). SDRAP is a new program that can

679 output the rearrangement annotations with minor differences from Chen et al. (1) but most
680 annotations are robust (Figure 3 - figure supplement 2). Scrambled and nonscrambled
681 junctions/IESs were annotated by custom python scripts (https://github.com/yifeng-
682 evo/Oxytricha_Tetmemena_Euplotes/tree/main/scrambled_nonscrambled_IES_pointer).
683
684 **MIC genome categories**
685 Each MIC genome region is assigned to only one category in Figure 2A-C, even if it
686 belongs to more than one category. The assignment is based on the following priority: MDS,
687 TBE/Tec, MIC genes (only available for *Oxytricha*, which has developmental RNA-seq data),
688 IES, tandem repeats, other repeats and non-coding non-repetitive regions. For example, a MIC
689 region can be a TBE in an IES, and it is only considered as TBE in Figure 2A-C.
690
691 **Ortholog comparison pipeline and Monte Carlo simulations**
692 Orthogroups of genes on telomeric MAC contigs were detected by OrthoFinder with "-S
693 blast" (103). Single-copy orthologs were aligned by Clustal Omega (104). Protein alignments
694 were reversely translated to CDS alignments by a modified script of pal2nal (105,
695 https://github.com/yifeng-
696 evo/Oxytricha_Tetmemena_Euplotes/tree/main/Ortholog_comparison/pal2nal.pl). Two
697 modifications were made in the script: 1) the modified script allows pal2nal to take different
698 genetic codes for three sequences (-codontable 6,6,10); 2) the script also fixed an error in the
699 original pal2nal script in which codontable 10 for the Euplotid nuclear code was the same as the
700 universal code. Visualization of pointer positions and intron locations on orthologs was
701 implemented by a custom python script (https://github.com/yifeng-

evo/Oxytricha_Tetmemena_Euplotes/blob/main/Ortholog_comparison/visualization_of_ortholog_comparison.py). Pointer positions or intron locations are considered conserved if they are within a 20 bp alignment window on the CDS alignment. Protein domains were annotated by HMMER (106). We performed Monte Carlo simulations by randomly shuffling pointer locations on the CDS but keeping their original position distribution. This was implemented by a custom python script, which transforms the CDS to a circle, rotates pointer positions on the circle and outputs the shuffled position on the re-linearized CDS (https://github.com/yifeng-evo/Oxytricha_Tetmemena_Euplotes/blob/main/Ortholog_comparison/shuffle_simulation.py). The null hypothesis of the Monte Carlo test is that pointers positions are conserved by chance. *P*-value of Monte Carlo test is given by $N_{expected>observed}/N_{total}$ ($N_{expected>observed}$ is the number of simulations when there are more conserved pointers in the simulation than the observation from real data, N_{total} =1000 in this study).

PCR validation of Russian doll locus

The complex Russian doll locus on MIC contig TMEMEN_MIC_21461 in *Tetmemena* was validated by PCR to confirm the *Tetmemena* MIC genome assembly. *Tetmemena* micronuclear DNA was purified as described previously and used as template for PCR using PrimeSTAR Max DNA polymerase (Takara Bio). 11 primer sets (Supplementary File 14) were designed to amplify products between 3 kb and 6 kb in length, with overlapping regions between consecutive primer pairs. The resulting PCR products were visualized through agarose gel electrophoresis and bands of the expected size were extracted using a Monarch® DNA Gel Extraction Kit (New England Biolabs). The purified gel bands were cloned using a TOPO™ XL-2 Complete PCR Cloning Kit (Invitrogen), transformed into One Shot™ OmniMAX™ 2 T1R *E.*

coli cells (Invitrogen), and individual clones were grown and their plasmids harvested with a QIAprep Spin Miniprep Kit (QIAGEN). The plasmid ends were Sanger sequenced, as well as the region where the *Oxytricha* MIC assembly contains inserted MDSs (Genewiz). Sanger sequencing reads were mapped to the *Tetmemena* MIC contig TMEMEN_MIC_21461 and visualized using Geneious Prime® 2021.1.1 (https://www.geneious.com).

Acknowledgements

We thank Toshinobu Suzaki (Kobe University) for the gift of *E. woodruffi* (strain Iz01 from Shizuoka Prefecture) and *Chlorogonium capillatum*. We thank Sheela George for laboratory support and help with cell collection. We thank David Dai, Eoghan Harrington, John Beaulaurier and Sissel Juul at Oxford Nanopore Technologies in New York for providing sequencing and advice. We thank Robert Sebra and Melissa Smith for advice on PacBio sequencing. We thank Takahiko Akematsu, Lorraine Symington and Lea Marie for helping with PFGE. We thank Kaiyi Zhu, Shaojie He, Molly Przeworski, Harmen Bussemaker and Nataša Jonoska for advice on Monte Carlo simulation. We also thank Samuel Sternberg, Bill Jack, Scott Roy, and all current and past Landweber lab members for discussion about the origin of scrambled genes and Margarita T. Angelova, Sindhuja Devanapally, Danylo Villano and Kehan Bao for comments on the manuscript. This work was supported by the National Institutes of Health, R35GM122555, and National Science Foundation, DMS1764366, and the National Center for Genome Analysis Support computing resources (supported by National Science Foundation DBI1062432, ABI1458641, and ABI1759906 to Indiana University). Rafik Neme was supported by the Pew Latin American Fellows Program.

Availability of data and materials

Custom scripts are public on https://github.com/yifeng-evo/Oxytricha_Tetmemena_Euplotes. DNA-seq reads and genome assemblies are available at GenBank under Bioprojects PRJNA694964 (*Tetmemena sp.*) and PRJNA781979 (*Euplotes woodruffi*). Genbank accession numbers for genomes are JAJKFJ000000000 (*Tetmemena sp.* Micronucleus genome), JAJLLS000000000 (*Euplotes woodruffi* Micronucleus genome), and JAJLLT000000000 (*Euplotes woodruffi* Macronucleus genome).

Three replicates of RNA-seq reads for vegetative cells are available at GenBank under accession numbers of SRR21815378, SRR21815379, SRR21815380 for *E. woodruffi* and SRR21817702, SRR21817703 and SRR21817704 for *Tetmemena sp.*.

MDSs annotations for three species are available at https://doi.org/10.5061/dryad.5dv41ns96 and https://knot.math.usf.edu/mds_ies_db/2022/downloads.html (please select species from the drop-down menu).

Declaration of interests

Leslie Y. Beh is an employee at Illumina Inc.

Xiao Chen is an employee at Pacific Biosciences.

References

1. Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM, Swart EC, Perlman DH, Doak TG, Stuart A, Amemiya CT, Sebra RP, Landweber LF. The architecture of a

scrambled genome reveals massive levels of genomic rearrangement during development. Cell. 2014 Aug 28;158(5):1187-98.

2. Mitreva M, Blaxter ML, Bird DM, McCarter JP. Comparative genomics of nematodes. Trends in Genetics. 2005 Oct 1;21(10):573-81.
3. Smith JJ, Baker C, Eichler EE, Amemiya CT. Genetic consequences of programmed genome rearrangement. Current Biology. 2012 Aug 21;22(16):1524-9.
4. Biederman MK, Nelson MM, Asalone KC, Pedersen AL, Saldanha CJ, Bracht JR. Discovery of the first germline-restricted gene by subtractive transcriptomic analysis in the zebra finch, *Taeniopygia guttata*. Current Biology. 2018 May 21;28(10):1620-7.
5. Beh LY, Debelouchina GT, Clay DM, Thompson RE, Lindblad KA, Hutton ER, Bracht JR, Sebra RP, Muir TW, Landweber LF. Identification of a DNA N6-adenine methyltransferase complex and its impact on chromatin organization. Cell. 2019 Jun 13;177(7):1781-96.
6. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, Khurana JS, Goldman AD, Nowacki M, Schotanus K, Jung S., Fulton RS, Ly A, McGrath S, Haub K, Wiggins JL, Storton D, Matese JC, Parsons L, Chang WJ, Bowen MS, Stover NA, Jones TA, Eddy SR, Herrick GA, Doak TG, Wilson RK, Mardis ER, Landweber LF. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biology. 2013 Jan 29;11(1):e1001473.
7. Chen X, Jung S, Beh LY, Eddy SR, Landweber LF. Combinatorial DNA rearrangement facilitates the origin of new genes in ciliates. Genome Biology and Evolution. 2015 Oct 1;7(10):2859-70.

8. Wang R, Xiong J, Wang W, Miao W, Liang A. High frequency of +1 programmed ribosomal frameshifting in *Euplotes octocarinatus*. Scientific Reports. 2016 Feb 19;6(1):1-2.
9. Chen X, Jiang Y, Gao F, Zheng W, Krock TJ, Stover NA, Lu C, Katz LA, Song W. Genome analyses of the new model protist *Euplotes vannus* focusing on genome rearrangement and resistance to environmental stressors. Molecular Ecology Resources. 2019 Sep;19(5):1292-308.
10. Lindblad KA, Pathmanathan JS, Moreira S, Bracht JR, Sebra RP, Hutton ER, Landweber LF. Capture of complete ciliate chromosomes in single sequencing reads reveals widespread chromosome isoforms. BMC Genomics. 2019 Dec;20(1):1-1.
11. Vinogradov DV, Tsoĭ OV, Zaika AV, Lobanov AV, Turanov AA, Gladyshev VN, Gel'fand MS. Draft macronuclear genome of a ciliate *Euplotes crassus*. Molekuliarnaia Biologiia. 2012 Mar 1;46(2):361-6.
12. Mitcham JL, Lynn AJ, Prescott DM. Analysis of a scrambled gene: the gene encoding alpha-telomere-binding protein in *Oxytricha nova*. Genes and Development. 1992 May 1;6(5):788-800.
13. Prescott DM. The DNA of ciliated protozoa. Microbiological Reviews. 1994 Jun;58(2):233-67.
14. Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, Landweber LF. RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature. 2008 Jan;451(7175):153-8.

15. Lindblad KA, Bracht JR, Williams AE, Landweber LF. Thousands of RNA-cached copies of whole chromosomes are present in the ciliate *Oxytricha* during development. RNA. 2017 Aug 1;23(8):1200-8.

16. Yerlici VT, Landweber LF. Programmed genome rearrangements in the ciliate *Oxytricha*. Microbiology Spectrum. 2014 Dec 5;2(6):2-6.

17. Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury JM, Denby Wilkes C, Garnier O, Labadie K, Lauderdale BE, Le Mouël A, Marmignon A. The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. PLoS Genetics. 2012;8(10):e1002984. doi: 10.1371/journal.pgen.1002984.

18. Maurer-Alcalá XX, Knight R, Katz LA. Exploration of the germline genome of the ciliate *Chilodonella uncinata* through single-cell omics (transcriptomics and genomics). MBio. 2018 Jan 9;9(1):e01836-17.

19. Maurer-Alcalá XX, Yan Y, Pilling OA, Knight R, Katz LA. Twisted tales: insights into genome diversity of ciliates using single-cell 'omics. Genome Biology and Evolution. 2018 Aug;10(8):1927-38.

20. Smith SA, Maurer-Alcalá XX, Yan Y, Katz LA, Santoferrara LF, McManus GB. Combined genome and transcriptome analyses of the ciliate *Schmidingerella arcuata* (Spirotrichea) reveal patterns of DNA elimination, scrambling, and inversion. Genome Biology and Evolution. 2020 Sep;12(9):1616-22.

21. Katz LA, Kovner AM. Alternative processing of scrambled genes generates protein diversity in the ciliate *Chilodonella uncinata*. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 2010 Sep 15;314(6):480-8.

