Cell

The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development

Xiao Chen,[1,9] John R. Bracht,[2,9,10] Aaron David Goldman,[2,11] Egor Dolzhenko,[3] Derek M. Clay,[1] Estienne C. Swart,[4] David H. Perlman,[5] Thomas G. Doak,[6] Andrew Stuart,[7,12] Chris T. Amemiya,[7] Robert P. Sebra,[8] and Laura F. Landweber[2,*]
[1]Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA
[2]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA
[3]Department of Mathematics and Statistics, University of South Florida, Tampa, FL 33620, USA
[4]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland
[5]Collaborative Proteomics and Mass Spectrometry Center, Molecular Biology Department and the Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
[6]Department of Biology, University of Indiana, Bloomington, IN 47405, USA
[7]Benaroya Research Institute at Virginia Mason, Seattle, WA 98101, USA
[8]Icahn Institute and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[9]Co-first author
[10]Present address: Department of Biology, American University, Washington, DC 20016, USA
[11]Present address: Department of Biology, Oberlin College, Oberlin, OH 44074, USA
[12]Present address: Seattle Biomedical Research Institute, Seattle, WA 98109, USA
*Correspondence: lfl@princeton.edu
http://dx.doi.org/10.1016/j.cell.2014.07.034

SUMMARY

Programmed DNA rearrangements in the single-celled eukaryote *Oxytricha trifallax* completely rewire its germline into a somatic nucleus during development. This elaborate, RNA-mediated pathway eliminates noncoding DNA sequences that interrupt gene loci and reorganizes the remaining fragments by inversions and permutations to produce functional genes. Here, we report the *Oxytricha* germline genome and compare it to the somatic genome to present a global view of its massive scale of genome rearrangements. The remarkably encrypted genome architecture contains >3,500 scrambled genes, as well as >800 predicted germline-limited genes expressed, and some posttranslationally modified, during genome rearrangements. Gene segments for different somatic loci often interweave with each other. Single gene segments can contribute to multiple, distinct somatic loci. Terminal precursor segments from neighboring somatic loci map extremely close to each other, often overlapping. This genome assembly provides a draft of a scrambled genome and a powerful model for studies of genome rearrangement.

INTRODUCTION

Genomes are dynamic structures. Humans possess genomic variation among tissues from the same individual (O'Huallachain et al., 2012). Furthermore, genome instability can be a common factor in cancer transformation (Stephens et al., 2011), when thousands of genome rearrangement events contribute to cancer-causing lesions. Curiously, programmed genome rearrangements occur during development in a variety of eukaryotes, with DNA elimination the most frequent type. Examples include chromatin diminution in the parasitic nematode *Ascaris* (Wang et al., 2012) and DNA loss in lamprey (Smith et al., 2012). In both cases, ~10%–20% of germline DNA is eliminated in somatic cells. Rejoining of flanking sequences follows DNA deletion in some cases, but sometimes whole chromosomes are discarded, as in sciarid flies (Goday and Esteban, 2001). Genome-wide DNA rearrangements are most exaggerated in ciliates, particularly in the model organism *Oxytricha trifallax*, which programs not only DNA deletion, but also total reorganization, through RNA-mediated events (Fang et al., 2012; Nowacki et al., 2008). Hence, *Oxytricha* presents a unique opportunity to study the intricate process of large-scale genome remodeling.

O. trifallax, like most ciliates, possesses two types of nuclei in a single cell: a germline *micronucleus* (MIC) and a transcriptionally-active somatic *macronucleus* (MAC) (Prescott, 1994). After sexual conjugation, the old MAC disintegrates and a new MAC develops from a copy of the diploid zygotic MIC through an elaborate cascade of events that delete >90% of the germline DNA and reorganize and join the remaining DNA pieces. The germline precursors of MAC gene loci are highly interrupted by short noncoding elements called internal eliminated sequences (IESs) (Figure 1) that are removed during development. The retained gene segments, called macronuclear-destined sequences (MDSs), are often disordered (scrambled) or inverted in the germline. Thus, macronuclear development requires the rearrangement of MDS segments by inversion or permutation to assemble functional genes. Pairs of short direct repeats, called

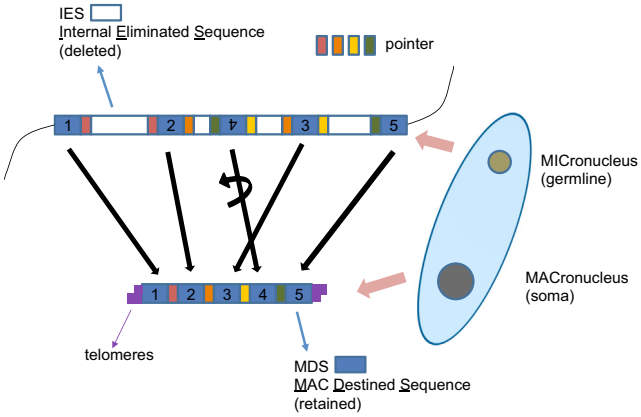

Figure 1. Development of the *Oxytricha* Macronuclear Genome from the Micronuclear Genome

In the micronucleus (MIC), macronuclear destined sequences (MDSs) are interrupted by internal eliminated sequences (IESs); MDSs may be disordered (e.g., MDS 3, 4, and 5) or inverted (e.g., MDS 4). During development after conjugation, IESs, as well as other MIC-limited DNA, are removed. MDSs are stitched together, some requiring inversion and/or unscrambling. Pointers are short identical sequences at consecutive MDS-IES junctions. One copy of the pointer is retained in the new macronucleus (MAC). The old macronuclear genome degrades. Micronuclear chromosome fragmentation produces gene-sized nanochromosomes (capped by telomeres) in the new macronuclear genome. DNA amplification brings nanochromosomes to a high copy number. See also Figure S1.

pointers, are present at consecutive MDS-IES junctions, and one copy of each pointer is retained in the MAC. It is unknown whether the recombination mechanism is a *cis* or *trans* process, or a combination of both, because the polytene chromosome stage during MAC development (Spear and Lauth, 1976) could permit recombination in *trans* between identical copies of MIC DNA molecules. Massive chromosome fragmentation breaks long MIC chromosomes into ~16,000 gene-sized "nanochromosomes" that are bound by short telomeres, average just 3.2 kb, and are each amplified to high copy number (Swart et al., 2013). Recent studies have revealed roles of noncoding RNAs in these events, suggesting that 27 nt Piwi-associated small RNAs (piRNAs) mark MDS regions for retention (Fang et al., 2012) and long, noncoding RNAs serve as epigenetic templates to program segment order and orientation (Nowacki et al., 2008), as well as DNA copy number (Nowacki et al., 2010), in the rearranged, somatic genome.

Previously, very little was known about the *Oxytricha* MIC genome, aside from a general understanding of a small number of scrambled genes and its vast excess of noncoding and repetitive DNA, including satellites and transposons. Among the latter, only the TBE elements (a group of abundant DNA transposons) were previously characterized and shown to participate in programmed deletion of themselves as well as IESs (Nowacki et al., 2009). The MIC genome has been a mysterious puzzle, harboring not only a labyrinth of hundreds of thousands of intricately organized gene segments but also other germline-limited elements that could be involved in genome rearrangement.

High quality draft somatic genome sequences have been reported for the main model ciliates *Tetrahymena* (Eisen et al., 2006), *Paramecium* (Aury et al., 2006), and *Oxytricha* (Swart et al., 2013). Although genome-wide IES studies have been reported for *Paramecium* and *Tetrahymena* (Arnaiz et al., 2012; Fass et al., 2011), neither of which has scrambled genes, no comprehensive germline genome has been described to date for any ciliate species, nor for any organism with a scrambled genome. Here, we present a draft assembly of the *Oxytricha* micronuclear genome and compare it to the somatic genome (Swart et al., 2013) to reveal an unprecedented level of programmed rearrangements and genomic complexity, arguably the most complex genome architecture of any known eukaryote. We demonstrate that the MIC genome sequence is fragmented into over 225,000 segments, tens of thousands of which are complexly scrambled and interwoven. Gene segments from neighboring loci are located in extreme proximity to each other, often overlapping. Furthermore, the discovery of more than 800 germline-restricted genes provides insights into genome rearrangement events.

RESULTS AND DISCUSSION

Genome Sequencing Reveals a Dispersed Set of Fragments that Produce a Somatic Nucleus

We isolated *Oxytricha* micronuclei using sucrose purification (Lauth et al., 1976) and sequenced the DNA to a coverage of ~110× using a shotgun Illumina method and ~15× with single molecule real-time (SMRT) DNA sequencing (Pacific Biosciences). De novo assembly of error-corrected PacBio reads with the Celera assembler (Miller et al., 2008) yields an ~496 Mb draft assembly (Table 1 and Table S1 available online). Additional data (BAC and Fosmid sequencing) were used to validate the assembly (Table S1; Extended Experimental Procedures). Previous studies using reassociation kinetics (Lauth et al., 1976) estimated the MIC genome as 0.3–2.3 Gb and the somatic-destined (MDS) portion to be 2.4%–18% (generally cited as ~5%) of the MIC genome (Prescott, 1994), but we infer it could be as high as 10%, based on read statistics (Figure S1; Extended Experimental Procedures). This small fraction of the germline gives rise to all functional somatic genes.

The MIC assembly contains 98.9% of all nucleotides in the MAC assembly. Of 18,405 MAC contigs with one or both telomeres (Swart et al., 2013), 18,097 (98.3%) are at least 90% covered in the MIC assembly, and 16,220 (88.1%) are at least 90% covered on single MIC contigs, suggesting completely resolved germline-somatic maps. MIC gene loci are typically interrupted by at least one IES, except for 548 IES-less nanochromosomes. Hence, most functional information is encrypted in the MIC, and macronuclear development is a process of decryption. Most IESs interrupt exons (84.7%), making their removal a strict requirement for gene expression.

