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

Yi Feng¹, Rafik Neme¹,², Leslie Y. Beh¹, Xiao Chen³, Jasper Braun⁴,⁵, Michael Lu¹, Laura F. Landweber¹*

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
is arguably the most complex of any model organism (1), requiring programmed deletion of over
90% of the germline DNA during development and massive descrambling of the resulting
fragments to construct a new MAC genome of over 18,000 chromosomes (10). This differs from
the distantly related Tetrahymena and Paramecium that both eliminate ~30% of the germline
genome (27, 28). Paramecium uses exclusively 2 bp pointers and lacks evidence of any
scrambled loci. A small number of scrambled loci (4 confirmed out of 2711 candidates) have
been reported in Tetrahymena (30, Figure 1B). Tetrahymena and Paramecium diverged from
Oxytricha over one billion years ago (31, 32), which leaves a large gap in our understanding of
the emergence of complex DNA rearrangements in the Oxytricha lineage.

Open questions include how did the Oxytricha germline genome acquire its high number
of IES insertions and how do scrambled loci arise and evolve. Three previous studies tackled
these questions at the level of single genes and orthologs, including DNA polymerase α, actin I
and TEBPα (33, 34, 35, 36). Here, we provide the first comparative genomic analysis of
Oxytricha trifallax and two other spirotrichous ciliates, Tetmemena sp. and Euplotes woodruffi.
Tetmemena sp. is a hypotrich similar to Tetmemena pustulata, formerly Stylonychia pustulata
(7), in the same family as Oxytricha trifallax (Figure 1B, 1, 7). Hypotrichs are noted for the
presence of scrambled genes, based on previous ortholog comparisons (7, 33, 34, 35, 36 Figure
1B). E. woodruffi, together with the hypotrichous ciliates, belong to the class Spirotrichea
(Figure 1B). Like hypotrichs, Euplotes also has gene-sized nanochromosomes in the MAC
genome (8, 9, 37), but this outgroup uses a different genetic code (UGA is reassigned to cysteine,
ref. 38) and little is known about its MIC genome. A partial MIC genome of Euplotes vannus
was previously assembled, and it contains highly conserved TA pointers (9), consistent with
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).
*Oxytricha* contains three families of TBEs. TBE3 appears to be the most ancient among hypotrichs, based on previous analysis of limited MIC genome data (44). We constructed phylogenetic trees using randomly subsampled TBE sequences for all three ORFs from *Oxytricha* and *Tetmemena* (Figure 2D-F). This confirmed that only TBE3 is present in the *Tetmemena* MIC genome, as proposed in (44). This also suggests that TBE1 and TBE2 expanded in *Oxytricha* after its divergence from other hypotrichous ciliates. As illustrated in Figure 2 - figure supplement 1, the MIC genome contexts of TBEs in *Oxytricha* and *Tetmemena* are similar, with many TE insertions within IESs, consistent with either IESs as hotspots for TE insertion or with the model (49) that some TE insertions may have generated IESs, as demonstrated in *Paramecium* (29, 52). Subsequent sequence evolution at the edges of IES/MDS pointers (36) can give rise to boundaries that no longer correspond precisely to TBE ends. For further discussion of the conservation of TBE locations, see the section, “*Oxytricha* and *Tetmemena* share conserved rearrangement junctions” below.

Additionally, Repeatmodeler/Repeatmasker identified that *Oxytricha* has more MIC repeats in the “Other” category than *Tetmemena* or *E. woodruffi* (Figure 2, subcategories of repeat content in Supplementary File 2). 214 Mbp of the *Oxytricha* MIC genome (43%, which is greater than 35.9% reported in ref. 1 that used earlier versions of the software) is considered repetitive (including TBEs, tandem repeats and other repeats in Figure 2), versus 31.7 Mbp for *Tetmemena* (13.4%) and 28.5 Mbp (16.8%) for *E. woodruffi*. *Oxytricha*’s additional ~180 Mbp in repeat content partially explains the significantly larger MIC genome size of *Oxytricha* versus 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, 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 $\rho=0.755$, $p<1e^{-5}$, 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 $\rho$ shown in Figure 4 - figure supplement...
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
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 (TATATCTATACCTTAACTT) 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*. 

18
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
protein, ref. 1) is similarly complex, with 195 scrambled junctions. Together, these observations are consistent with our interpretation that *E. woodruffi* reflects an evolutionary intermediate stage, as it contains both fewer scrambled loci and fewer scrambled junctions within its scrambled loci. The observation that the most scrambled locus differs in each species is also consistent with the conclusion that complex gene architectures may continue to elaborate independently.