22. Gao F, Song W, Katz LA. Genome structure drives patterns of gene family evolution in ciliates, a case study using *Chilodonella uncinata* (Protista, Ciliophora, Phyllopharyngea). Evolution. 2014 Aug;68(8):2287-95.
23. Denby Wilkes C, Arnaiz O, Sperling L. ParTIES: a toolbox for *Paramecium* interspersed DNA elimination studies. Bioinformatics. 2016 Feb 15;32(4):599-601.
24. Zheng W, Chen J, Doak TG, Song W, Yan Y. ADFinder: accurate detection of programmed DNA elimination using NGS high-throughput sequencing data. Bioinformatics. 2020 Jun 1;36(12):3632-6.
25. Feng Y, Beh LY, Chang WJ, Landweber LF. SIGAR: Inferring Features of Genome Architecture and DNA Rearrangements by Split-Read Mapping. Genome Biology and Evolution. 2020 Oct;12(10):1711-8.
26. Seah BK, Swart EC. BleTIES: Annotation of natural genome editing in ciliates using long read sequencing. bioRxiv. 2021 Jan 1. doi.org/10.1101/2021.05.18.444610
27. Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, Hadjithomas M, Krishnakumar V, Badger JH, Caler EV, Russ C. Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. Elife. 2016 Nov 28;5:e19090.
28. Guérin F, Arnaiz O, Boggetto N, Wilkes CD, Meyer E, Sperling L, Duharcourt S. Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements. BMC Genomics. 2017 Dec;18(1):1-7.
29. Sellis D, Guérin F, Arnaiz O, Pett W, Lerat E, Boggetto N, Krenek S, Berendonk T, Couloux A, Aury JM, Labadie K, Malinsky S, Bhullar S, Meyer E, Sperling L, Duret L,

Duharcourt S. Massive colonization of protein-coding exons by selfish genetic elements in *Paramecium* germline genomes. Plos Biology. 2021 Jul 29;19(7):e3001309.

30. Sheng Y, Duan L, Cheng T, Qiao Y, Stover NA, Gao S. The completed macronuclear genome of a model ciliate *Tetrahymena thermophila* and its application in genome scrambling and copy number analyses. Science China Life Sciences. 2020 Oct;63(10):1534-42.f
31. Parfrey LW, Lahr DJ, Knoll AH, Katz LA. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proceedings of the National Academy of Sciences. 2011 Aug 16;108(33):13624-9.
32. Bracht JR, Fang W, Goldman AD, Dolzhenko E, Stein EM, Landweber LF. Genomes on the edge: programmed genome instability in ciliates. Cell. 2013 Jan 31;152(3):406-16.
33. Hogan DJ, Hewitt EA, Orr KE, Prescott DM, Müller KM. Evolution of IESs and scrambling in the actin I gene in hypotrichous ciliates. Proceedings of the National Academy of Sciences. 2001 Dec 18;98(26):15101-6.
34. Chang WJ, Bryson PD, Liang H, Shin MK, Landweber LF. The evolutionary origin of a complex scrambled gene. Proceedings of the National Academy of Sciences. 2005 Oct 18;102(42):15149-54.
35. Wong LC, Landweber LF. Evolution of programmed DNA rearrangements in a scrambled gene. Molecular Biology and Evolution. 2006 Apr 1;23(4):756-63.
36. DuBois MI, Prescott DM. Scrambling of the actin I gene in two *Oxytricha* species. Proceedings of the National Academy of Sciences. 1995 Apr 25;92(9):3888-92.

878 37. Chen W, Zuo C, Wang C, Zhang T, Lyu L, Qiao Y, Zhao F, Miao M. The hidden
879 genomic diversity of ciliated protists revealed by single-cell genome sequencing. BMC
880 Biology. 2021 Dec;19(1):1-3.
881 38. Meyer F, Schmidt HJ, Plümper E, Hasilik A, Mersmann G, Meyer HE, Engström A,
882 Heckmann K. UGA is translated as cysteine in pheromone 3 of *Euplotes octocarinatus*.
883 Proceedings of the National Academy of Sciences. 1991 May 1;88(9):3758-61.
884 39. Klobutcher LA, Herrick G. Consensus inverted terminal repeat sequence of *Paramecium*
885 IESs: resemblance to termini of Tc1-related and *Euplotes* Tec transposons. Nucleic Acids
886 Research. 1995 Jun 11;23(11):2006-13.
887 40. Tan M, Brünen-Nieweler C, Heckmann K. Isolation of micronuclei from *Euplotes*
888 *octocarinatus* and identification of an internal eliminated sequence in the micronuclear
889 gene encoding γ-tubulin 2. European Journal of Protistology. 1999 Jun 21;35(2):208-16.
890 41. Wang W, Zhi H, Chai B, Liang A. Cloning and sequence analysis of the micronuclear
891 and macronuclear gene encoding Rab protein of *Euplotes octocarinatus*. Bioscience,
892 Biotechnology, and Biochemistry. 2005 Jan 1;69(3):649-52.
893 42. Syberg-Olsen MJ, Irwin NA, Vannini C, Erra F, Di Giuseppe G, Boscaro V, Keeling PJ.
894 Biogeography and character evolution of the ciliate genus *Euplotes* (Spirotrichea,
895 Euplotia), with description of *Euplotes curdsi* sp. nov. PloS One. 2016 Nov
896 9;11(11):e0165442.
897 43. Gao F, Roy SW, Katz LA. Analyses of alternatively processed genes in ciliates provide
898 insights into the origins of scrambled genomes and may provide a mechanism for
899 speciation. MBio. 2015 Feb 3;6(1):e01998-14.

44. Chen X, Landweber LF. Phylogenomic analysis reveals genome-wide purifying selection on TBE transposons in the ciliate *Oxytricha*. Mobile DNA. 2016 Dec;7(1):1-0.

45. Baird SE, Fino GM, Tausta SL, Klobutcher LA. Micronuclear genome organization in *Euplotes crassus*: a transposonlike element is removed during macronuclear development. Molecular and Cellular Biology. 1989 Sep 1;9(9):3793-807.

46. Krikau MF, Jahn CL. Tec2, a second transposon-like element demonstrating developmentally programmed excision in *Euplotes crassus*. Molecular and Cellular Biology. 1991 Sep;11(9):4751-9.

47. Jahn CL, Doktor SZ, Frels JS, Jaraczewski JW, Krikau MF. Structures of the *Euplotes crassus* Tec1 and Tec2 elements: identification of putative transposase coding regions. Gene. 1993 Oct 29;133(1):71-8.

48. Jahn CL, Krikau MF, Shyman S. Developmentally coordinated en masse excision of a highly repetitive element in *E. crassus*. Cell. 1989 Dec 22;59(6):1009-18.

49. Klobutcher LA, Herrick GL. Developmental genome reorganization in ciliated protozoa: the transposon link. Progress in Nucleic Acid Research and Molecular Biology. 1997 Mar 21;56:1-62.

50. Doak TG, Witherspoon DJ, Jahn CL, Herrick G. Selection on the genes of *Euplotes crassus* Tec1 and Tec2 transposons: evolutionary appearance of a programmed frameshift in a Tec2 gene encoding a tyrosine family site-specific recombinase. Eukaryotic cell. 2003 Feb 1;2(1):95-102.

51. Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG, Landweber LF. A functional role for transposases in a large eukaryotic genome. Science. 2009 May 15;324(5929):935-8.

52. Feng Y, Landweber LF. Transposon debris in ciliate genomes. PLoS Biology. 2021 Aug 24;19(8):e3001354.

53. Braun J, Neme R, Feng Y, Landweber LF, Jonoska N. SDRAP for annotating scrambled or rearranged genomes. bioRxiv 2022.10.24.513505

54. Lynn D. The ciliated protozoa: characterization, classification, and guide to the literature. Springer Science & Business Media; 2008 Jun 24.

55. Gao F, Warren A, Zhang Q, Gong J, Miao M, Sun P, Xu D, Huang J, Yi Z, Song W. The all-data-based evolutionary hypothesis of ciliated protists with a revised classification of the phylum Ciliophora (Eukaryota, Alveolata). Scientific Reports. 2016 Apr 29;6(1):1-4.

56. Landweber LF, Kuo TC, Curtis EA. Evolution and assembly of an extremely scrambled gene. Proceedings of the National Academy of Sciences. 2000 Mar 28;97(7):3298-303.

57. Burns J, Kukushkin D, Chen X, Landweber LF, Saito M, Jonoska N. Recurring patterns among scrambled genes in the encrypted genome of the ciliate *Oxytricha trifallax*. Journal of Theoretical Biology. 2016 Dec 7;410:171-80.

58. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Current Biology. 2003 Sep 2;13(17):1512-7.

59. Möllenbeck M, Cavalcanti AR, Jönsson F, Lipps HJ, Landweber LF. Interconversion of germline-limited and somatic DNA in a scrambled gene. Journal of Molecular Evolution. 2006 Jul;63(1):69-73.

60. Braun J, Nabergall L, Neme R, Landweber LF, Saito M, Jonoska N. Russian doll genes and complex chromosome rearrangements in *Oxytricha trifallax*. G3: Genes, Genomes, Genetics. 2018 May 1;8(5):1669-74.

946 61. Thomas Jr CA. The genetic organization of chromosomes. Annual review of genetics.
947 1971 Dec;5(1):237-56.
948 62. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of
949 eukaryotic genome content. Philosophical Transactions of the Royal Society B:
950 Biological Sciences. 2015 Sep 26;370(1678):20140331.
951 63. Mourier T, Jeffares DC. Eukaryotic intron loss. Science. 2003 May 30;300(5624):1393-
952 1393.
953 64. Parenteau J, Maignon L, Berthoumieux M, Catala M, Gagnon V, Abou Elela S. Introns
954 are mediators of cell response to starvation. Nature. 2019 Jan;565(7741):612-7.
955 65. Morgan JT, Fink GR, Bartel DP. Excised linear introns regulate growth in yeast. Nature.
956 2019 Jan;565(7741):606-11.
957 66. Hoffman DC, Prescott DM. Evolution of internal eliminated segments and scrambling in
958 the micronuclear gene encoding DNA polymerase α in two *Oxytricha* species. Nucleic
959 acids research. 1997 May 1;25(10):1883-9.
960 67. Prescott JD, DuBois ML, Prescott DM. Evolution of the scrambled germline gene
961 encoding α-telomere binding protein in three hypotrichous ciliates. Chromosoma. 1998
962 Nov;107(5):293-303.
963 68. Landweber LF. Why genomes in pieces?. Science. 2007 Oct 19;318(5849):405-7.
964 69. Landweber LF. Making sense of scrambled genomes. Science. 2008 Feb
965 15;319(5865):901-2.
966 70. Schmidbaur H, Kawaguchi A, Clarence T, Fu X, Hoang OP, Zimmermann B, Ritschard
967 EA, Weissenbacher A, Foster JS, Nyholm SV, Bates PA. Emergence of novel

968 cephalopod gene regulation and expression through large-scale genome reorganization.
969 Nature communications. 2022 Apr 21;13(1):1-1.
970 71. Albertin CB, Medina-Ruiz S, Mitros T, Schmidbaur H, Sanchez G, Wang ZY, Grimwood
971 J, Rosenthal JJ, Ragsdale CW, Simakov O, Rokhsar DS. Genome and transcriptome
972 mechanisms driving cephalopod evolution. Nature communications. 2022 May 4;13(1):1-
973 4.
974 72. Lauth MR, Spear BB, Heumann J, Prescott DM. DNA of ciliated protozoa: DNA
975 sequence diminution during macronuclear development of *Oxytricha.* Cell. 1976 Jan
976 1;7(1):67-74.
977 73. Akematsu T, Fukuda Y, Garg J, Fillingham JS, Pearlman RE, Loidl J. Post-meiotic DNA
978 double-strand breaks occur in *Tetrahymena,* and require Topoisomerase II and Spo11.
979 Elife. 2017 Jun 16;6:e26176.
980 74. Bracht JR, Perlman DH, Landweber LF. Cytosine methylation and hydroxymethylation
981 mark DNA for elimination in *Oxytricha trifallax*. Genome Biology. 2012 Oct;13(10):1-
982 23.
983 75. Gorovsky MA, Hattman S, Pleger GL. [6N] Methyl adenine in the nuclear DNA of a
984 eucaryote, *Tetrahymena pyriformis*. The Journal of Cell Biology. 1973 Mar 1;56(3):697-
985 701.
986 76. Cummings DJ, Tait A, Goddard JM. Methylated bases in DNA from *Paramecium*
987 *aurelia*. Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis.
988 1974 Nov 20;374(1):1-1.
989 77. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
990 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV. SPAdes: a new genome assembly

algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 2012 May 1;19(5):455-77.

78. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011 Jul;29(7):644-52.

79. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Research. 1999 Sep 1;9(9):868-77.

80. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Research. 2002 Apr 1;12(4):656-64.

81. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec;10(1):1-9.

82. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150-2.

83. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research. 2003 Oct 1;31(19):5654-66.

84. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology. 2008 Sep;9(1):1-22.

85. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinformatics. 2021 Mar;3(1):lqaa108.

86. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005 Dec;6(1):1-1.
87. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210-2.
88. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular Biology and Evolution. 2021 Oct;38(10):4647-54.
89. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research. 1997 Mar 1;25(5):955-64.
90. Wang JR, Holt J, McMillan L, Jones CD. FMLRC: Hybrid long read error correction using an FM-index. BMC Bioinformatics. 2018 Dec;19(1):1-1.
91. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nature Methods. 2020 Feb;17(2):155-8.
92. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One. 2014 Nov 19;9(11):e112963.
93. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Research. 2016 Nov 2;44(19):e147-e147.

94. Boscaro V, Husnik F, Vannini C, Keeling PJ. Symbionts of the ciliate *Euplotes*: diversity, patterns and potential as models for bacteria–eukaryote endosymbioses. Proceedings of the Royal Society B. 2019 Jul 24;286(1907):20190693.

95. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology. 2019 May;37(5):540-6.

96. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications. 2016 Apr 13;7(1):1-9.

97. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012 Apr;9(4):357-9.

98. Smit AF, Hubley R. RepeatModeler Open-1.0. 2008-2015 <http://www.repeatmasker.org>.

99. Smit AF, Hubley R, Green P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.

100. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004 Mar 1;32(5):1792-7.

101. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1;25(15):1972-3.

102. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology. 2010 May 1;59(3):307-21.

103. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology. 2019 Dec;20(1):1-4.

104. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011;7(1):539.

105. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research. 2006 Jul 1;34(suppl_2):W609-12.

106. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research. 2011 May 18;39(suppl_2):W29-37.

107. Miller RV, Neme R, Clay DM, Pathmanathan JS, Lu MW, Yerlici VT, Khurana JS, Landweber LF. Transcribed germline-limited coding sequences in *Oxytricha trifallax*. G3. 2021 Jun;11(6):jkab092.

108. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, Tallon LJ. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biology. 2006 Sep;4(9):e286.

109. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 2006 Nov;444(7116):171-8.

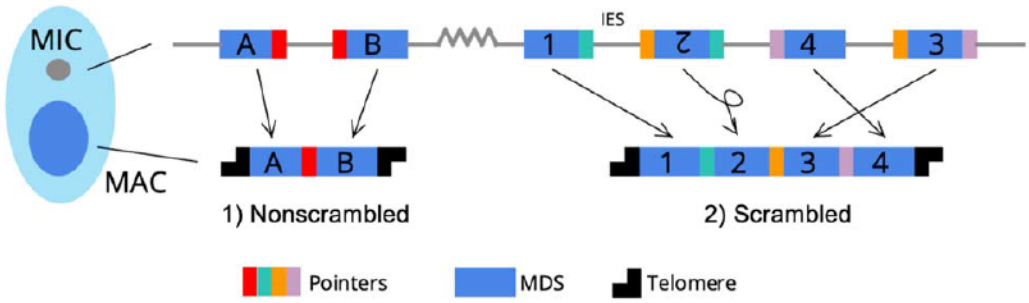
110. Riley JL, Katz LA. Widespread distribution of extensive chromosomal fragmentation in ciliates. Molecular Biology and Evolution. 2001 Jul 1;18(7):1372-7.

Table 1. Statistics of MAC and MIC genomes in three species

	Oxytricha trifallax		*Tetmemena sp.*		*Euplotes woodruffi*	
	MAC[6,b]	MIC[1]	MAC[7]	MIC[c]	MAC[c]	MIC[c]
genome size (Mbp)	67.1	496	60.6	237	72.2	172
N50 (bp)	3,745	27,807	3,339	14,722	2,702	44,656
GC%	31.36	28.44	37.05	32.17	36.56	35.31
number of contigs[a]	22,426	25,720	25,206	28,446	35,099	17,655
Two-telomere contigs	14,225	-	15,802	-	19,061	-
Telomeric contigs	20,336	-	21,165	-	28,294	-
Single-gene telomeric contigs	76.1%	-	75.5%	-	68.5%	-
Maximum number of genes on a telomeric contig	8	-	7	-	8	-

[a] TBE contaminants in MAC contigs were removed (Methods). Therefore, 24 *Oxytricha* MAC contigs and 13 *Tetmemena* MAC contigs were removed from the published versions.
[b] This study used the MAC genome of *Oxytricha* from Swart et al. (6) instead of the long-read assembly in Lindblad et al. (10), because the short MAC genomes in the present study were primarily assembled from Illumina reads, as in (6). (10) updated (6) by including nanochromosomes captured in single long reads, which are currently not available for the other two species. The MIC genomes of *Tetmemena* and *E. woodruffi* were assembled to a similar N50 as the reference *O. trifallax* genome (1) for comparative analysis.
[c] Data from this study.

A



B

	MAC gene-sized nanochromosomes	MIC genome assembly	Average pointer length	Scrambled genes
Oxytricha trifallax	Yes[6,10]	Yes[1]	scrambled 11bp nonscrambled 5bp[1]	Yes[1]
Tetmemena sp.	Yes[7]	**Yes**	**scrambled 10bp nonscrambled 5bp**	**Yes**
Euplotes woodruffi	**Yes**	**Yes**	**scrambled 13bp nonscrambled 4bp ***	**Yes**
Tetrahymena thermophila	No[30, 108]	Yes[27]	0 - 4bp AT-rich variable positions[27]	Yes[30]
Paramecium tetraurelia	No[109]	Yes[28]	2bp (TA)[17]	No
Chilodonella uncinata	Yes[110]	Partial genome assembled using single-cell omics[18]	scrambled 9bp nonscrambled 7bp[18]	Yes[18, 21, 22]
Plasmodium falciparum	NA	NA	NA	NA

Figure 1. Genome rearrangements in representative ciliate species. A) Diagram of genome rearrangement in *Oxytricha*. Each ciliate cell contains a somatic macronucleus (MAC) and a germline micronucleus (MIC). During development, the MAC genome rearranges from a copy of the MIC genome. 1) Nonscrambled genes rearrange simply by joining consecutive macronuclear destined sequences (MDSs, blue boxes) and removing internal eliminated sequences (IESs, thin lines). 2) Rearrangement of scrambled genes requires MDS translocation and/or inversion.

Pointers are microhomologous sequences (colored vertical bars) present in two copies in the MIC and only one copy in the MAC where consecutive MDSs recombine. B) Comparison of genome rearrangement features of representative ciliates and the non-ciliate *Plasmodium falciparum* as an outgroup (phylogenetic information is based on refs. 31 and 32). Conclusions from this study are shown in bold. * indicates that some scrambled pointers in *E. woodruffi* are much longer, as discussed in the results. Statistics for pointers <=30bp in *E. woodruffi* are shown.

Table information derives from the following sources:

[1] Chen, Xiao, et al. "The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development." *Cell* 158.5 (2014): 1187-1198. [6] Swart, Estienne C., et al. "The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes." *PLoS biology* 11.1 (2013): e1001473. [7] Chen, Xiao, et al. "Combinatorial DNA rearrangement facilitates the origin of new genes in ciliates." *Genome biology and evolution* 7.10 (2015): 2859-2870. [10] Lindblad, Kelsi A., et al. "Capture of complete ciliate chromosomes in single sequencing reads reveals widespread chromosome isoforms." *BMC genomics* 20.1 (2019): 1-11. [17] Arnaiz, Olivier, et al. "The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences." (2012): e1002984. [18] Maurer-Alcalá, Xyrus X., Rob Knight, and Laura A. Katz. "Exploration of the germline genome of the ciliate *Chilodonella uncinata* through single-cell omics (transcriptomics and genomics)." *MBio* 9.1 (2018): e01836-17. [21] Katz, Laura A., and Alexandra M. Kovner. "Alternative processing of scrambled genes generates protein diversity in the ciliate *Chilodonella uncinata*." *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 314.6 (2010): 480-488. [22] Gao, Feng, Weibo Song,

and Laura A. Katz. “Genome structure drives patterns of gene family evolution in ciliates, a case study using *Chilodonella 52uncinate* (Protista, Ciliophora, Phyllopharyngea).” *Evolution* 68.8 (2014): 2287-2295. [27] Hamilton, Eileen P., et al. "Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome." *eLife* 5 (2016): e19090. [28] Guérin, Frédéric, et al. "Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements." *BMC genomics*18.1 (2017): 1-17. [30] Sheng, Yalan, et al. "The completed macronuclear genome of a model ciliate *Tetrahymena thermophila* and its application in genome scrambling and copy number analyses." *Science China Life Sciences* 63.10 (2020): 1534-1542. [108] Eisen, Jonathan A., et al. "Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote." *PLoS Biology* 4.9 (2006): e286. [109] Aury, Jean-Marc, et al. "Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*." *Nature* 444.7116 (2006): 171-178. [110] Riley, Jennifer L., and Laura A. Katz. "Widespread distribution of extensive chromosomal fragmentation in ciliates." *Molecular Biology and Evolution* 18.7 (2001): 1372-1377.

Full citation information for refs. 31 and 32:

[31] Parfrey, Laura Wegener, et al. "Estimating the timing of early eukaryotic diversification with multigene molecular clocks." *Proceedings of the National Academy of Sciences* 108.33 (2011): 13624-13629.

[32] Bracht, John R., et al. "Genomes on the edge: programmed genome instability in ciliates." *Cell* 152.3 (2013): 406-416.

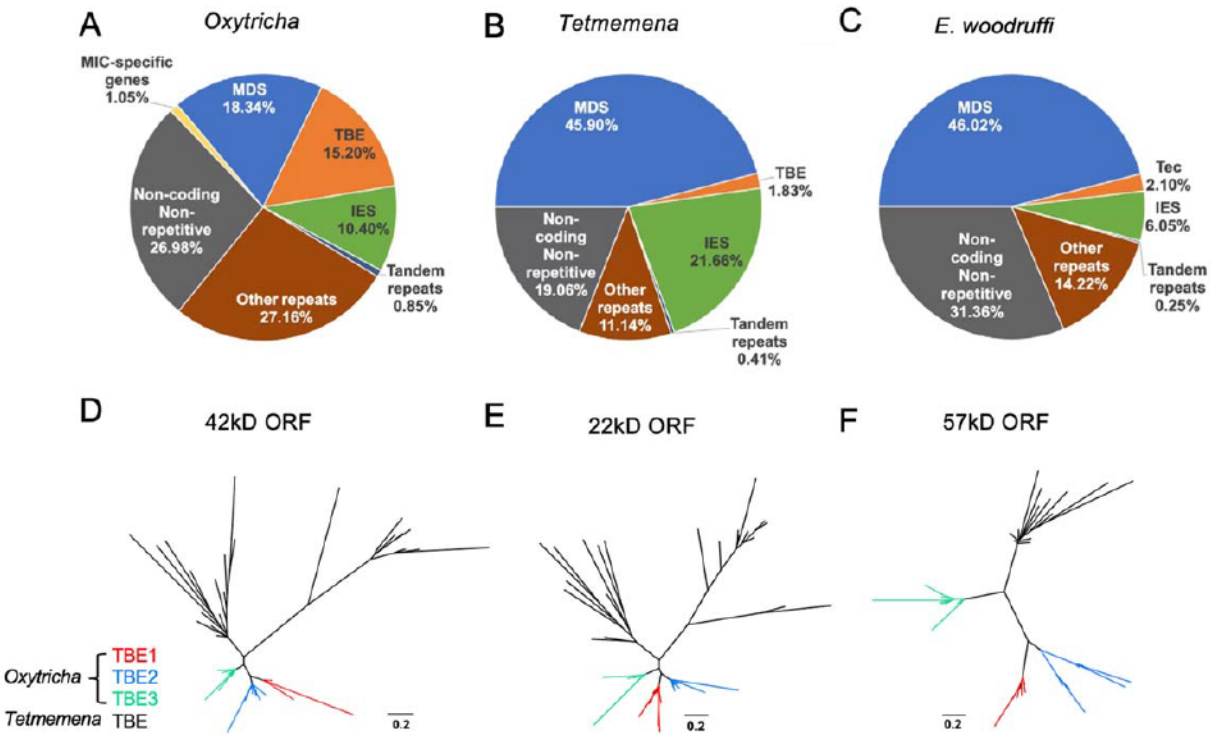


Figure 2. The three MIC genomes differ in repeat content, especially transposable elements. A-C) MIC genome categories for (A) *Oxytricha trifallax*, (B) *Tetmemena sp.*, and (C) *E. woodruffi*. *Oxytricha* displays the greatest proportion of repetitive elements (TBE, Other repeats, and Tandem repeats) relative to the other species. *Oxytricha* MIC-specific genes were annotated in (1, 107). D-F) Phylogenetic analysis of the three TBE ORFs in *Oxytricha* and *Tetmemena*: (D) 42kD, (E) 22kD, and (F) 57kD, suggest that TBE3 (green) is the ancestral transposon family in *Oxytricha*. For each ORF, 30 protein sequences from each species were randomly subsampled and Maximum Likelihood trees constructed using PhyML (102).