The germline genome is fragmented into over 225,000 precursor DNA segments (MDSs) that massively rearrange during development to produce nanochromosomes containing approximately one gene each. Note that this number is on the same order of magnitude as the total number of exons in the human genome (Harrow et al., 2012), but these segments fuse via DNA splicing at short direct repeats (*pointers*) at their ends. The six tiniest of these segments (0 bp MDSs) are merely a splint

Table 1. Assembly Statistics from *Oxytricha* Micronuclear and Macronuclear Genomes

	MIC Genome Assembly	MAC Genome Assembly[a]
Estimated genome size (Mb)	~490–500	~50[b]
Total assembly size (Mb)	496.2	55.4
Contig number	25,720	19,152
Number ≥ 200 bp	25,720	19,078 (18,405)[c]
Number ≥ 10 kb	15,942	249
N50 (bp) (≥ 200 bp)	27,807	3,597
Longest (kb)	381	66
GC (%)	28.4	31.0
Repeat (%)	35.9	0
Gene number	810[d]	20,883 (~18,400)[e]

See also Table S1.

[a]The MAC genome assembly was clustered using CD-HIT (Fu et al., 2012) at 95% identity to remove redundancy before calculation of statistics. Repetitive contigs assembled from MIC contamination and bacterial contigs were also removed.

[b]Taken from Swart et al. (2013).

[c]Containing one or both telomeres.

[d]Not including IES-less genes.

[e]Estimated number (18,400) of nonredundant genes. Taken from Swart et al. (2013).

joining two other MDSs. We identified six strong cases of 0 bp MDSs (four nonscrambled and two scrambled; Figure 2C). These comprise just two tandem pointers with no intervening MAC sequence, underscoring the minimalist role of these MDSs as a splint between two adjacent regions that would otherwise share no pointer repeat between them (and be misannotated as 0 bp pointers).

Of all the millions of piRNA sequences (Fang et al., 2012) that map to the MIC genome assembly, 96.0% map to MDS regions. Among the remaining 4% that map to non-MDS regions, the majority (2.4% of total) map to the MAC genome assembly of another strain, JRB510, suggesting that they belong to MIC regions that are either MAC-destined in the other strain or were missed in the JRB310 MAC assembly. The remaining 311,012 (1.6%) reads could also derive from MAC-destined regions that are present on lower copy number nanochromosomes and absent from either MAC assembly. Just 0.11% (21,946) of piRNAs map to MIC-limited repeats, such as TBE transposons and satellites. Therefore, *Oxytricha* piRNAs rarely map to IESs or other MIC-limited sequences. These observations support the model in Fang et al. (2012) that piRNAs mark the precise regions of the MIC genome for retention during genome rearrangement.

Massive Genome Reorganization

In addition to the intense dispersal of all somatic coding information into >225,000 DNA fragments in the germline, a second unprecedented feature of the *Oxytricha* MIC genome is its remarkable level of scrambling (disordered or inverted MDSs). The germline maps of at least 3,593 genes, encoded on 2,818 nanochromosomes, are scrambled. No other sequenced genome bears this level of structural complexity.

Scrambled nanochromosomes are typically longer and contain more MDSs (average 4.9 per kb) than nonscrambled nanochromosomes (average 3.7 per kb; Figure 2A). Among the 2,818 scrambled MAC chromosomes, 1,676 contain at least one inverted segment and 644 contain extended regions that partition odd- and even-numbered segments, a pattern previously observed for a limited number of scrambled genes (Prescott, 1994). The most scrambled gene is a 22 kb MIC locus fragmented into 245 precursor segments that assemble to produce a 13 kb nanochromosome encoding a dynein heavy chain family protein (Figure 2B).

Scrambled MDSs are typically much shorter than nonscrambled MDSs (Figure 2C; median 81 bp for scrambled MDSs, 181 bp for nonscrambled MDSs), presumably reflecting the increased fragmentation of scrambled genes (Figure 2A). While IESs are generally short and GC poor (18% GC), scrambled IESs (median 27 nt) are also much shorter on average than nonscrambled IESs (median 68 nt, Figure 2D). Like the smallest MDS, the shortest IES is 0 nt, just two adjacent scrambled pointers. Figure 2E plots the length distribution of nonscrambled IESs plus one copy of the pointer, which is the full length deleted, since one copy of a pointer is retained. (Scrambled IESs, on the other hand, are flanked by different pointers.) Several weak peaks are present in this length distribution, with a periodicity of approximately 10 bp, similar to one turn of a DNA double helix. A stronger trend in length distribution among nonscrambled IESs in *Paramecium* (Arnaiz et al., 2012) suggests DNA loop formation during assembly of a transposase-containing excision complex. Despite the prevalence of small IESs, gene scrambling permits consecutive MDSs in the MAC to be far apart in the MIC, up to 208 kb (Figure 2F, median distance 2.9 kb).

Scrambled pointers are also longer and more GC-rich than those flanking nonscrambled MDSs (Figure S2A; average scrambled pointer 11 bp and 30% GC, average nonscrambled pointer 5 bp and 19% GC). These longer, more GC-rich pointers may facilitate pointer alignment and unscrambling, even over a distance.

Most 2 bp pointers are TA (Figures S2B and S2C), the only pointer sequence in *Paramecium* IESs and the most common among *Euplotes* IESs, all of which appear to be nonscrambled and have a short terminal consensus sequence resembling the ends of Tc1/*mariner* transposons (Jacobs and Klobutcher, 1996; Klobutcher and Herrick, 1995). This suggests that such IES may be relics from ancient transposon invasion (Klobutcher and Herrick, 1997). Sequences at the ends of TA IESs in *Oxytricha* do not match the *Paramecium* consensus (Figure S2H) but do display a complementary overall base composition that resembles an inverted repeat (Figure S2I).

Among 3 bp pointers, an A nucleotide is overrepresented at the 5′ most position and a T at the 3′ most position (Figures S2D and S2E). ANT is the target duplication site of TBE transposons, present in thousands of copies in the *Oxytricha* MIC genome. Such IESs may also be relics of nonfunctional transposons (Klobutcher and Herrick, 1997) or reflect constraints on splice sites processed by transposon-derived machinery (Nowacki et al., 2009). Their terminal sequences (Figure S2J) also bear a weak resemblance to the first few bases of the CA_4C_4 telomeric repeats at TBE transposon ends (Figure S2K).

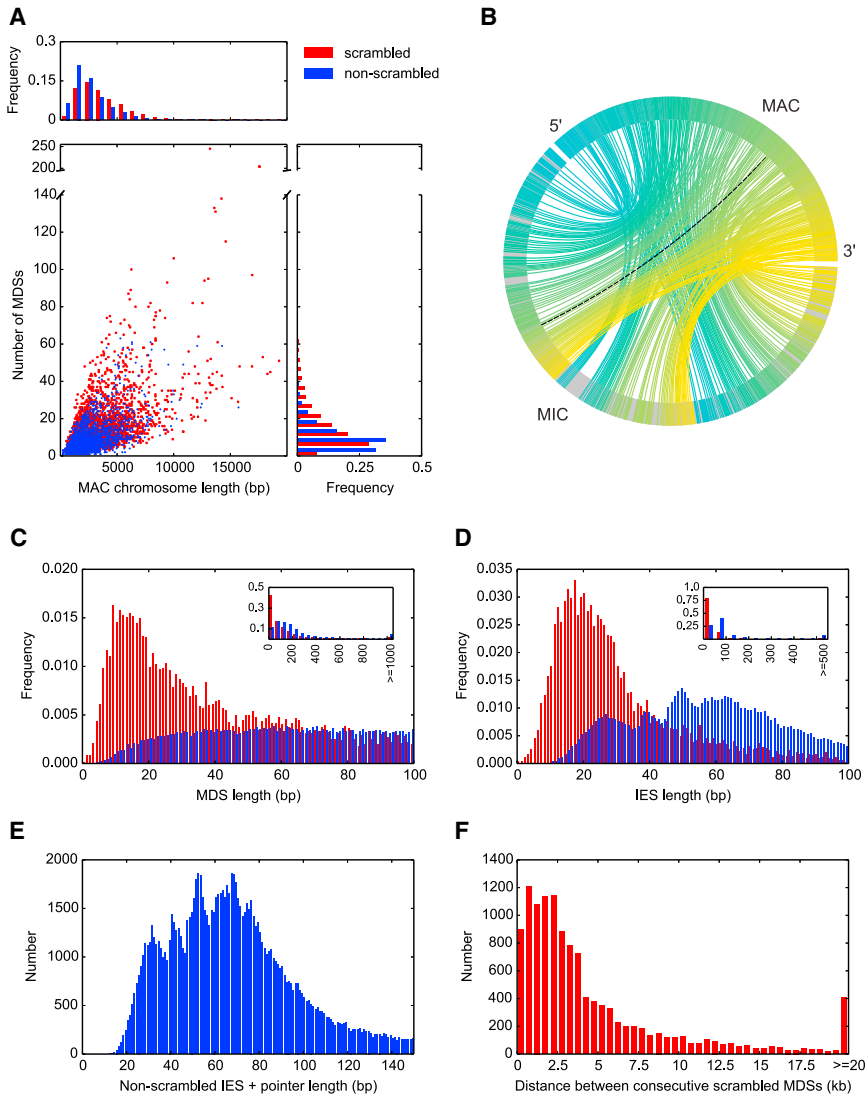

Figure 2. The MIC Genome Is Fragmented into Hundreds of Thousands of Segments with Massive Levels of Scrambling

(A) Comparison of chromosome length and number of MDS segments between 13,910 non-scrambled (blue) and 2,310 scrambled nanochromosomes (red) completely covered on single MIC contigs.

(B) A chord diagram mapping MIC ctg7180000068801 to its rearranged form, MAC Contig17454.0. Lines connect precursor (MIC) and product (MAC) MDS locations; black dotted line, an inverted MDS; all 245 MDSs (242 are scrambled) drawn in a blue to yellow MAC color gradient; IESs, gray.