We observed that scrambled genes in each species tend to have more paralogs than nonscrambled genes. Similarly, in *Chilodonella uncinata* (18), a distantly-related ciliate in the class *Phyllopharyngea* that also has scrambled genes, scrambled gene families (orthogroups) contain more genes (~2.9) than nonscrambled gene families (~1.3) (18). Apart from duplications at the gene level, *E. woodruffi* often contains partial MDS duplications at scrambled junctions, annotated as unusually long “pointers” (Figure 4 - figure supplement 1). We also demonstrate that odd-even scrambled patterns could readily arise from local duplications (Figure 4 - figure supplement 2). These observations are most consistent with a simple model (43) in which local duplications permit combinatorial DNA recombination between paralogous germline regions, and mutation accumulation in either paralog establishes an odd-even scrambled pattern that can propagate by weaving together segments from paralogous sources. Other proposed models include Hoffman and Prescott’s (66) IES-invasion model that suggested that pairs of IESs could invade an MDS, and then subsequently recombine with another IES to yield odd-even scrambled regions; however, a previous examination did not find support for this model (34). Prescott et al. (67) also proposed that some odd-even scrambled loci could arise suddenly via reciprocal recombination with loops of A/T-rich DNA, but this does not exploit paralogy, only the high A/T 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$_2$HPO$_4$, 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
MAC genome sequencing. Paired-end sequencing was performed on an Illumina Hiseq2000 at the Princeton University Genomics Core Facility. MIC DNA was enriched from whole cell DNA and sequenced via three sequencing platforms (Illumina, Pacific Biosciences and Oxford Nanopore Technologies). We used conventional and pulse-field gel electrophoresis to enrich MIC DNA:

1) High-molecular-weight DNA was separated from whole cell DNA by gel-electrophoresis (0.25% agarose gel at 4 °C, 120 V for 4 hr). The top band was cut from the gel and purified with the QIAGEN QIAquick kit. The purified high-molecular-weight DNA was directly sent to the group of Dr. Robert Sebra at the Icahn School of Medicine at Mount Sinai for library construction and sequencing. BluePippin (Sage Science) separation was used before sequencing to select DNA >10kb. DNA was sequenced on two platforms: Illumina HiSeq2500 (150 bp paired-end reads) and PacBio Sequel (SMRT reads).

2) High-molecular-weight DNA was also enriched by pulsed-field gel electrophoresis (PFGE). *E. woodruffi* cells were mixed with 1% low-melt agarose to form plugs according to Akematsu et al. (73), with addition of 1 hr incubation with 50 μg/ml RNase (Invitrogen AM2288) in 10 mM Tris-HCl (pH7.5) at 37 °C for RNA depletion. After three washes of 1 hr with 1x TE buffer, the DNA plugs were incubated in 1mM PMSF to inactivate proteinase K, followed by MspJI (New England Biolabs) digestion at mCNNR(9/13) sites to remove contaminant DNA (mC indicates C5-methylation or C5- hydroxymethylation). Previous reports have shown that no methylcytosine is detectable in vegetative cells of *Oxytricha* (74), *Tetrahymena* (75) and *Paramecium* (76), suggesting that C5-methylation and C5- hydroxymethylation are rarely involved in the vegetative growth of the ciliate lineage. We also 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
FMLRC (90) using Illumina reads with default parameters. Corrected PacBio reads were aligned to both the MAC genome and the Illumina MIC assembly with BLASTN. Reads were removed if they start or end with telomeres or are aligned better to the MAC. The remaining reads were assembled with wtdbg2 (91, parameters “-x rs”). The PacBio assembly was polished by Pilon (92) with the "--diploid" option. The Illumina and PacBio assemblies were merged by quickmerge (93) with the "-l 5000" option.

MIC genome assembly of *E. woodruffi*

The MIC genome of *E. woodruffi* was assembled using a similar procedure as described above for *Tetmemena*. *E. woodruffi* reads were filtered to remove bacterial contamination, including abundant high-GC contaminants, possibly endosymbionts (94). Nanopore reads with GC content >= 55% were assembled by Flye (95) with the parameter "--meta" for metagenomic assembly of bacterial contigs. We used kaiju (96) to identify bacteria taxa for these contigs. 9 of 10 top-covered contigs derive from Proteobacteria, from which many *Euplotes* symbionts derive (94). Bacterial contamination was removed from Illumina reads if perfectly mapping to these metagenomic contigs by Bowtie2 (97). The cleaned Illumina reads were then assembled by SPAdes with "-k 21,33,55,77,99,127" (77). Pacbio raw reads and Nanopore raw reads with GC content < 55% were aligned to a concatenated database containing both the MAC genome and the Illumina MIC assembly with BLASTN. Reads were removed if they start or end with telomeres or align better to the MAC. Remaining PacBio/Nanopore reads were assembled by Flye with "--meta" mode. The PacBio-Nanopore assembly was polished by Pilon with the "--diploid" option. Illumina and PacBio-Nanopore assemblies were merged by quickmerge with the "-l 10000" option. Contigs shorter than 1kb were removed.
MIC genome decontamination

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

Repeat identification

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

TBE/Tec detection

Representative Oxytricha TBE ORFs (Genbank accession AAB42034.1, AAB42016.1 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
output the rearrangement annotations with minor differences from Chen et al. (1) but most annotations are robust (Figure 3 - figure supplement 2). Scrambled and nonscrambled junctions/IESs were annotated by custom python scripts (https://github.com/yifeng-evo/Oxytricha_Tetmemena_Euplotes/tree/main/scrambled_nonscrambled_IES_pointer).