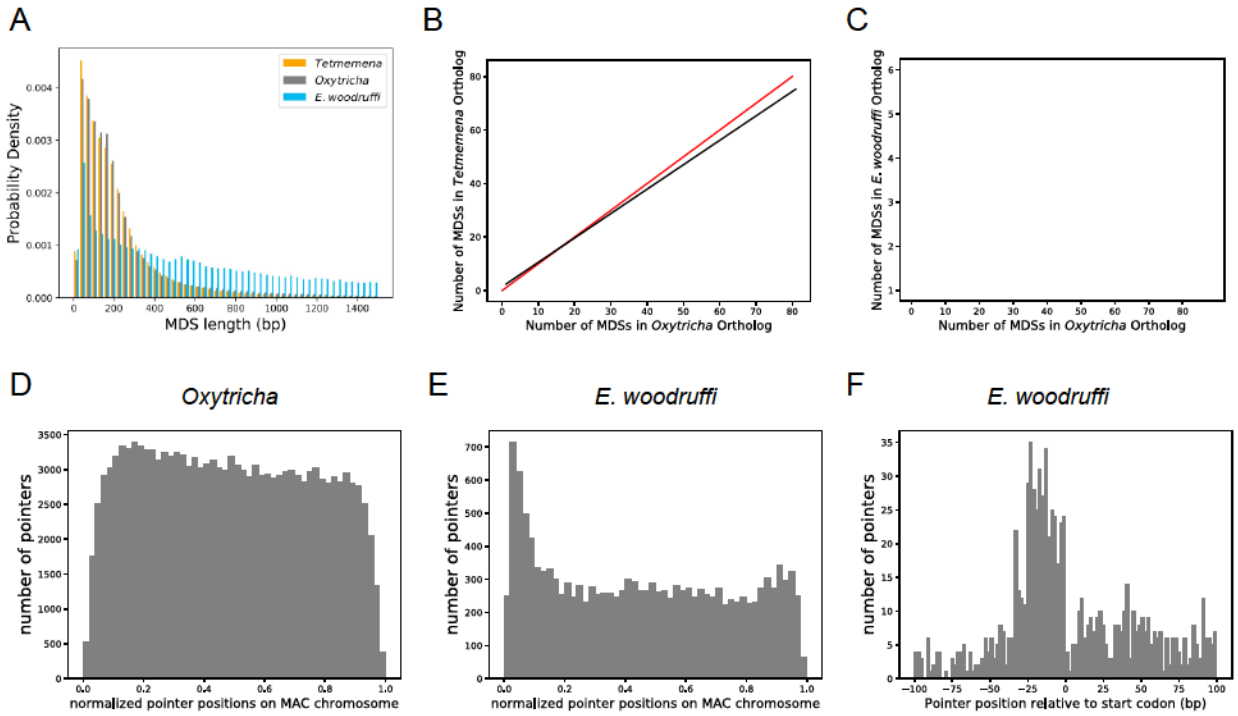


Figure 3. The three MIC genomes are interrupted by IESs at different levels. A) MDSs of *E. woodruffi* are longer compared to *Oxytricha* or *Tetmemena*. B) Positive correlation between the numbers of MDSs for orthologous genes in *Tetmemena* and in *Oxytricha* for 903 single-gene orthologs. Black line is the function of linear regression ($R^2 = 0.75$). Red line is y=x. C) Orthologs in *E. woodruffi* have fewer MDSs compared to *Oxytricha*, with no correlation ($R^2 = 0.003$). Note that many highly discontinuous genes in *Oxytricha* are IES-less in *E. woodruffi* (present on one MDS). 917 single-gene orthologs are shown. D) Distribution of pointers on single-gene MAC chromosomes in *Oxytricha vs.* E) *E. woodruffi*, with MAC chromosomes oriented in gene direction. Pointers significantly accumulate at the 5' end of single-gene MAC chromosomes in *E. woodruffi*. (F) Pointer positions on 3684 two-MDS MAC chromosomes demonstrate a preference upstream of the start codon.

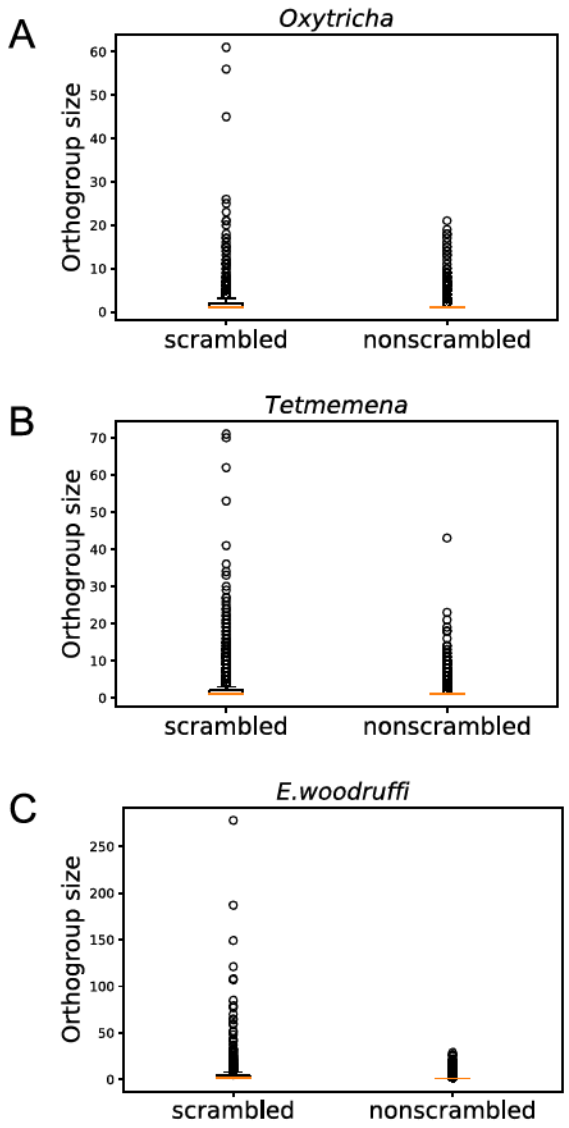
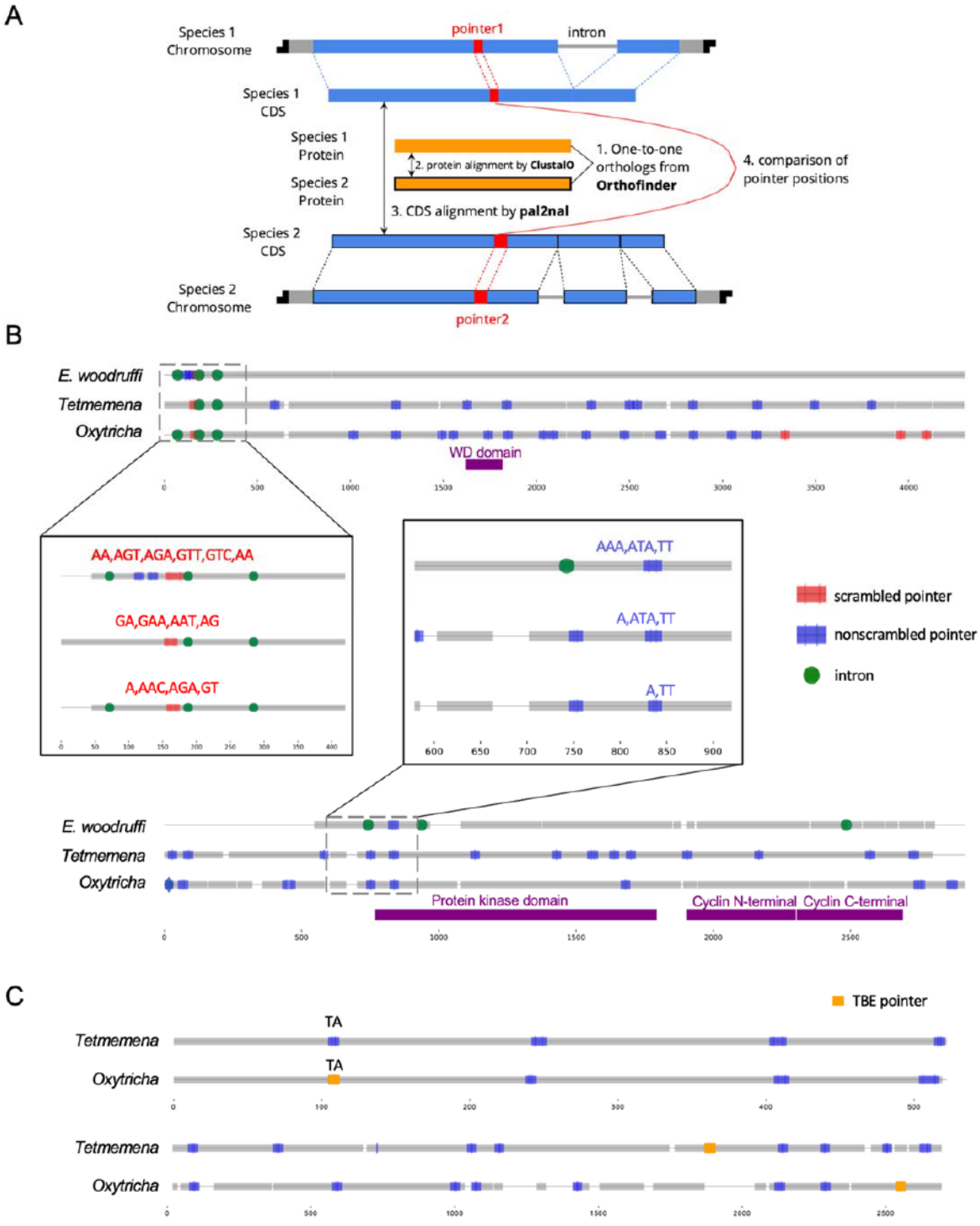


Figure 4. Scrambled genes have more paralogs than nonscrambled genes in the three species. Orthogroups containing at least one scrambled gene ("scrambled") are larger than orthogroups that lack scrambled genes ("nonscrambled") in A) *Oxytricha*, B) *Tetmemena* and C) *E. woodruffi*.



1180

1181 **Figure 5.** Identification and examples of conserved pointers. A) Pipeline for comparison of

1182 pointer positions in orthologs. Orthologs are first grouped by OrthoFinder (103) and protein

1183 sequences of single-copy orthologs aligned by Clustal Omega (104). Then the protein alignments

are reverse translated to coding sequence (CDS) alignments by a modified script of pal2nal (105, Methods). Pointers are annotated on the CDS alignments for comparison between any two orthologs. B) Two examples of pointer conservation across three species. Gray lines represent the alignment of orthologous CDSs and boxes show magnified regions containing conserved pointers. The top panel shows a conserved scrambled pointer (*Oxytricha*: Contig889.1.g68; *Tetmemena*: LASU02015390.1.g1; *E. woodruffi*: EUPWOO_MAC_30105.g1). The bottom panel shows a conserved nonscrambled pointer (*Oxytricha*: Contig19750.0.g98; *Tetmemena*: LASU02002033.1.g1; *E. woodruffi*: EUPWOO_MAC_31621.g1). Pointer sequences are noted and commas indicate reading frame. Protein domains detected by HMMER (106) are marked in purple. C) Examples of TBE insertions in nonscrambled IESs. The upper pair of sequences show an *Oxytricha* TBE pointer (orange insertion of an incomplete TBE2 transposon containing the 42kD and 57kD ORFs) conserved with a *Tetmemena* non-TBE pointer (*Oxytricha*: Contig736.1.g130; *Tetmemena*: LASU02012221.1.g1). Both species have a TA pointer at this junction. The bottom pair of sequences illustrate a case of nonconserved TBE pointers (*Oxytricha*: Contig17579.0.g71; *Tetmemena*: LASU02007616.1.g1).