(C) Length distribution of 44,191 nonscrambled and 9,841 scrambled MDSs <100 nt (excluding pointers). Inset: 150,615 nonscrambled MDSs and 16,350 scrambled MDSs excluding pointers, showing the most typical length of scrambled MDSs is <50 nt.

(D) Length distribution of 101,345 nonscrambled and 8,333 scrambled high-confidence IESs <100 nt (excluding pointers and IESs that contain other MDSs). Inset: 147,122 nonscrambled versus 9,040 scrambled IESs excluding pointers and IESs that contain other MDSs. We identified six strong cases of 0 bp MDSs (four nonscrambled [Contig7827.0 MDS3, Contig11190.0.1 MDS18, Contig13633.0 MDS3, and Contig9208.0.0 MDS18] and two scrambled [Contig6325.0.0 MDS58 and Contig1267.1 MDS7]).

(E) Length distribution of 112,125 nonscrambled IESs (excluding those that contain other MDSs) <150 nt, with one copy of the pointer included (i.e., the total length of DNA deleted).

(F) MIC genomic distance between scrambled MDSs that are consecutive in the MAC (n = 12,197); distance calculated from the pointer flanking MDS N to its paired pointer flanking MDS N+1.

See also Figure S2.

Thousands of Interwoven Gene Loci

A third exceptional feature we noted is 1,537 cases (1,043 of which are scrambled) of nested genes, with the precursor MDS segments for multiple different MAC chromosomes interwoven on the same germline locus, such that IESs for one gene contain MDSs for another. Previously, in an earlier diverged genus, *Uroleptus*, Kuo et al. (2006) discovered that the precursor of a two-gene MAC chromosome exhibits a nested structure, with a precursor segment of one gene present among those for the other gene—but these segments descramble into only one chromosome. The finding in *Oxytricha* of nested structures occurring even among segments for different MAC chromosomes implicates a massive scale and coordination of genome rearrangement to assemble separate MAC chromosomes. Though simple cases of nested genes exist in other eukaryotes; for example, 158 human genes reside completely within the intron of another gene (Kumar, 2009; Yu et al., 2005), the nested gene segments in *Oxytricha* interweave with each other in an elaborately entangled order and orientation. Figure 3A shows a germline locus that contains precursors of 5 nanochromosomes, whose MDSs are not only heavily scrambled themselves, but also deeply interwoven with each other. This type of interwoven architecture also contributes to the large micronuclear distances between MDS segments that are consecutive in the MAC (Figure 2F).

Alternative Processing of MDSs Produces Multiple Genes and Chromosomes

A fourth notable feature arising from this radical genome architecture is that a single MDS in the MIC may contribute to multiple, distinct MAC chromosomes. Like alternative splicing, this modular mechanism of "MDS shuffling" (proposed in Prescott, 1999 and suggested in Katz and Kovner, 2010) can be a source of genetic variation, producing different nanochromosomes and even new genes and scrambled patterns (Figures 3B and 3C). At least 1,267 MDSs from 105 MIC loci are reused, contributing to 240 distinct MAC chromosomes. A single MDS can contribute

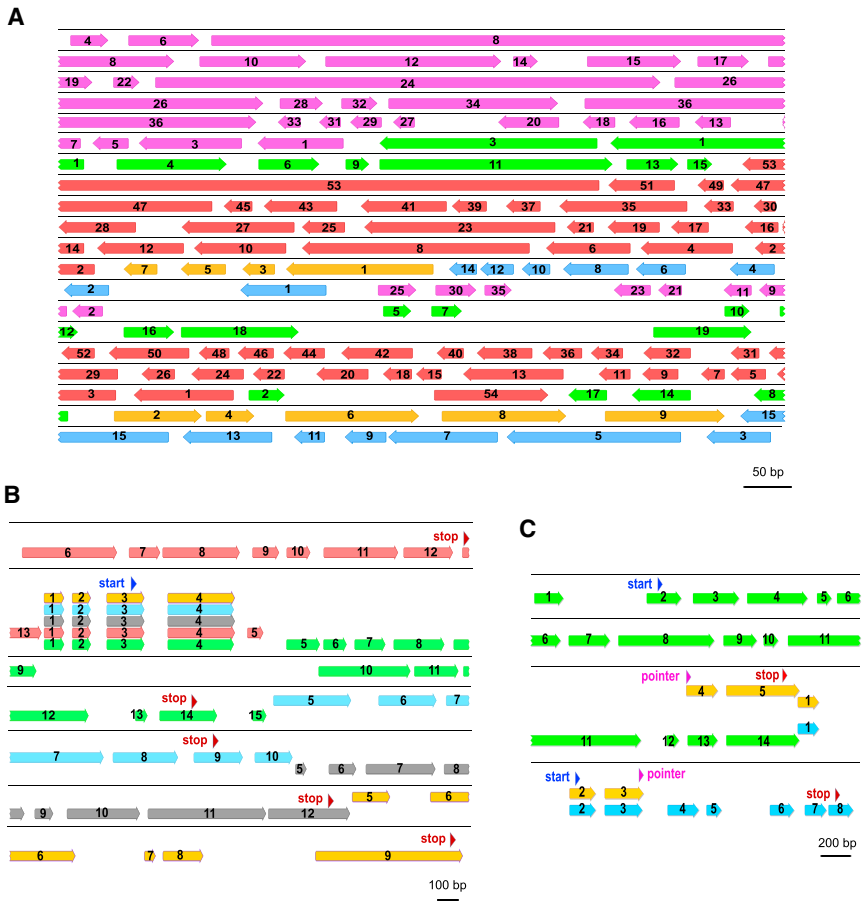

Figure 3. Gene Segments for Multiple Distinct MAC Chromosomes Are Sometimes Interwoven or Reused

(A) Germline map of MIC ctg7180000067411 (drawn to scale), containing precursor MDSs (bars, orientation as shown, including pointers) for five MAC chromosomes (purple, Contig1267.1; green, Contig18709.0; red, Contig20652.0; gold, Contig18297.0; blue, Contig6980.0) whose MDSs are scrambled and interwoven with each other. IES regions are gaps. MDS numbers are consecutive in the MAC.

(B) Germline map of a MIC region (ctg7180000067243) with four shared MDSs that assemble into five distinct MAC chromosomes with identical 5′ ends (red, scrambled Contig14686.0; green, Contig7507.0; blue, Contig7395.0; gray, Contig15152.0; gold, Contig4858.0); start/stop codons annotated in blue and red, respectively.

(C) Germline map (ctg7180000068430) depicting a scrambled MAC chromosome (gold, Contig19716.0) that arose by recombination between MDSs from two different gene loci (green, Contig16277.0; blue, Contig22490.0) at a new pointer (11 bp direct repeat, magenta triangles). Note that the green Contig16277.0 is an alternatively processed chromosome, itself, with two predicted stop codons; the shorter, more abundant isoform (not shown) terminates at an alternative telomere addition site between MDS 12-13, upstream of an intron 3′ splice site. This creates an earlier, in-frame stop codon within the retained portion of the unspliced intron (Swart et al., 2013).

to the assembly of as many as five different nanochromosomes. Figure 3B shows an example where the precursors of five single-gene nanochromosomes (one scrambled) share 4 MDSs at the 5′ end of their encoded genes but have different sets of 3′ MDSs. The shared MDSs preserve reading frame and use the same start codon.

MDS sharing can also produce new MAC chromosome architectures. For example, in Figure 3C, the gold nanochromosome fuses the first three MDSs from the 5′ end of the downstream gene (blue) to the last two MDSs from the upstream gene (green) via recombination at an 11 bp direct repeat (magenta triangle, labeled "pointer"). This single event creates a novel chimeric, scrambled nanochromosome from two precursor nonscrambled loci, supporting the ability of MDS shuffling (Prescott, 1999) to contribute to the origins of both new genes and new scrambled genetic architectures.

We examined a set of potential new genes that are produced through combinatorial assembly of both reused and unique MDSs. Of the 105 MIC loci that share some MDSs, 55 encode paralogous proteins (i.e., some of their unique MDSs also share sequence similarity at the predicted protein level). Such cases might derive from duplication and divergence of MDSs but not complete gene loci. Thirty-two cases lack paralogy outside the shared MDS, resulting in genuinely chimeric proteins that fuse identical sequence blocks to completely unique blocks. In one case, the shared MDSs do not extend into the coding regions. In 12 cases, entire reading frames reside within the shared MDSs, producing no new predicted gene structures. The remaining five cases have no predicted protein-coding genes.