**MIC genome categories**

Each MIC genome region is assigned to only one category in Figure 2A-C, even if it belongs to more than one category. The assignment is based on the following priority: MDS, TBE/Tec, MIC genes (only available for *Oxytricha*, which has developmental RNA-seq data), IES, tandem repeats, other repeats and non-coding non-repetitive regions. For example, a MIC region can be a TBE in an IES, and it is only considered as TBE in Figure 2A-C.

**Ortholog comparison pipeline and Monte Carlo simulations**

Orthogroups of genes on telomeric MAC contigs were detected by OrthoFinder with “-S blast” (103). Single-copy orthologs were aligned by Clustal Omega (104). Protein alignments were reversely translated to CDS alignments by a modified script of pal2nal (105, https://github.com/yifeng-evo/Oxytricha_Tetmemena_Euplotes/tree/main/Ortholog_comparison/pal2nal.pl). Two modifications were made in the script: 1) the modified script allows pal2nal to take different genetic codes for three sequences (-codontable 6,6,10); 2) the script also fixed an error in the original pal2nal script in which codontable 10 for the Euplotid nuclear code was the same as the universal code. Visualization of pointer positions and intron locations on orthologs was implemented by a custom python script (https://github.com/yifeng-evo/Oxytricha_Tetmemena_Euplotes/tree/main/Ortholog_comparison/pal2nal.pl).
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


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.


43. Gao F, Roy SW, Katz LA. Analyses of alternatively processed genes in ciliates provide insights into the origins of scrambled genomes and may provide a mechanism for speciation. MBio. 2015 Feb 3;6(1):e01998-14.


cephalopod gene regulation and expression through large-scale genome reorganization.


98. Smit AF, Hubley R. RepeatModeler Open-1.0. 2008-2015


<table>
<thead>
<tr>
<th></th>
<th><em>Oxytricha trifallax</em></th>
<th><em>Tetmemena sp.</em></th>
<th><em>Euplotes woodruffi</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAC</strong></td>
<td>67.1</td>
<td>60.6</td>
<td>72.2</td>
</tr>
<tr>
<td><strong>MIC</strong></td>
<td>496</td>
<td>237</td>
<td>172</td>
</tr>
<tr>
<td><strong>N50 (bp)</strong></td>
<td>3,745</td>
<td>3,339</td>
<td>2,702</td>
</tr>
<tr>
<td><strong>GC%</strong></td>
<td>31.36</td>
<td>37.05</td>
<td>36.56</td>
</tr>
<tr>
<td><strong>number of contigs</strong></td>
<td>22,426</td>
<td>25,206</td>
<td>35,099</td>
</tr>
<tr>
<td></td>
<td>25,720</td>
<td>28,446</td>
<td>17,655</td>
</tr>
<tr>
<td><strong>Two-telomere contigs</strong></td>
<td>14,225</td>
<td>-</td>
<td>19,061</td>
</tr>
<tr>
<td><strong>Telomeric contigs</strong></td>
<td>20,336</td>
<td>21,165</td>
<td>28,294</td>
</tr>
<tr>
<td><strong>Single-gene telomeric</strong></td>
<td>76.1%</td>
<td>75.5%</td>
<td>68.5%</td>
</tr>
<tr>
<td><strong>Maximum number of genes on a telomeric contig</strong></td>
<td>8</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

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.
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:


Full citation information for refs. 31 and 32:


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. 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) Tetemena and C) E. woodruffi.
Figure 5. Identification and examples of conserved pointers. A) Pipeline for comparison of pointer positions in orthologs. Orthologs are first grouped by OrthoFinder (103) and protein 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
partial TBE insertion. B) *Oxytricha* MIC contig OXYTRI/MIC_69233 vs. *Tetmemena* MIC contig TMEMEN/MIC_22886. Pointer sequences bridging the nested MDSs of orange and green genes are highlighted. The underlined pointer portions are conserved between species, e.g. the last 8bp of the *Oxytricha* pointer, TAAGTTC\texttt{CAAAGTAG}, are identical to the first 8bp of CAA\texttt{AGTAGCTCAATC} in *Tetmemena*, illustrating pointer sliding (36), or gradual shifting of MDS/IES boundaries. White star indicates a decayed TBE with no ORF identified.
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 $\leq 30$bp in A) *Oxytricha*, B) *Tetmemena* and C) *E. woodruffi*. D-F) Length distribution of scrambled and nonscrambled IESs in D) *Oxytricha* ($\leq 100$bp), E) *Tetmemena* ($\leq 100$bp) and F) *E. woodruffi* ($\leq 300$bp).
**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].)

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 + $1.5 \times$ interquartile range (IQR) and the lower whisker shows the first quartile – $1.5 \times$ 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 *Tetmema*. 
Figure S1

A

B
Figure S2 (relate to figure 2)
Figure 3

A

B

C

D

E

F
Figure S6

Russian doll genes
10808  04166  09313  00618 (139,443 – 140,646 bp)  08901/11158

Other genes
02638  00564  11844  04978  04942  08167  00880  07156/03489

PCR Primers
07156/03489  05053  05069  02448  03934