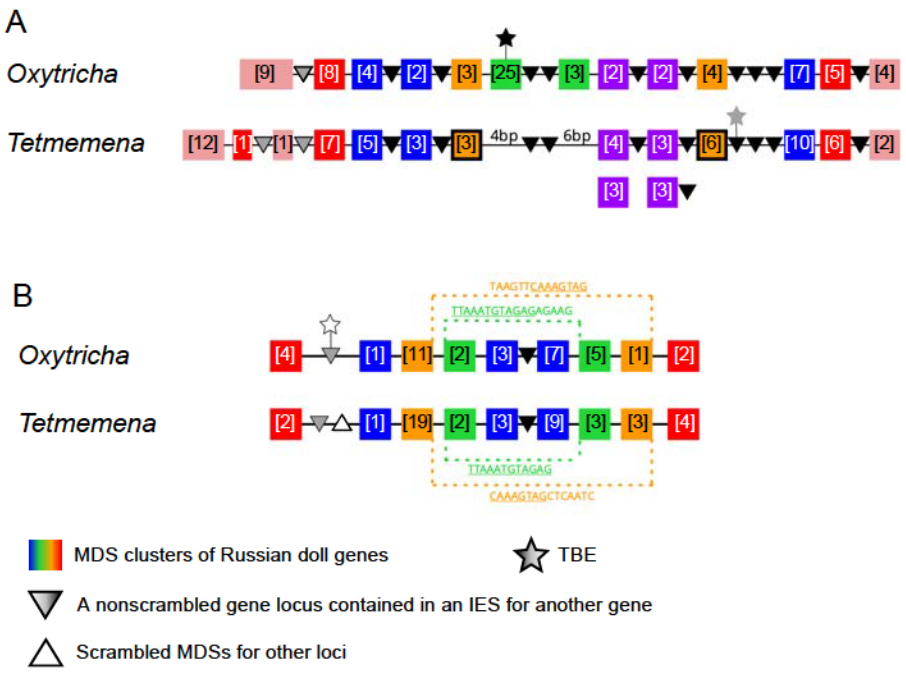


Figure 6. Synteny in "Russian doll" loci in *Oxytricha* and *Tetmemena*. A) Schematic comparison of the Russian doll gene cluster on *Oxytricha* MIC contig OXYTRI_MIC_87484 vs. *Tetmemena* MIC contig TMEMEN_MIC_21461. Boxes of the same color represent clusters of MDSs for orthologous genes (detailed map in Figure 6 - figure supplement 1 and Figure 6 - figure supplement 2). Numbers in brackets indicate the number of MDSs in each cluster, grouped by MAC chromosome. One nested gene (green) in *Oxytricha* is absent from *Tetmemena*. A two-gene chromosome (orange) that derives from seven MDSs in *Oxytricha* is processed as two single-gene chromosomes in *Tetmemena* instead (indicated by black border around orange boxes). The purple gene in *Oxytricha* has two paralogs in *Tetmemena*. Black triangles represent conserved, orthologous, nonscrambled gene loci inserted between nested Russian doll genes. Empty triangle represents scrambled MDSs for other loci. Gray triangles, complete nonscrambled MAC loci embedded between gene layers in one species with no orthologous gene detected in the other species. Black star, a complete TBE transposon insertion. Gray star, a

1213 partial TBE insertion. B) *Oxytricha* MIC contig OXYTRI_MIC_69233 vs. *Tetmemena* MIC
1214 contig TMEMEN_MIC_22886. Pointer sequences bridging the nested MDSs of orange and
1215 green genes are highlighted. The underlined pointer portions are conserved between species, e.g.
1216 the last 8bp of the *Oxytricha* pointer, TAAGTTCAAAGTAG, are identical to the first 8bp of
1217 CAAAGTAGCTCAATC in *Tetmemena*, illustrating pointer sliding (36), or gradual shifting of
1218 MDS/IES boundaries. White star indicates a decayed TBE with no ORF identified.
1219

Figure 2 - figure supplement 1. Comparison of MIC genome context of A) TBE/Tec transposons and B) other transposable elements in the three species. Complete and partial TBE/Tec elements were annotated by MIC context. Other transposable elements include all subcategories shown in Supplementary File 2. Boundary (light blue): edges of assembled MIC contigs. MIC-specific contig (orange): no MDS identified on the MIC contig so it cannot be annotated as intergenic or a long IES. Intergenic (green): MIC regions between MDSs for different MAC contigs. IES paralogous (yellow): TE insertions between duplicate (paralogous)MDSs, so they are neither scrambled nor nonscrambled. IES nonscrambled (dark blue): TE insertions that map between consecutive, nonscrambled MDSs for the same MAC contigs. IES scrambled (magenta): MIC regions between nonconsecutive (scrambled) MDSs for the same MAC contig. Note that TEs in IESs or intergenic regions could be flanked by other MIC-limited sequences extending beyond the TE ends.

Figure 2 - figure supplement 2. Length distribution of assembled MAC nanochromosomes in the three species. Chromosomes over 11 kb are excluded from the plot.

Figure 3 - figure supplement 1. A and B) CDS lengths correlate for *Oxytricha*, *Tetmemena* and *E. woodruffi* orthologs (related to Figure 3). A) *Tetmemena* CDS length positively correlates with that of *Oxytricha* orthologs (R^2=0.96). Black line is the linear regression fitting function. Red line shows y=x. B) *E. woodruffi* CDS length positively correlates with that of *Oxytricha* orthologs (R^2=0.83). C) The distribution of pointers on single-gene MAC chromosomes in *Tetmemena* displays a weak 5' bias (related to Figure 3).

Figure 3 - figure supplement 2. Scrambled and nonscrambled loci have distinct length distributions of IESs and pointers. A-C) Length distribution of scrambled and nonscrambled pointers <= 30bp in A) *Oxytricha*, B) *Tetmemena* and C) *E. woodruffi*. D- F) Length distribution of scrambled and nonscrambled IESs in D) *Oxytricha* (<=100bp), E) *Tetmemena* (<=100bp) and F) *E. woodruffi* (<=300bp).

Figure 4 - figure supplement 1. An example of an *E. woodruffi* scrambled gene locus containing paralogous MDSs. A) The upper panel is the map of a scrambled MIC locus (EUPWOO_MIC_17325). Below is the corresponding map of the MAC chromosome (EUPWOO_MAC_29939). Pointers between MDSs are labeled above or below the MAC contig (nonscrambled pointer length in blue and scrambled pointers labeled in red). B) A model for the evolutionary origin of this scrambled MIC locus by partial duplication and subsequent decay. Stage 1: The ancestral MIC locus contains three nonscrambled MDSs (labeled proto-MDSs because they are precursors for the modern state). Stage 2: The region containing two proto-MDSs duplicated in the MIC genome. Stage 3: Nucleotide substitutions accumulated in both paralogous copies at different positions (shown in gray dashed boxes) leading to the fixation of some regions as MDSs, while the regions that accumulated more mutations decayed into IESs, which are removed during genome rearrangement. (Figure 4 - figure supplement 1B has been adapted from a general model in Figure 3 from Gao et al., 2015 [43].)

[43]: Gao, Feng, Scott W. Roy, and Laura A. Katz. "Analyses of alternatively processed genes in ciliates provide insights into the origins of scrambled genomes and may provide a mechanism for speciation." *MBio* 6.1 (2015): e01998-14.

Figure 4 - figure supplement 2. The trend of scrambled loci to contain odd-even patterns may arise from partial duplication followed by mutation accumulation. A) A diagram describing a typical scrambled region with an odd-even pattern. We propose that the IES (S1) between MDS *n* and MDS *n+2* may be ancestrally paralogous to MDS *n+1* (S2) which evolved by duplication of MDS *n+1* before it was scrambled. S1 and S2 would therefore be homologous in this model. B) The lengths of modern IES (S1) and MDS (S2) display a strong positive correlation in *E. woodruffi* (504 pairs). Many data points fall on the y=x (red line). All MDS and IES pairs were only considered if they are on the same MIC contig, to exclude alleles. C) Character mapping of scrambled loci on a phylogeny: 1. Examples of scrambled loci uniquely present in one species (only showing for *Oxytricha* and *Tetmemena*; most scrambled genes in *E. woodruffi* have no ortholog detectable in the other two species, possibly because the long genetic distance obscured homology, see main text and Supplementary File 4); 2. Scrambled loci shared between *Oxytricha* and *Tetmemena*, but not *E. woodruffi*; 3. Scrambled loci shared in three species. The lengths of IES (S1) and MDS (S2) in typical odd-even regions display a moderately positive correlation in *Oxytricha* (D) and *Tetmemena* (E). Newer scrambled loci correlate more strongly. Red line represents y=x. Note that S1 and S2 are flanked by identical pointers, *a* and *b*, in all annotated pairs.

Figure 4 - figure supplement 3. Expression level of scrambled and nonscrambled genes in A) *Oxytricha*, B) *Tetmemena* and C) *E. woodruffi*. P-values of Mann-Whitney U tests are shown in blue. The line in orange shows the median. The box shows the range between the first and third quartiles. The upper whisker represents the third quartile + 1.5 × interquartile range (IQR) and the lower whisker shows the first quartile – 1.5 × IQR. Numbers in brackets indicate genes which have a coefficient of variation of TPM (transcripts per million) less than 1.

Figure 5 - figure supplement 1. Examples of Intron-IES conversion across three species. A) Four intron positions in *E. woodruffi* (orange boxes in magnified regions) overlap locations of nonscrambled pointers in the orthologous genes in *Oxytricha* and *Tetmemena* (*Oxytricha*: Contig13378.0.g40; *Tetmemena*: LASU02004100.1.g1; *E. woodruffi*: EUPWOO_MAC_08218.g1) consistent with a possible trend of some ancestral introns becoming IESs in the hypotrich lineage. Two positions fall within a conserved protein domain of unknown function (DUF3591). B) An orthologous gene with two intron-IES conversions in reciprocal directions (*Oxytricha*: Contig16930.0.g77; *Tetmemena*: LASU02013377.1.g1; *E. woodruffi*: EUPWOO_MAC_15089.g1). Colors and annotation as in Figure 5.

Figure 6 - figure supplement 1. Detailed illustration of both Russian Doll regions in Figure 6. MDS indices are annotated here for each MAC locus. Overlined numbers represent inverted MDSs. MAC contig numbers for the MDSs are listed below and shown in corresponding color patterns (the *Oxytricha* loci were previously characterized in ref. 60).

Figure 6 - figure supplement 2. Details of the Russian Doll region in *Tetmemena* (TMEMEN_MIC_21461, Figure 6A). The whole region (~50 kb) was validated by 11 PCRs. The two black arrows indicate the absence of a Russian doll gene (green in Figure 6A) that is present in *Oxytricha*. Legend lists the 20 *Tetmemena* MAC contigs that contain the corresponding MDSs.

Supplementary File 1. Sequencing depth statistics for MIC genome assemblies

*Sequencing data from Chen et al. (1).

**Raw reads were mapped to the MIC genome assembly by Minimap2 and Bowtie2 (97). Average coverage was calculated with BBmap (sourceforge.net/projects/bbmap/) pileup.sh for MDS-containing contigs in the MIC genome assembly.

Supplementary File 2. Subcategories of repeat content in the three species.

Repeat content of the three genomes, as annotated by Repeatmasker (99) with additional manual annotation of TBE/Tec elements. The numbers may differ from Figure 2A-C because some repeats are assigned as other MIC categories in the pie charts (Methods). For example, a MIC region which is both an IES and satellite, is assigned as IES in Figure 2A-C, but is counted as a satellite in this table.

Supplementary File 3. TBE/Tec ORFs in three species

* Differs from 10,109 in Chen et al. (44) because we used different versions of BLAST and custom python scripts to identify complete TBEs (See Methods).