Precursors of Chromosome Ends Often Overlap

In addition to removal of MIC-limited sequences and descrambling of MDSs, macronuclear development also involves fragmentation of the long MIC chromosomes and addition of telomeres at the new termini, producing gene-sized nanochromosomes in the MAC. Among the 22,875 terminal MDSs we identified in the MIC genome, most (20,012) are adjacent to a terminal MDS of another nanochromosome, while 2,863 reside next to an MDS that is an internal segment of another nanochromosome. A fifth remarkable feature of *Oxytricha*'s MIC genome, deriving from the proximity between terminal segments for different nanochromosomes, is that these MDSs frequently overlap. This creates vanishingly short intergenic regions, to the point where the median distance between 10,006 pairs of terminal MDSs that are adjacent to each other in the MIC (Figure 4, black bars) is precisely 0 bp. (The range is −34 bp to 19 kb, where a negative value indicates the length of overlap.) This is in striking contrast with the absence of gene linkage on most somatic chromosomes (Swart et al., 2013). While preliminary studies of the related ciliate, *O. nova*, hinted that MDS-containing regions

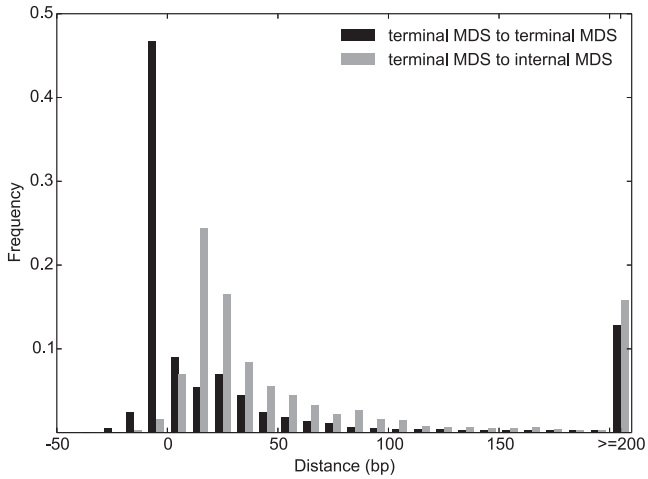

Figure 4. The Distance between Adjacent Terminal MDSs Is Much Smaller Than that between Terminal MDSs and Adjacent Internal MDSs for a Different MAC Chromosome

Negative values represent the length of overlapping regions, with the peak distance between terminal MDSs from −1 to −10 bp (10,006 pairs, black) and the peak between terminal MDS to internal MDS between 10–19 bp (2,863 pairs, gray).

See also Figure S3.

cluster in the germline (Boswell et al., 1983; Klobutcher, 1987; Klobutcher et al., 1988), this type of gene density is exaggerated to the point where nearly half of nanochromosome ends actually overlap. The variable length of overlap is consistent with a mechanism of micronuclear chromosome fragmentation that involves staggered cuts to allow production of both chromosomes from one precursor or production of two chromosomes with overlapping ends from different polytene chromosomes (Klobutcher et al., 1988).

If IES removal and chromosome fragmentation are separate events, then during the stage of IES removal, two adjacent terminal MDSs for different nanochromosomes could sometimes be processed as a single MDS, with chromosome fragmentation and telomere addition to follow. This would economically require just a single cut when the distance between terminal MDSs is <1 bp. Terminal MDSs next to an internal MDS, on the other hand, must be processed separately, because they need to be joined to consecutive MDSs. Correspondingly, the distance between a terminal MDS and an adjacent internal MDS is generally much larger (median 30 bp, Figure 4, gray bars). Sequence features flanking chromosome fragmentation sites are discussed in the Extended Experimental Procedures and Figure S3.

Germline-Restricted Protein-Coding Genes

With TBE transposases a notable exception (Nowacki et al., 2009), the deleted portion of *Oxytricha*'s germline has generally been considered transcriptionally inactive. However, a sixth main conclusion of this study was the discovery and confirmation of hundreds of expressed germline-limited genes, many with predicted functions that could relate to genome rearrangement. In addition to 548 IES-less nanochromosomes, some of which appear expressed from both the MIC and the MAC (see next section), we predicted 810 expressed, nonrepetitive (single copy) MIC-limited protein-coding genes (including one MT-A70 gene family; see below). Sixty-eight of these MIC-limited genes fully reside within an IES of another MAC-destined locus. Therefore, IESs, often considered to be AT-rich "junk" DNA, can not only harbor MDSs of other genes, but they also bear germline-limited genes that are discarded during genome rearrangement.

Based on RNA sequencing (RNA-seq) data (Swart et al., 2013) these 810 germline-limited genes are almost exclusively expressed during conjugation (peaking 40–60 hr after the onset of conjugation, with 98% expressed only at 40 and/or 60 hr and little transcription in asexually dividing (vegetative) cells or at the "0 hr" time point when cells of compatible mating types are mixed to initiate conjugation, Figure 5A). The developmentally-limited expression of these germline genes is naturally abrogated by DNA diminution, which has been proposed as a mechanism of germline gene regulation in lamprey (*Petromyzon marinus*) (Smith et al., 2012) and *Ascaris suum* (Wang et al., 2012). The ciliate *Euplotes crassus* possesses a telomerase gene that is expressed only during development, after activation by IES deletion, but the gene itself is absent from the vegetative MAC, suggesting that programmed gene elimination may shut off its expression (Karamysheva et al., 2003). In lamprey and *Ascaris*, the genes eliminated from somatic cells are mostly expressed during gametogenesis or early embryogenesis. They are often associated with basic cellular functions such as transcription, translation and chromatin remodeling, suggesting that these genes are likely to be involved in development or maintenance of the germline. In the unicellular *Oxytricha*, however, in addition to germline differentiation and maintenance, germline-limited genes could also provide functions in early somatic differentiation, specifically in macronuclear development and genome rearrangement, bridging the interval from degradation of the parental MAC through production of new macronuclei. This also suggests that, despite unicellularity, these microbial eukaryotes may harbor orthologs of genes required for the evolution of differentiated multicellularity, at least a refined germline-soma distinction. MIC-limited genes could technically be expressed from either the micronucleus or the developing macronucleus (before they are eliminated). Because RNA-seq data (Swart et al., 2013) derive from whole cells, we are unable to deduce at this time whether expression derives exclusively from either organelle.

Table S2 compares the properties of predicted *Oxytricha* germline-limited genes to the categories of both IES-less genes and MAC-specific genes on completely sequenced nanochromosomes (Swart et al., 2013). Predicted genes and introns are both shorter in MIC-limited genes than in the MAC. In addition, the MIC genome is much more intron-poor (chi-square test, p value $<2.2 \times 10^{-16}$). A possible explanation may be that the micronucleus or the developing macronucleus could have limited access to intron splicing machinery. Among MIC genes expressed at 40 or 60 hr, their expression levels are not significantly different from expressed MAC genes (2 sample t test, p value = 0.3148, 0.1285 for 40 and 60 hr, respectively). Among 810 predicted MIC genes, only 311 are located on MIC contigs that contain MDSs for MAC loci, while the others map either to

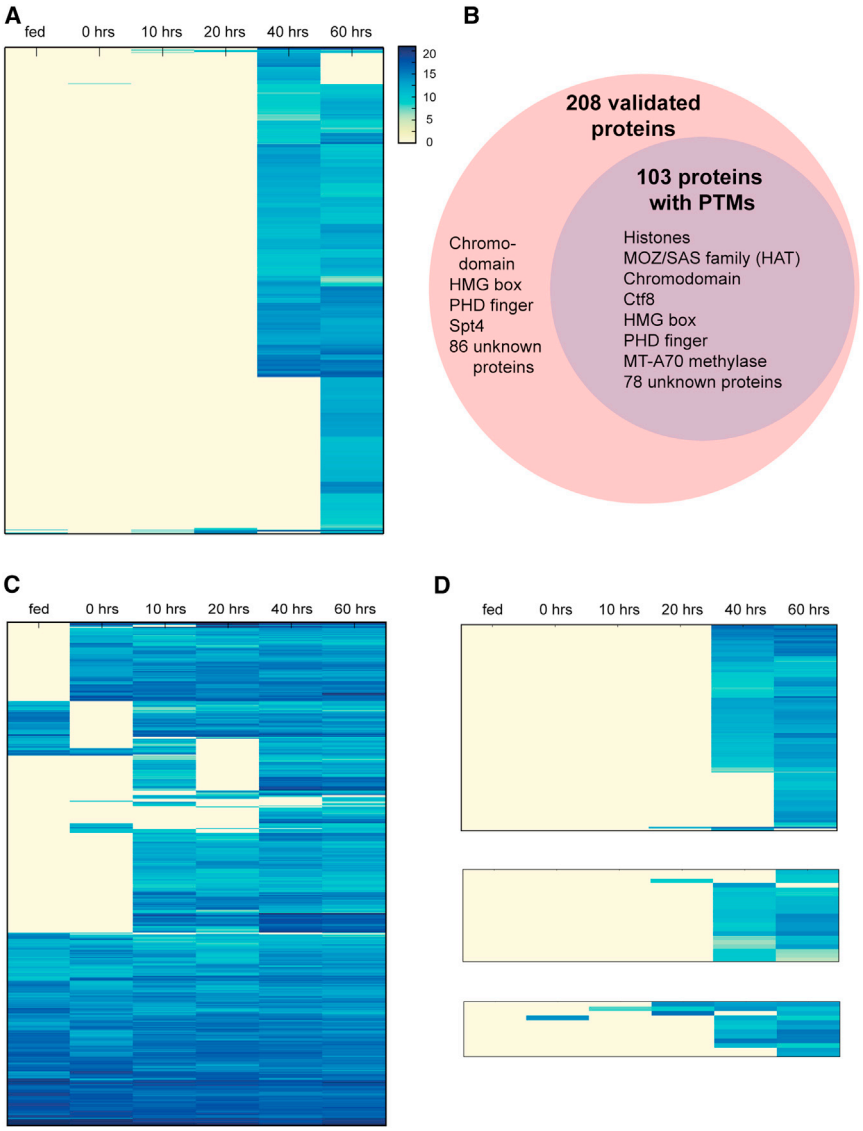

Figure 5. MIC-Limited Genes and Transposons Are Preferentially Expressed during Conjugation, while IES-less Genes Are More Constitutive but Show Universally High Expression during Conjugation

(A) Clustered expression profile of 810 germline-limited nonrepetitive genes across different time points (vegetative stage (fed); 0, 10, 20, 40, and 60 hr during the conjugating time course). Gene expression levels are represented by log2 (100,000 × normalized RNA-seq counts/coding sequence length).

(B) Mass spectrometry validated 208 MIC-limited genes (outer, pink circle) and 103 were found to contain posttranslational modifications (PTMs) (inner, purple circle). Representative members of each group are shown within the circles.

(C) Clustered expression profiles of 530 IES-less genes.

(D) Clustered expression profiles of MIC-limited transposon-associated genes. Upper: 275 reverse transcriptase and endonuclease domain proteins encoded by LINEs; Middle: 21 *Helitron*-associated helicases; Lower: 12 DDE_Tnp_IS1595 (ISXO2-like transposase) domain proteins from insertion sequences (ISs).