Supplementary File 4. Orthology among scrambled and nonscrambled genes in the three species

* Ciliate database is generated by extracting all protein sequences in phylum Ciliophora (taxid: 5878) from NR database.

Supplementary File 5. Summary of orthologs in each pair of species

The (*i*,*j*) cell shows the number of genes in species *i* with an ortholog in species *j*.

* Genes with no ortholog detected by OrthoFinder (103) in the other two species.

Supplementary File 6. More scrambled MAC contigs contain at least one paralogous MDS that may be involved in alternative rearrangement.

Supplementary File 7. MDS-IES pairs share homologous sequences in the three species (related to Figure 4 - figure supplement 2).

Supplementary File 8. Genes with expression support in the three species

Supplementary File 9. Presence of conserved pointers in three species, with Monte Carlo simulations

Supplementary File 10. Scrambled pointers are more conserved than nonscrambled pointers.

Supplementary File 11. Most pointers conserved in position are different in sequence

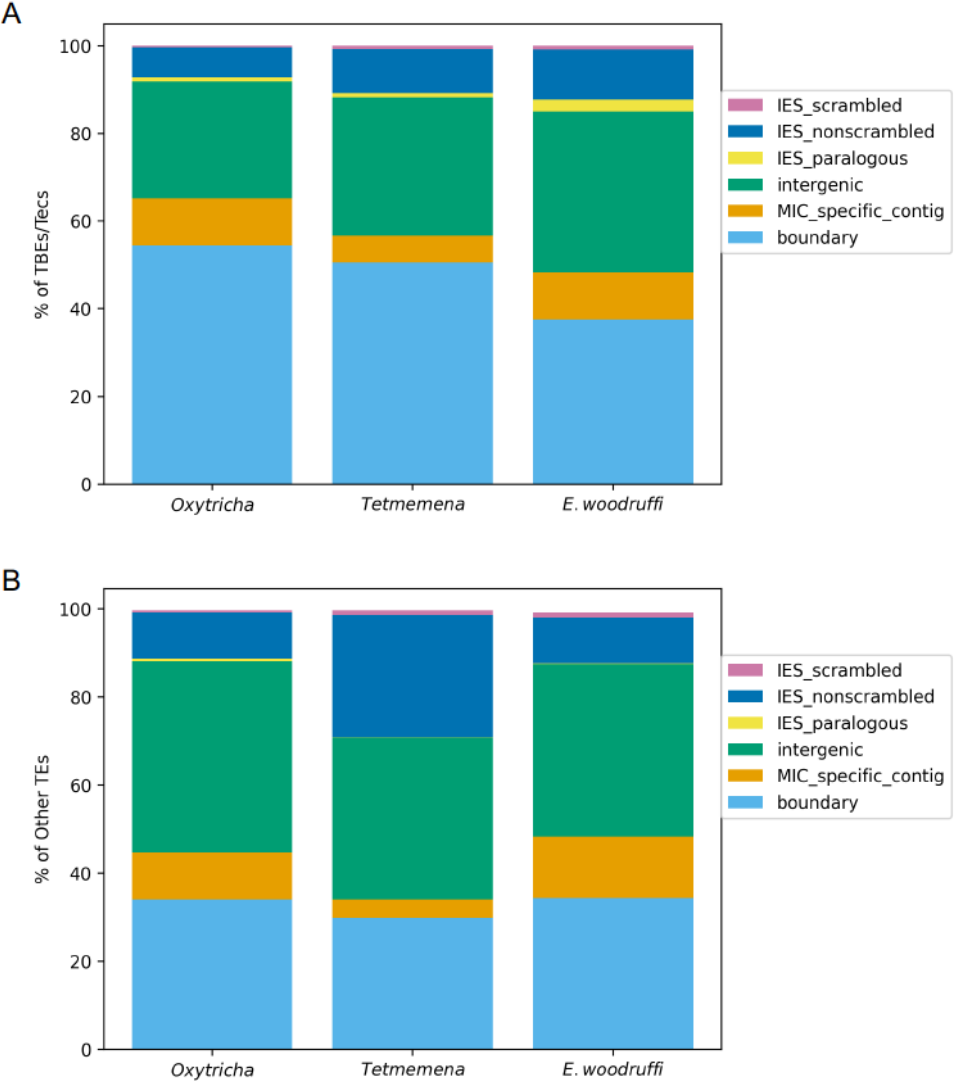
Supplementary File 12. Intron-IES conversion comparison in three species and Monte Carlo simulations

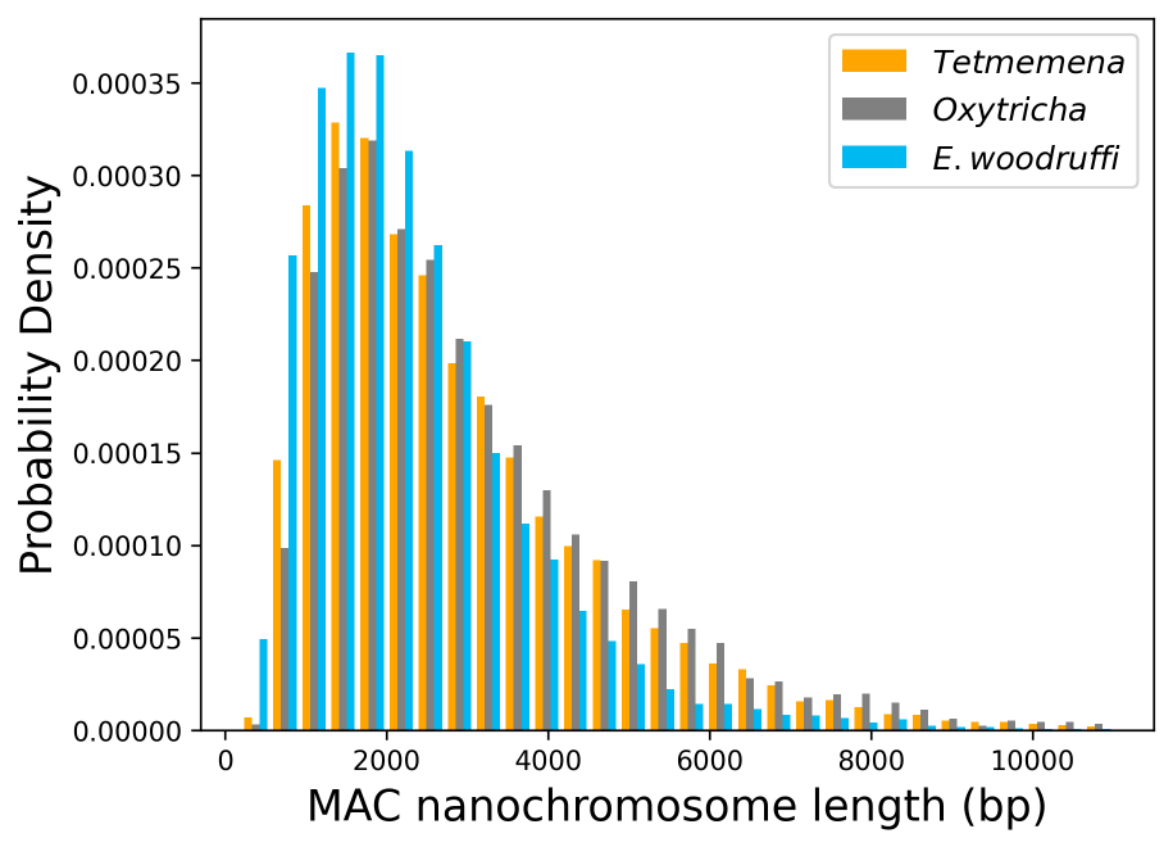
Supplementary File 13. Pairwise intron-IES conversion comparisons and Monte Carlo simulations

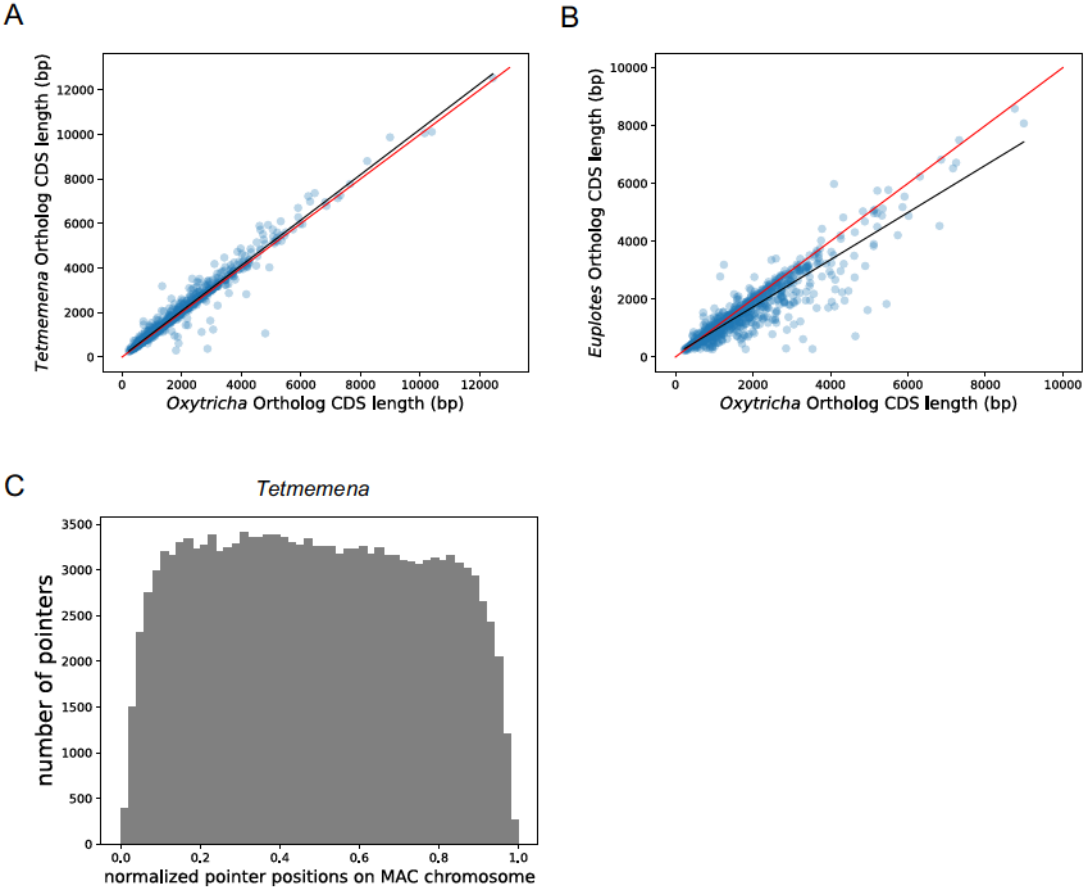
Supplementary File 14. PCR primers for validation of the Russian doll region in *Tetmemena* MIC DNA (Figure 6A)

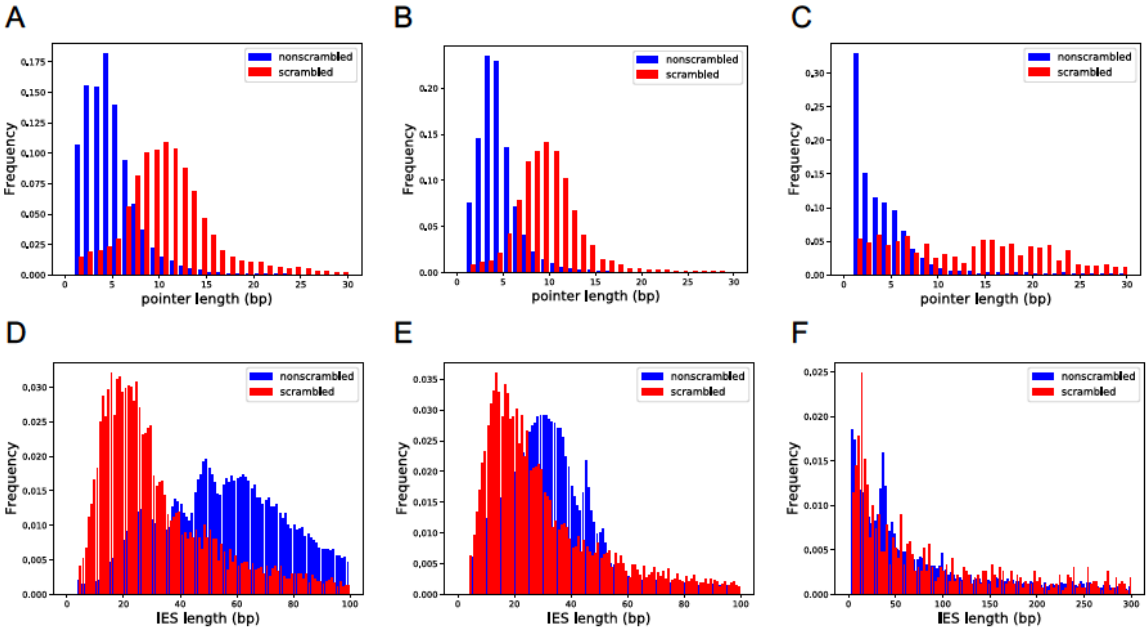
Figure 5 - source data 1. Pointers conserved in all three species.