See also Tables S2, S3, S4, S5, S6, S7, and S8 and Figure S4.

short contigs or to entirely MIC-limited regions of the genome, including contigs with repetitive elements.

Proteomic analysis unambiguously supported translation of 208 predicted germline-limited genes (26% of the 810; Figure 5B; Tables S3 and S4). High-resolution, accurate-mass ultra-high performance liquid chromatography mass spectrometry (UPLC-MS) analysis of *Oxytricha* 40 hr nuclear lysate with minimal upfront fractionation by SDS-PAGE detected MIC-encoded proteins, based on high-confidence MS characterization of over 100 peptides per protein in some cases (Table S3). With over one million tandem mass spectra from MAC and MIC peptides assigned to >6,900 proteins, this analysis was sufficiently deep to reveal significant stoichiometric and substoichiometric post-translational modifications (PTMs) on half (103) of the 208 validated proteins. These modifications include methylation, acetylation, and phosphorylation, consistent with functional regulation of the MIC-encoded proteins. Both phosphorylation and acetylation occur on proteins with predicted diverse cellular functions, as well as unknown proteins (Figures S4A–S4D) that account for most (74%) of the phosphorylated cases. Such candidates might participate in signaling pathways that coordinate genome rearrangements.

While 118 predicted MIC genes have paralogs in the *Oxytricha* MAC genome, several MIC-specific protein domains are not identified in the MAC (Table S5). An example is Ctf8 (chromosome transmission fidelity 8), a component of a DNA clamp loader involved in sister chromatid cohesion and usually associated with mitosis/meiosis, both specific to the MIC, although it could also associate with polytene chromosomes during differentiation (Spear and Lauth, 1976). Proteomic analysis confirmed Ctf8 expression, with seven unique peptides, one of which contained serine 36 phosphorylation (Table S3), suggesting regulation. Most other MIC-specific protein domains are virus-associated, although the hits are often weak, as suggested by high E-values. This could be due to viral integration into the *Oxytricha* MIC genome. Moreover, these virus-associated genes use the *Oxytricha* genetic code and some of the contigs containing them bear MDSs and/or TBEs, suggesting that they are not contaminants. In addition, mass spectrometry validated the presence of four viral proteins (Filoviridae VP35 domain and three Parvovirus nonstructural protein NS1 domain; Table S3). Mass spectrometry also identified specific phosphorylation or methylation of three of these proteins.

Table 2. GO Terms Enriched in Predicted Germline-Limited Genes

GO Term	Description	Ratio in MIC-Limited Genes	Ratio in All Genes (MIC + MAC)	Fold Enrichment	Bonferroni-Corrected p Value
GO:0008168	methyltransferase activity	56/810	171/21,693	8.8	1.50×10^{-8}
GO:0016741	transferase activity, transferring one-carbon groups	56/810	173/21,693	8.7	1.99×10^{-8}
GO:0032259	methylation	56/810	165/21,693	9.1	3.38×10^{-8}
GO:0006139	nucleobase-containing compound metabolic process	64/810	760/21,693	2.3	6.79×10^{-7}
GO:0006725	cellular aromatic compound metabolic process	64/810	784/21,693	2.2	1.46×10^{-6}
GO:0046483	heterocycle metabolic process	64/810	786/21,693	2.2	2.40×10^{-6}
GO:1901360	organic cyclic compound metabolic process	64/810	793/21,693	2.2	2.81×10^{-6}
GO:0034641	cellular nitrogen compound metabolic process	64/810	798/21,693	2.1	3.26×10^{-6}

Similar to lamprey and *Ascaris*, the *Oxytricha* germline genome encodes a repertoire of chromatin-associated genes (Table S6), and these are significantly enriched in the phosphorylated and acetylated protein sets, relative to the total predicted MIC-limited genes (GO term enrichment test, p value = 7.55×10^{-5} and 1.63×10^{-3}, respectively). These include a complete set of core histones (one H2A, two H2B, one H3, and one H4; H1 has not been identified in either the *Oxytricha* MAC or MIC) and genes associated with histone modification (three PHD, one SET, and six chromodomain proteins), suggesting a direct involvement of chromatin and chromatin remodeling proteins in genome rearrangement and germline maintenance. Histone N-terminal tails were the most heavily modified MIC-limited peptides in the proteomic analysis (Table S3). For example, Histone H3 contains both repressive marks (e.g., H3K9 and H3K27 trimethylation) and activating marks (e.g., H3K9 acetylation) plus H3K4 monomethylation, which can be either activating or repressing (Cheng et al., 2014). Histone H4 was heavily modified (61 detected PTMs) with both activating and repressive marks identified on the same residues. The most heavily modified MIC-limited gene was an H2B variant, with 83 PTMs. We note that a *Euplotes* development-specific histone H3 (encoded in the MAC) has conjugation-specific expression in the developing MAC (Ghosh and Klobutcher, 2000). The presence of *Oxytricha* MIC-specific chromatin components and modifiers could allow changes in nucleosome composition and chromatin structure to regulate genome rearrangement. Development-specific histones or their modifications could either mark DNA segments for genome restructuring or alter the chromatin structure to allow access to machinery for DNA deletion and rearrangement. Proteomic analysis confirmed expression of a histone acetyltransferase of the MOZ/SAS family (18 unique peptides) with its own acetylation at lysines 7 and 257 (Table S3), suggesting the possibility of self-regulation through autoacetylation. Thirty unique peptides also confirmed expression of a MIC-limited SET domain histone methyltransferase. In addition to known epigenetic modifiers, the MIC genome encodes other genes that could directly manipulate DNA during genome rearrangement, such as a DNA topoisomerase (that could regulate DNA unwinding during recombination), a helicase, and an HTH_Tnp_Tc5 (Tc5 transposase DNA-binding domain) protein.

GO terms associated with predicted MIC-limited genes are especially enriched in activities related to methylation (Table 2). The MIC encodes a large set of 61 MT-A70 domain proteins (RNA adenosine methyltransferases, four of which are confirmed by proteomic analysis) that could participate in many steps during RNA-guided DNA rearrangement. During conjugation, both long, noncoding RNAs and 27 nt piRNAs are produced from the parental MAC and transported to the developing MAC, providing essential information about which sequences to retain and their rearrangement order (Fang et al., 2012; Nowacki et al., 2008). RNA adenosine methylation is a widespread and dynamically regulated posttranscriptional RNA modification (Meyer et al., 2012). It might function in *Oxytricha* to regulate noncoding RNAs or to mark specific sites or sequences on noncoding RNAs that guide genome rearrangement during development.

IES-less Genes

IES-less genes are a second category of genes that can be expressed from the micronucleus or developing macronucleus, but also the MAC, itself. While they display different expression levels during a conjugating time-series, most IES-less genes are highly expressed at 40 and 60 hr (Figure 5C). Furthermore, they have significantly higher expression during conjugation than genes whose MIC precursors contain IESs (2 sample t test, p value = 2.585×10^{-9} and 0.0003037 for 40 and 60 hr, respectively). Their gene features lie between those of MAC-specific and MIC-specific genes (Table S2). In particular, they contain fewer introns per gene than genes encoded on MAC nanochromosomes with IESs (chi-square test, p value = 3.38×10^{-7}). Among the 530 genes encoded on 548 IES-less nanochromosomes, 278 lack introns. It is possible that expression from the MIC contributes to their high expression levels, especially when the parental MAC is degraded. To query MIC-specific expression, we examined RNA-seq reads that mapped to MIC loci for these 530 genes and found 21 cases where reads mapped beyond all detected MAC telomere addition sites, consistent with transcription from the MIC or from incorrectly processed MAC chromosome ends.

While GO term enrichment analysis suggests that these IES-less genes are not enriched for specific functions (except carboxylesterase activity, GO:0004091), future studies can address

whether the absence of IESs specifically regulates their early expression or is a requirement for function during genome rearrangement.

Repetitive Elements

Finally, we analyzed the types and percentages of various repetitive elements in the genome, with the caveat that genome assembly of repetitive regions poses a special challenge and may be an underrepresentation. The MIC genome contains four types of germline-limited transposable elements, of which only TBEs were previously described in *Oxytricha* (Doak et al., 1994; Herrick et al., 1985; Nowacki et al., 2009): TBE transposons (DDE family cut-and-paste DNA transposons), LINEs, *Helitrons* (rolling-circle transposons), and insertion sequences (Table S8). *Oxytricha*'s germline genome appears less transposon-rich than our own, which is roughly half transposon-derived (de Koning et al., 2011; Lander et al., 2001). With the caveat that transposable elements may be degenerate and individual sequences not assembled accurately, we predicted hundreds of genes associated with these transposons (reverse transcriptase and endonuclease domain genes, *Helitron*-associated helicases and DDE_Tnp_IS1595 [ISXO2-like transposase] domain genes). Their expression is also limited to 40–60 hr into conjugation (Figure 5D), suggesting that, like TBE transposases (Nowacki et al., 2009), they could be recruited to function during genome rearrangement. Mass spectrometry confirmed expression at 40 hr of *Helitron*-associated helicases and LINE-associated genes, both of which are often posttranslationally modified, suggesting regulated functions (Table S7).

The largest class of repetitive elements, TBE transposons frequently map near MDSs and interrupt at least 6,776 nanochromosome gene loci. They encode three genes: a 42 kDa transposase, implicated in genome rearrangement (Nowacki et al., 2009), a 22 kDa unknown ORF, and a 57 kDa gene with a zinc finger/kinase domain (Witherspoon et al., 1997) and fall into three classes: TBE1, TBE2, and TBE3, with two subfamilies that we identified within TBE2 (see Extended Experimental Procedures). Long PacBio reads allowed us to successfully assemble as many as 16 complete TBE transposons on a single MIC contig (three TBE1, seven TBE2, and six TBE3), demonstrating the power of this approach to resolve repetitive regions.