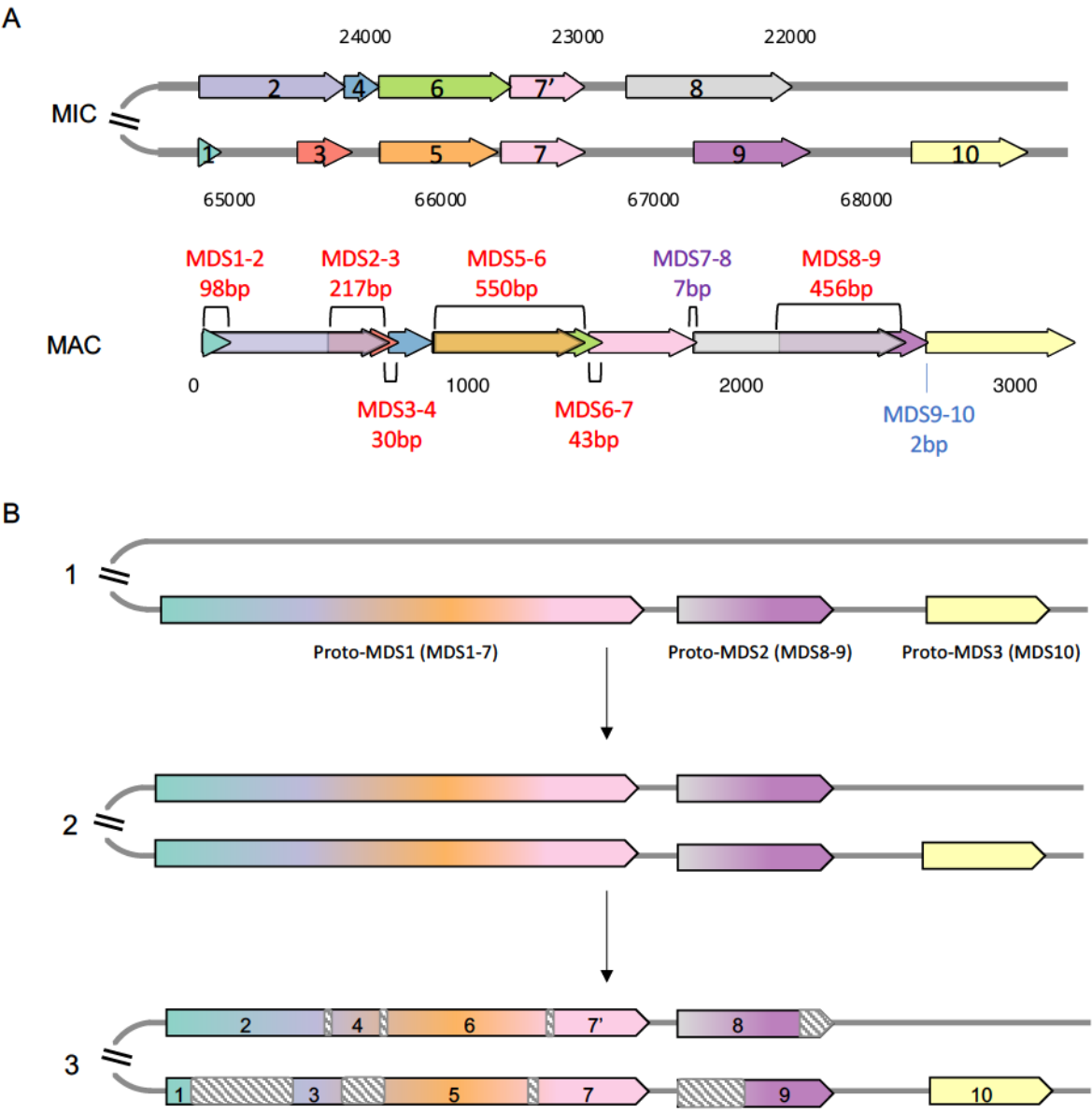
Figure 5 - source data 2. The TBE pointers in *Oxytricha* that are conserved with non-TBE pointers in *Tetmemena.*

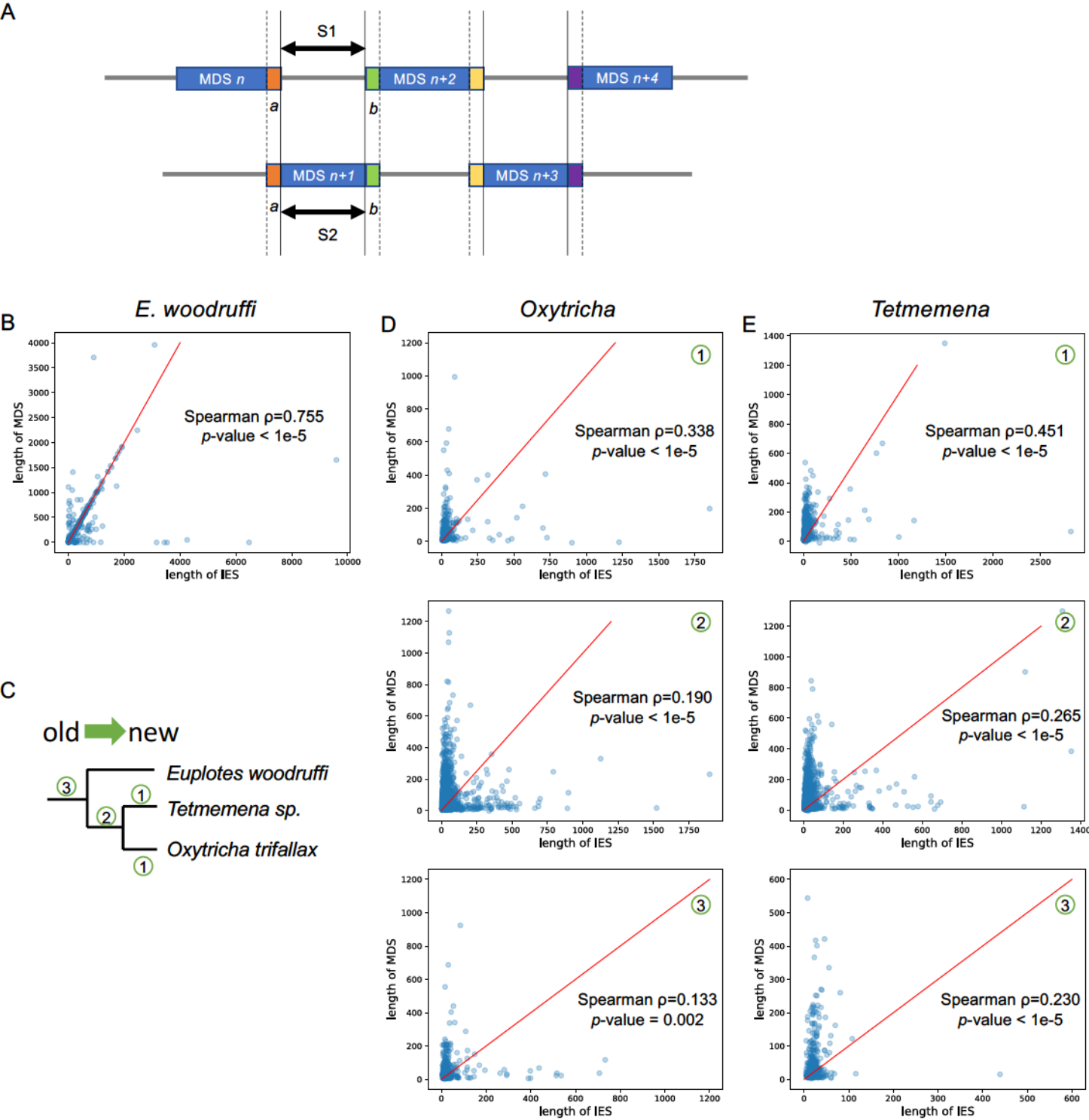

A
% of TBEs/Tecs
100
80
60
40
20
0
Oxytricha
Tetmemena
E. woodruffi
IES_scrambled
IES_nonscrambled
IES_paralogous
intergenic
MIC_specific_contig
boundary
B
% of Other TEs
100
80
60
40
20
0
Oxytricha
Tetmemena
E. woodruffi
IES_scrambled
IES_nonscrambled
IES_paralogous
intergenic
MIC_specific_contig
boundary

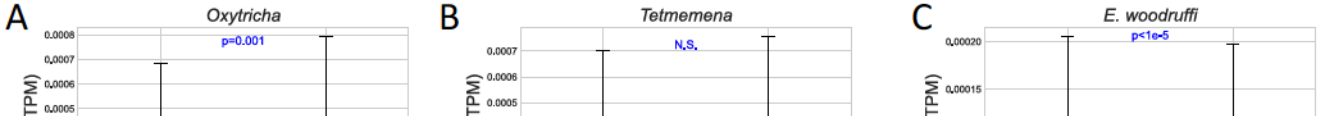

Tetmemena
Oxytricha
E. woodruffi
Probability Density
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2000
4000
6000
8000
10000
MAC nanochromosome length (bp)


A
Tetmemena Ortholog CDS length (bp)
Oxytricha Ortholog CDS length (bp)
B
Euplotes Ortholog CDS length (bp)
Oxytricha Ortholog CDS length (bp)
C
Tetmemena
number of pointers
normalized pointer positions on MAC chromosome

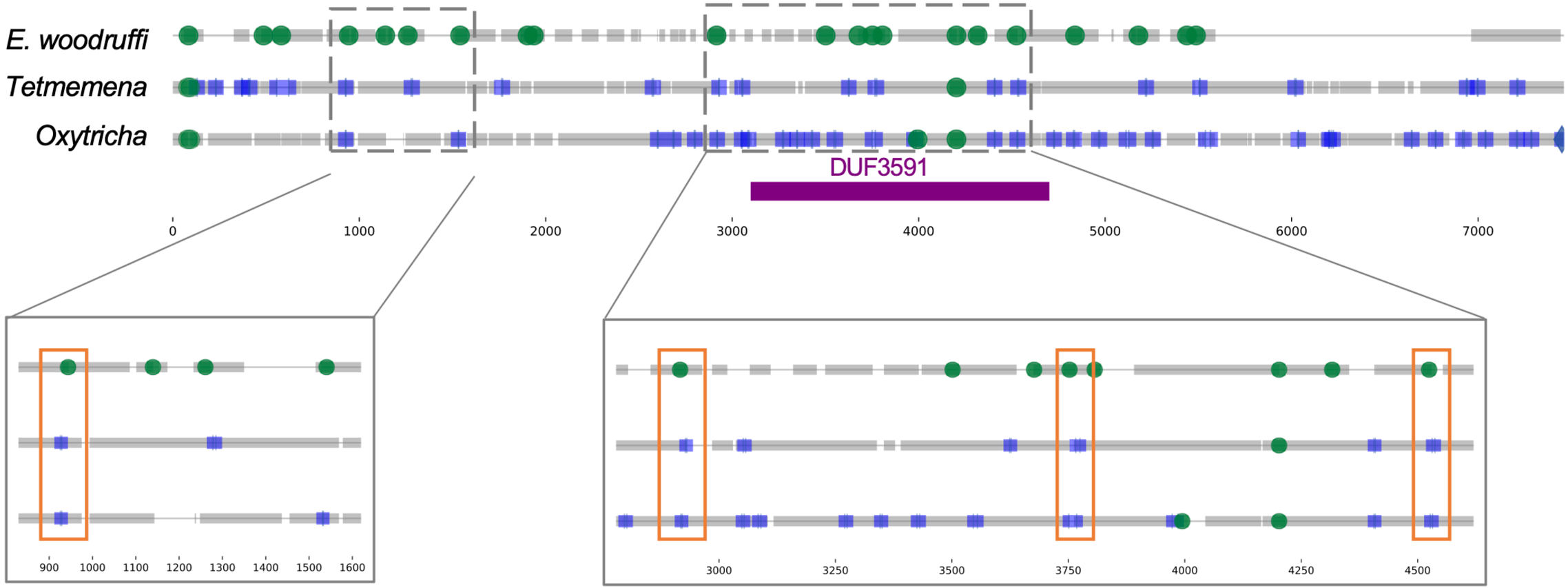

A
B
C
D
E
F
nonscrambled
scrambled
Frequency
pointer length (bp)
IES length (bp)