LINE elements, also present in the *Tetrahymena* germline (Fillingham et al., 2004), interrupt at least nine *Oxytricha* precursor gene loci. In the *Oxytricha* MAC genome, telomerase is the only protein containing an RVT_1 (reverse transcriptase) domain, commonly associated with telomerases or retrotransposons. Curiously, in the MAC genomes of both *Paramecium* and *Tetrahymena*, the RVT_1 domain is present in other genes besides telomerase. *Paramecium* gene model GSPATP00023049001 matches the RVT_1 domain with a HMMER E-value of 1.3×10^{-4} and does not show significant differential expression during conjugation (http://Paramecium.cgm.cnrs-gif.fr/db/feature/217802). In *Tetrahymena*, gene model TTHERM_00129610 matches RVT_1 domain with a HMMER E-value of 8.9×10^{-30} and is upregulated during conjugation (http://tfgd.ihb.ac.cn/search/detail/gene/TTHERM_00129610). It is possible that their LINE-associated reverse transcriptases were domesticated in the MAC, unless those genome assemblies are contaminated by MIC sequences (that is less likely because the two RVT_1 domain genes are located on long MAC contigs [>300 kb]). *Helitrons*, on the other hand, do not appear to interrupt any *Oxytricha* precursor gene loci. Seven *Helitron*-associated genes are also present in the *Tetrahymena* MIC, suggesting a possible deeper evolutionary origin.

Swart et al. (2013) previously predicted 21 proteins containing Phage_integrase, DDE_Tnp_IS1595, or MULE transposase domains in the MAC proteome of *Oxytricha* but not *Tetrahymena* or *Paramecium*. All these *Oxytricha* transposase domain genes show highly conjugation-specific expression. Their MIC precursors all contain IESs, however, and none are full-length transposons. We did not find any germline transposons bearing Phage_integrase or MULE domain genes, but we did identify hundreds of MIC insertion sequences (0.12% – 0.2% of the MIC genome; Table S8) that carry DDE_Tnp_IS1595 domain genes. These sequences interrupt at least 24 precursor gene loci in the MIC. Although usually rare in eukaryotes, this protein domain is present in *Stylonychia*'s MAC genome and *Perkinsus* (Swart et al., 2013), but absent from *Paramecium*'s MAC genome and also appears absent from both the MAC and MIC genomes of *Tetrahymena*. Full-length insertion sequences are also rare in eukaryotes. Hence, the discovery of insertion sequences bearing these transposase genes and with conjugation-specific expression suggests they could be another class of domesticated transposase recruited to genome rearrangement.

Two major classes of satellite repeats that were previously identified by hybridization, but not sequencing (Dawson et al., 1984) are also present in the germline, with repeat units of 170 and 380 bp, respectively (Table S8; Extended Experimental Procedures). They rarely interrupt MDSs (just 11 cases) and some fosmid clones have the same repeat sequence present at both ends, suggesting that large, satellite repeat-dense regions may cluster independently of MDS-rich clusters. This satellite organization may facilitate their complete elimination during development.

Conclusions

The assembled *Oxytricha* micronuclear genome greatly expands our perspective on the limits of genome complexity, displaying an unprecedentedly fragmented and scrambled genome architecture, with thousands of scrambled genes. We provide complete germline-somatic maps for the majority of genes and a window into nuclear development at a whole-genome level.

The correct interpretation of complex MDS-IES structures relies on the accuracy of genome assembly. Because macronuclear contamination was nearly absent from our micronuclear DNA preparations (see Extended Experimental Procedures), interference of macronuclear sequences in the assembly was kept to a minimum. We also validated portions of the assembly via several different approaches (see Extended Experimental Procedures). The assembly agrees with the validation data in all cases in nonrepetitive regions. Therefore, we conclude that this assembly is an accurate representation of the *Oxytricha* germline genome. Repetitive regions offer a significant challenge for the assembly of any complex genome. Despite the fragmentary nature of the assembly at repetitive regions, we were able to achieve a relatively continuous assembly of the nonrepetitive

regions, especially the MDS-containing regions, from which we derive most of our biological conclusions.

The discovery of hundreds of germline-limited nonrepetitive genes is unique among unicellular eukaryotes, so far, and elegantly suggests a cache of functional genes that support somatic differentiation and genome rearrangement when the maternal somatic genome is destroyed, consistent with the proposed use of chromatin diminution for germline gene regulation in lamprey (Smith et al., 2012) and *Ascaris* (Wang et al., 2012). Validation of 26% of these germline-limited genes by MS-based proteomics, plus identification of posttranslational modifications affecting half of the validated genes, hints at a complex protein regulatory network during somatic differentiation. The *Oxytricha* germline genome assembly presented here provides a valuable resource for comparative genomics, even within a single cell, a window into the extreme limits of eukaryotic genome architecture, and a platform for future studies of genome remodeling.

EXPERIMENTAL PROCEDURES

See the Extended Experimental Procedures for detailed protocols.

Nuclei Isolation, DNA Extraction, and Genome Sequencing

We grew vegetative cultures of *Oxytricha* strain JRB310 and isolated micronuclei using sucrose gradient centrifugation, as described in Lauth et al. (1976). Different libraries were prepared from extracted DNA and sequenced with Illumina HiSeq 2000 and PacBio platforms.

Genome Assembly

Illumina unitigs assembled from MaSuRCA (Zimin et al., 2013) were used to error correct PacBio reads with ectools (https://github.com/jgurtowski/ectools). Corrected PacBio reads were assembled using the Celera assembler (Miller et al., 2008).

Identifying Genome Rearrangement Junctions and MDS, IES, and Pointer Designations

We mapped the MAC genome assembly (excluding telomeres) onto the MIC assembly using BLASTN (BLAST+ [Camacho et al., 2009] parameters: -ungapped -word_size 20 -outfmt 6). Paralogous MDS regions were filtered out if they had poor sequence similarity (<98%) to the MAC and no pointers between consecutive matches. Custom Python scripts were used to extract MDSs, IESs, and pointers from the BLAST output. MDSs, IESs, and pointers are defined below. Complete MDS-IES maps are available at http://trifallax.princeton.edu/cms/raw-data/genome/mic/Oxytricha_trifallax_micronuclear_genome_MDS_IES_maps.gff.

Pointers

Pointers are short sequences of microhomology repeated at MDS-IES junctions and present as one copy at consecutive MDS-MDS junctions in the MAC. No mismatch is allowed between the two copies of each pointer.

MDSs

MDSs are sequence blocks retained in the MAC, excluding the pointers.

IESs

IESs are sequence blocks that separate MDSs in the MIC and are absent from the MAC, excluding the pointers.

Nonscrambled MDSs

Nonscrambled MDSs are MDSs that are in the same orientation and order in the MIC relative to the MAC.

Scrambled MDSs

Scrambled MDSs are MDSs that are not in the same orientation or order in the MIC relative to the MAC.

Nonscrambled IESs

Nonscrambled IESs are IESs between two MDSs that are in the same orientation and order in the MIC relative to the MAC. They are flanked by identical pointer repeats.

Scrambled IESs

Scrambled IESs are IESs between two MDSs that are not in the same order or orientation in the MIC relative to the MAC. They are flanked by different pointer sequences.

Gene Prediction

We gathered RNA-seq reads from Swart et al. (2013) and assembled a transcriptome using only reads that do not map to the MAC genome assembly with SOAPdenovo-Trans (Li et al., 2010), Trinity (Grabherr et al., 2011) and PASA (Haas et al., 2003). We predicted gene models with AUGUSTUS (version 2.5.5) (Stanke and Morgenstern, 2005) using assembled transcripts as hints.

ACCESSION NUMBERS

The GenBank accession number for the genome assembly and the raw sequencing data reported in this paper is ARYC00000000.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.07.034.

AUTHOR CONTRIBUTIONS

J.R.B. optimized experimental procedures and isolated micronuclei and extracted genomic DNA. X.C. assembled the genome, conducted the bioinformatic analyses in the paper, and drafted the manuscript. T.G.D., J.R.B., A.D.G., and L.F.L. conceived the project, which L.F.L. supervised. R.P.S. performed PacBio library preparation and sequencing. D.H.P performed mass spectrometry experiments. J.R.B. and D.H.P. analyzed the proteomic data. J.R.B., A.D.G., A.S., and C.T.A. prepared and sequenced the fosmid and BAC clones. D.M.C. performed assembly validation by PCR. E.D. produced chord diagrams for visualization of germline-soma maps. E.C.S. provided advice on genome analysis and edited the manuscript. J.R.B., D.H.P., and R.P.S. contributed to the writing of the manuscript, which X.C., J.R.B., and L.F.L. extensively edited.

ACKNOWLEDGMENTS

We thank the late David Prescott for suggesting sucrose purification of the *Oxytricha* micronuclei, Jingmei Wang for laboratory assistance, Jessica Wiggins, Wei Wang, and Donna Storton of the Princeton Sequencing Core Facility for assistance with Illumina library preparation and sequencing, Gintaras Deikus for assistance with PacBio sequencing, Klaas Schotanus, Mariusz Nowacki, Wenwen Fang, and all current laboratory members for discussion. We thank Jue Ruan at Beijing Institute of Genomics for advice on the assembly strategy. We are grateful to National Center for Genome Analysis Support (NCGAS) computing resources (supported by National Science Foundation [NSF] grant ABI-1062432 to Indiana University). This study was supported by NIH grant GM59708 and GM109459 and NSF grants 0900544 and 0923810 to L.F.L.. J.R.B. was supported by NIH fellowship 1F32GM099462 and A.D.G. was a National Aeronautics and Space Administration (NASA) postdoctoral fellow.

Received: October 8, 2013
Revised: May 18, 2014
Accepted: July 3, 2014
Published: August 28, 2014

REFERENCES

Arnaiz, O., Mathy, N., Baudry, C., Malinsky, S., Aury, J.-M., Wilkes, C.D., Garnier, O., Labadie, K., Lauderdale, B.E., Le Mouël, A., et al. (2012). The *Paramecium* germline genome provides a niche for intragenic parasitic DNA:

evolutionary dynamics of internal eliminated sequences. PLoS Genet. *8*, e1002984.

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature *444*, 171–178.

Boswell, R.E., Jahn, C.L., Greslin, A.F., and Prescott, D.M. (1983). Organization of gene and non-gene sequences in micronuclear DNA of *Oxytricha nova*. Nucleic Acids Res. *11*, 3651–3663.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Cheng, J., Blum, R., Bowman, C., Hu, D., Shilatifard, A., Shen, S., and Dynlacht, B.D. (2014). A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. Mol. Cell *53*, 979–992.

Dawson, D., Buckley, B., Cartinhour, S., Myers, R., and Herrick, G. (1984). Elimination of germ-line tandemly repeated sequences from the somatic genome of the ciliate *Oxytricha fallax*. Chromosoma *90*, 289–294.

de Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. *7*, e1002384.

Doak, T.G., Doerder, F.P., Jahn, C.L., and Herrick, G. (1994). A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. USA *91*, 942–946.

Eisen, J.A., Coyne, R.S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J.R., Badger, J.H., Ren, Q., Amedeo, P., Jones, K.M., et al. (2006). Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. *4*, e286.

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

Fass, J.N., Joshi, N.A., Couvillion, M.T., Bowen, J., Gorovsky, M.A., Hamilton, E.P., Orias, E., Hong, K., Coyne, R.S., Eisen, J.A., et al. (2011). Genome-scale analysis of programmed DNA elimination sites in *Tetrahymena thermophila*. G3 (Bethesda) *1*, 515–522.

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

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Ghosh, S., and Klobutcher, L.A. (2000). A development-specific histone H3 localizes to the developing macronucleus of *Euplotes*. Genesis *26*, 179–188.

Goday, C., and Esteban, M.R. (2001). Chromosome elimination in sciarid flies. BioEssays *23*, 242–250.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Jr., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. *31*, 5654–5666.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

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

Jacobs, M.E., and Klobutcher, L.A. (1996). The long and the short of developmental DNA deletion in *Euplotes crassus*. J. Eukaryot. Microbiol. *43*, 442–452.

Karamysheva, Z., Wang, L., Shrode, T., Bednenko, J., Hurley, L.A., and Shippen, D.E. (2003). Developmentally programmed gene elimination in *Euplotes crassus* facilitates a switch in the telomerase catalytic subunit. Cell *113*, 565–576.

Katz, L.A., and Kovner, A.M. (2010). Alternative processing of scrambled genes generates protein diversity in the ciliate *Chilodonella uncinata*. J. Exp. Zoolog. B Mol. Dev. Evol. *314*, 480–488.

Klobutcher, L.A. (1987). Micronuclear organization of macronuclear genes in the hypotrichous ciliate *Oxytricha nova*. J. Protozool. *34*, 424–428.

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

Klobutcher, L.A., and Herrick, G. (1997). Developmental genome reorganization in ciliated protozoa: the transposon link. In Progress in Nucleic Acid Research and Molecular Biology, W.E. Cohn and K. Moldave, eds. (Waltham, MA: Academic Press), pp. 1–62.

Klobutcher, L.A., Huff, M.E., and Gonye, G.E. (1988). Alternative use of chromosome fragmentation sites in the ciliated protozoan *Oxytricha nova*. Nucleic Acids Res. *16*, 251–264.

Kumar, A. (2009). An overview of nested genes in eukaryotic genomes. Eukaryot. Cell *8*, 1321–1329.

Kuo, S., Chang, W.-J., and Landweber, L.F. (2006). Complex germline architecture: two genes intertwined on two loci. Mol. Biol. Evol. *23*, 4–6.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

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

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. *20*, 265–272.

Meyer, K.D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C.E., and Jaffrey, S.R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell *149*, 1635–1646.

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics *24*, 2818–2824.

Nowacki, M., Vijayan, V., Zhou, Y., Schotanus, K., Doak, T.G., and Landweber, L.F. (2008). RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature *451*, 153–158.

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

Nowacki, M., Haye, J.E., Fang, W., Vijayan, V., and Landweber, L.F. (2010). RNA-mediated epigenetic regulation of DNA copy number. Proc. Natl. Acad. Sci. USA *107*, 22140–22144.

O'Huallachain, M., Karczewski, K.J., Weissman, S.M., Urban, A.E., and Snyder, M.P. (2012). Extensive genetic variation in somatic human tissues. Proc. Natl. Acad. Sci. USA *109*, 18018–18023.

Prescott, D.M. (1994). The DNA of ciliated protozoa. Microbiol. Rev. *58*, 233–267.

Prescott, D.M. (1999). The evolutionary scrambling and developmental unscrambling of germline genes in hypotrichous ciliates. Nucleic Acids Res. *27*, 1243–1250.

Smith, J.J., Baker, C., Eichler, E.E., and Amemiya, C.T. (2012). Genetic consequences of programmed genome rearrangement. Curr. Biol. *22*, 1524–1529.

Spear, B.B., and Lauth, M.R. (1976). Polytene chromosomes of *Oxytricha*: biochemical and morphological changes during macronuclear development in a ciliated protozoan. Chromosoma *54*, 1–13.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. *33*, W465–W467.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27–40.

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

Wang, J., Mitreva, M., Berriman, M., Thorne, A., Magrini, V., Koutsovoulos, G., Kumar, S., Blaxter, M.L., and Davis, R.E. (2012). Silencing of germline-expressed genes by DNA elimination in somatic cells. Dev. Cell *23*, 1072–1080.

Witherspoon, D.J., Doak, T.G., Williams, K.R., Seegmiller, A., Seger, J., and Herrick, G. (1997). Selection on the protein-coding genes of the TBE1 family of transposable elements in the ciliates *Oxytricha fallax* and *O. trifallax*. Mol. Biol. Evol. *14*, 696–706.

Yu, P., Ma, D., and Xu, M. (2005). Nested genes in the human genome. Genomics *86*, 414–422.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics *29*, 2669–2677.

EXTENDED EXPERIMENTAL PROCEDURES

Nuclei Isolation, DNA Extraction, and Illumina Sequencing

Vegetative cultures of *Oxytricha* were grown in Pringsheim media, with the volume doubled each day. *Chlamydomonas reinhardtii* was the primary food source, supplemented occasionally with fresh overnight cultures of *Klebsiella pneumoniae* to boost growth. *Oxytricha* were grown to approximately 20 × 10^6 total cells and then moved to a 4°C room for a 3 day starvation period to allow the cells to digest the majority of bacterial and algal food. After the first day at 4°C, the cells were filtered through gauze to eliminate algae clumps and remove excess uneaten algae. At the end of 3 days at 4°C, the cells were again gauze-filtered, collected on a 10μm Nytex filter and washed 2X with Pringsheim media on-filter. The washed, starved *Oxytricha* cells were then collected in two 50ml conical tubes on ice, spun at 200 g for 1 min followed by removal of the supernatant (leaving approximately 2ml of packed cells). Approximately 45ml of Lysis buffer (3% sucrose, 0.2% Triton X-100, 0.01% Spermidine-trihydrochloride (Lauth et al., 1976)) was added to each tube and lysis was allowed to proceed, with gentle inversion mixing and occasional vortex, for 45min-1hr on ice. The completeness of the lysis was verified by microscopic examination with 5ng/ul DAPI, and the lysed material was layered onto 10%–40% discontinuous sucrose gradients exactly as described (Lauth et al., 1976), and centrifuged at 250 g for 10min in a Sorvall SH3000. The 10% sucrose fractions were collected (avoiding the 10%–40% interface as well as the 3%–10% interface) into fresh 50ml conical tubes and spun at 4700rpm for 10min. The supernatant was gently poured off and the resulting pellet was resuspended in buffer T1 of the Nucleospin Tissue Kit (#740952.250, Macherey-Nagel, Bethlehem, PA, USA). The standard protocol was followed for DNA preparation (proteinase K treatement, lysis and purification over a column). DNA was then run on a 0.3% SeaKem Gold agarose gel to purify large molecular weight components. The gel was imaged with Ethidium Bromide on a UV transiluminator to identify the limit-mobility band, which was excised and eluted from the gel using the QIAGEN Gel purification kit. The resulting material was submitted to the Princeton University sequencing facility for Illumina library construction and paired-end sequencing. An initial MIC DNA purification (from approximately 20 × 10^6 total cells) was used to generate libraries with inserts of 300 and 500bp, and a second MIC DNA purification (again from approximately 20 × 10^6 total cells) was used to create libraries of 350, 400, and 450 bp. These libraries were run in multiplex mode on two lanes of a HiSeq 2000 Illumina machine.

PacBio Sequencing of Micronucleus-Enriched Genomic DNA

Oxytricha trifallax DNA library preparation and sequencing was performed according to the manufacturer's instructions with the P5-C3 sequencing enzyme and chemistry, respectively. 5μg of sucrose-purified, high-quality, micronuclear-enriched DNA was verified using high sensitivity Qubit analysis to quantify the mass of double-stranded DNA present. After quantification, DNA was diluted to 150 μl using QIAGEN elution buffer at 33 μg/μL. The 150 μl aliquot was individually pipetted into the top chamber of a Covaris G-tube spin column and sheared gently for 60 s at 4500 rpm using an Eppendorf 5424 bench top centrifuge. Once completed, the spin column was flipped after verifying that all DNA was now in the lower chamber. Then, the column was spun for another 60 s at 4500 rpm to further shear the DNA and place the aliquot back into the upper chamber, resulting in a 10,000 to 20,000 bp DNA shear, verified using a DNA 12000 Agilent Bioanalyzer gel chip. The sheared DNA isolates were then repurified using a 0.45X AMPure XP purification step (0.45X AMPure beads added, by volume, to each DNA sample dissolved in 200 μl EB, vortexed for 10 min at 2,000 rpm, followed by two washes with 70% alcohol and finally diluted in EB). This AMPure XP purification step assures removal of any small fragment and/or biological contaminant.