A
24000
23000
22000
MIC
2
4
6
7'
8
1
3
5
7
9
10
65000
66000
67000
68000
MDS1-2
98bp
MDS2-3
217bp
MDS5-6
550bp
MDS7-8
7bp
MDS8-9
456bp
MAC
0
1000
2000
3000
MDS3-4
30bp
MDS6-7
43bp
MDS9-10
2bp
B
1
Proto-MDS1 (MDS1-7)
Proto-MDS2 (MDS8-9)
Proto-MDS3 (MDS10)
2
3
2
4
6
7'
8
1
3
5
7
9
10

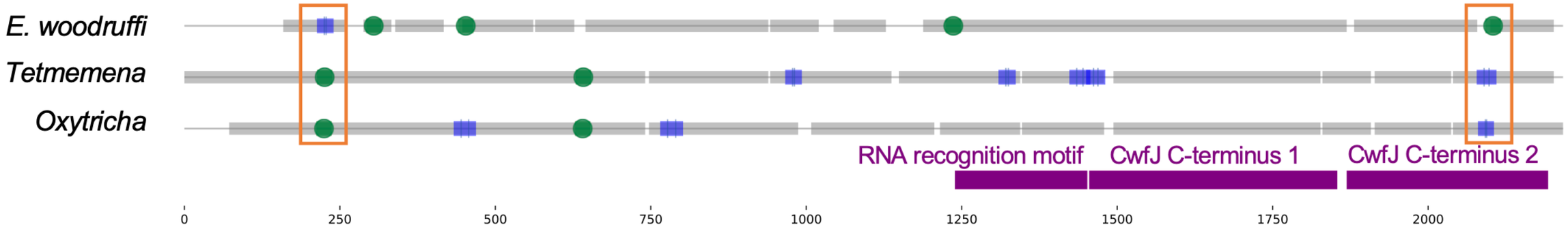

A
S1
MDS n
MDS n+2
MDS n+4
a
b
MDS n+1
MDS n+3
a
b
S2
B
E. woodruffi
Spearman ρ=0.755
p-value < 1e-5
length of MDS
length of IES
D
Oxytricha
①
Spearman ρ=0.338
p-value < 1e-5
②
Spearman ρ=0.190
p-value < 1e-5
③
Spearman ρ=0.133
p-value = 0.002
E
Tetmemena
①
Spearman ρ=0.451
p-value < 1e-5
②
Spearman ρ=0.265
p-value < 1e-5
③
Spearman ρ=0.230
p-value < 1e-5
C
old
new
Euplotes woodruffi
Tetmemena sp.
Oxytricha trifallax


A
Oxytricha
p=0.001
RNAseq (TPM)
0,0008
0,0007
0,0006
0,0005
0,0004
0,0003
0,0002
0,0001
0,0000
scrambled
(3062)
nonscrambled
(16636)
B
Tetmemena
N.S.
RNAseq (TPM)
0,0007
0,0006
0,0005
0,0004
0,0003
0,0002
0,0001
0,0000
scrambled
(3252)
nonscrambled
(20370)
C
E. woodruffi
p<1e-5
RNAseq (TPM)
0,00020
0,00015
0,00010
0,00005
0,00000
scrambled
(1633)
nonscrambled
(25839)

A


E. woodruffi
Tetmemena
Oxytricha
DUF3591
0
1000
2000
3000
4000
5000
6000
7000
900
1000
1100
1200
1300
1400
1500
1600
3000
3250
3500
3750
4000
4250
4500

B


E. woodruffi
Tetmemena
Oxytricha
RNA recognition motif
CwfJ C-terminus 1
CwfJ C-terminus 2
0
250
500
750
1000
1250
1500
1750
2000

A

Oxytricha

1 - 4 | 6 - 7 | 9 | 11 - 12 | 2̄ - 1̄ | 11̄ - 3̄ (odd) | 4̄ | 2̄ - 1̄ | 6 | 8 | 10 - 11

3̄ - 1̄ | 3̄ | 1̄ | 1 - 7 | 5 - 7 | 25̄ - 1̄ | 13̄ - 1̄ | 5̄ - 1̄ | 28̄ - 26̄ | 4̄ - 3̄

1̄ | 2̄ - 1̄ | 1 - 6 | 2 - 4 | 1 | 3̄ - 1̄ | 1 - 14 | 1 | 4̄ | 2̄ | 13̄ - 12̄

9̄ - 5̄ (odd) | 6 - 12 (*even*) | 13 | 5̄ | 13̄ | 10̄ | 8̄ | 5̄

MAC lists

11893.0 | 11684.0 | 17142.0 | 19333.0 | 12294.0 | 12507.0 | 16759.0 | 17089.0 | 12730.0

2933.0 | 1447.1 | 18346.0 | 2924.0 | 18809.0 | 17915.0 | 14396.0 | 16323.0

1 - 4 | 6 - 8 | 10 - 14 | 3̄ | 10̄ - 1̄ | 15 | 8̄ - 1̄ | 11̄ - 10̄ | 8̄ - 7̄ | 4̄ | 2̄ - 1̄

10 – 14 (*even*) | 15 - 16 | 2̄ - 1̄ | 4̄ | 2̄ - 1̄ | 1 - 9 | 3̄ - 1̄ | 20̄ - 1̄ | 5̄ - 1̄

Tetmemena

7̄ - 4̄ | 1̄ | 3̄ - 1̄ | 1 - 3 | 5 - 6 | 1 - 4 | 7̄ - 1̄ | 1 - 17 | 1 - 5 | 6̄ - 5̄ | 3̄

6̄ - 4̄ | 3̄ - 1̄ | 1 - 2

18̄ - 17̄ | 13̄ - 9̄ (odd) | 8̄ - 7̄ | 5 | 6 | 9 | 12 - 14 | 7̄ | 9̄ | 5̄

MAC lists

10808 | 04166 | 02638 | 00564 | 09313 | 11844 | 04978 | 00618 | 04942

08167 | 08901/11158 | 00880 | 07156/03489 | 05053 | 05069 | 02448 | 03934

B

7 - 9

Oxytricha

1 - 10 | 6̄ - 3̄ | 1 - 11 | 1̄ | 1 - 11 | 6 - 7 | 7̄ - 3̄ (odd) | 1 - 9 | 2 – 6 (*even*)

8 – 11 | 5̄ - 1̄ | 12̄ | 1 - 2

MAC lists

10733.0 | 20394.0 | 4520.0 | 18903.0 | 14894.0 | 9031.0 | 20347.0

1 - 3

1 | 3

2 - 6

1 - 10

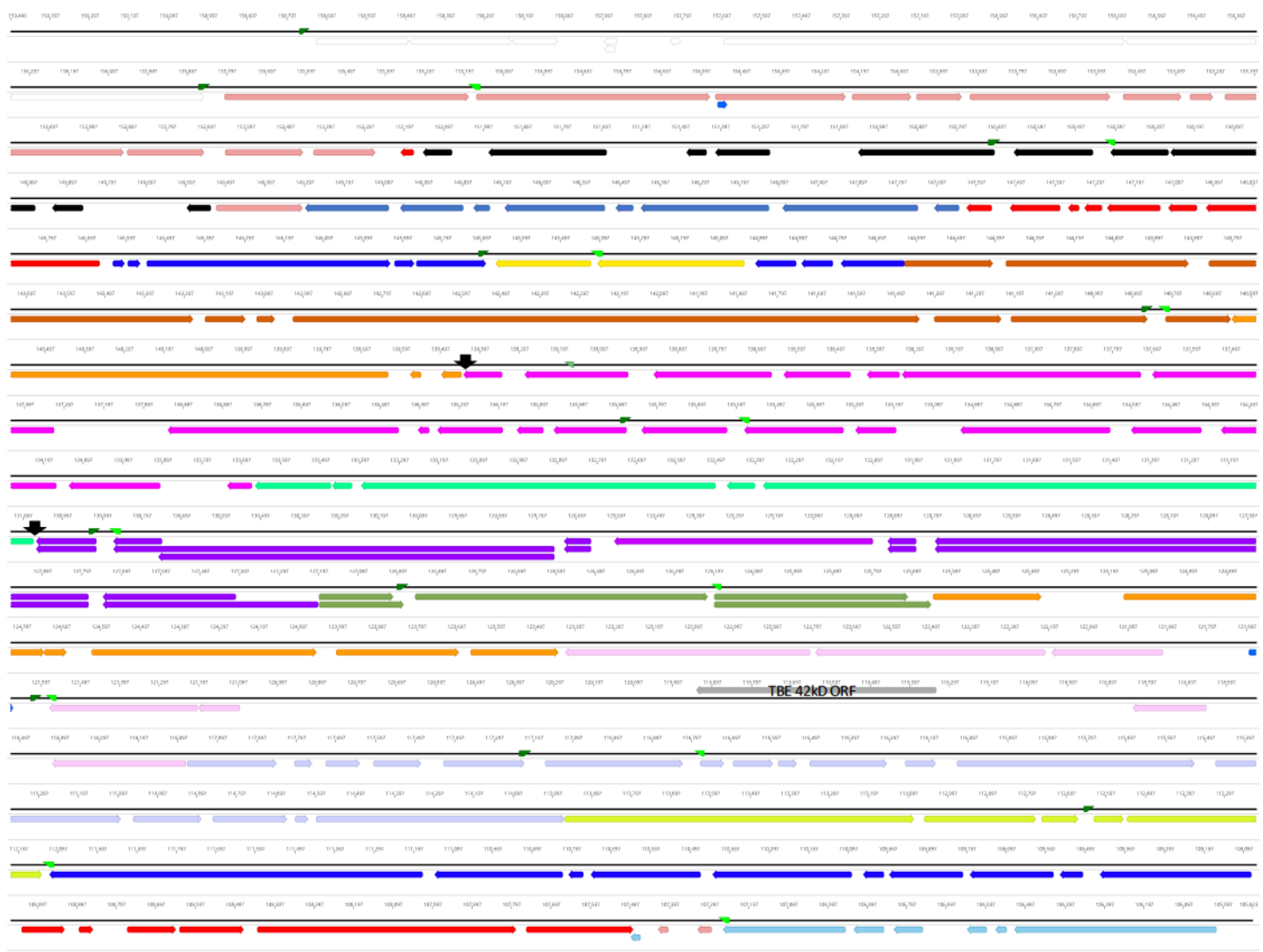
Tetmemena

1 | 4 - 11 | 6̄ - 5̄ | 1̄ | 2 - 3 | 4 - 6 | 1̄ | 1 - 19 | 4 - 5 | 8̄ | 5̄ | 3̄

1 - 12 | 2 – 6 (*even*) | 7 | 9 – 13 | 3̄ - 1̄ | 22̄ - 20̄ | 1 - 4

MAC lists

10239 | 02366 | 02365 | 13678 | 18935 | 16334 | 08537 | 15889 | 07670 | 19678


TBE 42kD ORF
Russian doll genes
10808
04166
09313
00618 (139,443 – 140,646 bp)
03666 (123,438 – 125,641 bp)
08901/11158
Other genes
02638
00564
11844
04978
04942
08167
00880
07156/03489
05053
05069
02448
03934
PCR Primers