After purification, ~1.9 ug of purified and sheared sample was taken into DNA damage and end-repair. Briefly, the DNA fragments were repaired using DNA damage repair solution (1X DNA damage repair buffer, 1X NAD+, 1 mM ATP high, 0.1 mM dNTP, and 1X DNA damage repair mix) with a volume of 21.1 μl and incubated at 37°C for 20 min. DNA ends were repaired next by adding 1X end repair mix to the solution, which was incubated at 25°C for 5 min, followed by the second 0.45X Ampure XP purification step. Next, 0.75 μM of blunt adaptor was added to the DNA, followed by 1X template preparation buffer, 0.05 mM ATP low and 0.75 U/μL T4 ligase to ligate (final volume of 47.5 μL) the SMRTbell adapters to the DNA fragments. This solution was incubated at 25°C overnight, followed by a 65°C 10 min ligase denaturation step. After ligation, the library was treated with an exonuclease cocktail to remove unligated DNA fragments using a solution of 1.81 U/μL Exo III 18 and 0.18 U/μL Exo VII, then incubated at 37°C for 1 hr. Two additional 0.45X Ampure XP purifications steps were performed to remove < 2000 bp molecular weight DNA and organic contaminant.

Upon completion of library construction, samples were validated as ~20 kb using another Agilent DNA 12000 gel chip. The micronucleus-enriched library was sufficient for additional size selection to remove any library molecules < 7,000 bp. This step was conducted using Sage Science Blue Pippin 0.75% agarose cassettes to select library in the range of 7,000–50,000 bp. This selection is necessary to narrow the library distribution and maximize the SMRTbell sub-readlength for the best de novo assembly possible. Without selection, smaller 2000–7000 bp molecules will dominate the zero-mode waveguide loading distribution, decreasing the sub-readlength. About 18% of the input library resulted after elution from the agarose cassette, which was enough yield to proceed to primer annealing and DNA sequencing on the PacBio RSII machine. Size-selection was confirmed by Bio-Analysis and the mass was quantified using the aforementioned Qubit assay.

Then, primer was annealed to the size-selected SMRTbell with the full-length libraries (80°C for 2 min 30 followed by decreasing the temperature by 0.1°/s to 25C°). The polymerase-template complex was then bound to the P5 enzyme using a ratio of 10:1 polymerase to SMRTbell at 0.5 nM for 4 hr at 30°C and then held at 4°C until ready for magbead loading, prior to sequencing. The magnetic

bead-loading step was conducted at 4°C for 60 min. The magbead-loaded, polymerase-bound, SMRTbell libraries were placed onto the RSII machine at a sequencing concentration of 50 pM and configured for a 180 min continuous sequencing run. Sequencing was conducted to ample coverage across two SMRTcells for each isolate. Data were then generated and filtered at 0.75 RQ and 500 bp sub-readlength.

Genome Assembly

Raw Illumina sequencing reads were trimmed with the preprocess module from the SGA package (Simpson and Durbin, 2012) (quality score threshold 30). Illumina reads were assembled into contigs using MaSuRCA (verson 2.1.0, cgwErrorRate = 0.15, USE_LINKING_MATES = 0) (Zimin et al., 2013). The MaSuRCA assembly contains 145,639 contigs with a N50 of 7.1kb (total bases 393.2Mb). The smaller genome size than estimated is likely due to collapsing of repetitive sequences. PacBio reads were first mapped with BLASTN (BLAST+ (Camacho et al., 2009)) to a combined database of the MAC genome and the MIC MaSuRCA assembly and assigned to either MAC or MIC sequences based on their best hit. BLASTN was used here instead of BLASR (Chaisson and Tesler, 2012) because BLASR allows many more gaps, but we rely on MDS-IES structures that break alignments to distinguish a read as a MAC or MIC derived sequence. 46,978 out of 1,774,906 PacBio templates longer than 3kb were assigned to MAC and filtered out. MIC derived PacBio reads longer than 3kb were error corrected using high identity unitigs (from the MaSuRCA output) with ectools (https://github.com/jgurtowski/ectools) using default parameters (PRE_DELTA_FILTER = false). Error corrected MIC reads were assembled with the Celera assembler (merSize = 14, unitigger = bogart, ovlMinLen = 1500) (Miller et al., 2008).

The final assembly covers most of the entire MAC genome assembly (98.9%) and complete MDS-IES maps were resolved for the majority (16,220, 88.1%) of 18,405 MAC contigs with one or both telomeres. The MIC precursors for 1615 (8.8%) MAC contigs are present on two MIC contigs; therefore, the distance separating them is unknown. For the remaining 570 (3.1%) MAC contigs, 262 (1.4%) map completely to more than two MIC contigs, while 308 (1.7%) are incompletely (<90%) covered in the current MIC genome assembly. This set of incompletely mapped MAC contigs could potentially include mis-assemblies or contaminations in the MAC genome assembly.

The final assembly was mapped to the NCBI nonredundant protein database using BLASTX (BLAST+ (Camacho et al., 2009), parameter: -query_gencode 6). An E-value threshold of e-9 was applied to filter out bacterial contigs. The assembly was also mapped to the *Oxytricha* mitochondrial genome (Swart et al., 2012) (GenBank accession: JN383843) and the *Chlamydomona reinhardtii* genome (Merchant et al., 2007) (version 4.0 - http://genome.jgi-psf.org/Chlre4/download/Chlre4_genomic_scaffolds.fasta.gz) to filter out these contaminants.

We used BWA (Li and Durbin, 2009) to map raw reads to the assembly with default parameters. SAMtools (Li et al., 2009) was used to process the alignment files and calculate the coverage at each base that is not masked by RepeatMasker. The median coverage for each contig can be found at: http://trifallax.princeton.edu/cms/raw-data/genome/mic/Oxytricha_trifallax_micronuclear_genome_contig_coverage.txt. SAMtools was used for SNP identification. Candidate SNPs were filtered with the following criteria: (1) total depth is $\geq$ 10 and $\leq$ 140; (2) root mean square of mapping quality is $\geq$ 20; (3) number of supporting reads for either the reference or the alternate bases should be $\geq$ 4; (4) the SNP is in unmasked regions. The identified SNPs in vcf format are documented at http://trifallax.princeton.edu/cms/raw-data/genome/mic/Oxytricha_trifallax_micronuclear_genome_SNP.vcf

Estimation of Genome Size

The MIC genome was previously estimated to be 0.3-2.3Gb through reassociation kinetics (Lauth et al., 1976). The size of the haploid genomes (G) in bp is estimated as G = B × P/X (Lander and Waterman, 1988), where B is the total bases in raw reads, P is the percentage of *Oxytricha* reads, and X is the estimated base coverage, assuming equal representation of all types of MIC elements in the raw reads. By mapping raw reads to the assembled bacterial contigs, we estimated the proportion of bacterial reads in the raw reads as 33.2% for insert sizes 300 and 500 bp; and 17.3% for insert sizes 350, 400 and 450 bp. Using this approach (Table S1), we estimated the haploid genome size to be ~490-500 Mb, which is close to half of the conventionally accepted estimate of the 1Gb diploid genome size (Prescott, 1994). The MIC genome assembly is 496Mb, which agrees with the estimated genome size.

Earlier studies reported approximately 120 germline chromosomes in *Oxytricha* (Spear and Lauth, 1976). We detected the MIC telomere sequence (long stretches of C4A4 repeats) on 2243 PacBio reads. After error correction, the remaining 1742 reads assembled into 149 distinct contigs corresponding to chromosome ends. This early analysis suggests that there exist at least 75 germline chromosomes.

BAC Library Construction and Sequencing

Oxytricha micronuclei were isolated as described under "Nuclei isolation, DNA extraction and sequencing" but the purified micronuclear pellet was gently resuspended in remnant liquid after the 4700rpm 10 min spin, transferred to a 1.5 ml Eppendorf tube for concentrating by gentle spin (4500rpm, 3min), and most liquid removed. A molten 2% low-melt InCert Agarose solution (Cambrex, Rockland, ME) was added at a 1:1 volume and mixed gently with the nuclear pellet, and immediately pipetted on ice into plug molds for pulse-field electrophoresis (BioRad, Hercules, CA) and the agarose was allowed to harden for 5 min. Plugs were released into 45ml of CTAB-PK solution (CTAB extraction buffer from Tek*nova*, Hollister, CA) supplemented with 0.2 mg/ml proteinase K and 1% SDS, incubated at 50°C for 2hr. The CTAB-PK was removed and fresh CTAB-PK was added, and incubated overnight. The plugs were then washed with 3-4 changes of TE, and transferred to 45ml of 1% lithium dodecyl sulfate, 10mM Tris pH 8.0, 100mM EDTA pH

8.0 and incubated at 37°C for 1hr. This buffer was then replaced and the incubation was continued at 37°C for 24-48hr. The blocks were then transferred to ~50ml of 20% N-laurylsarcosine, 2mM Tris pH 8.0, 140mM EDTA pH 8.0 and equilibrated for 2hr at room temperature on a gyratory shaker. This buffer was then replaced with fresh buffer, and the plugs were analyzed directly by pulse-field electrophoresis to isolate DNA of 100-200 kb. This material was cloned into the EcoRI site of the pCC1BAC vector (Epicenter, Madison WI) and transformed into DH10BT1 cells. Only a few hundred colonies were observed when DNA was prepared in this manner, prompting us to investigate fosmid strategies for large-insert library creation (see below). 96 BACs were Sanger end sequenced, including both randomly-selected clones with inserts and a set that hybridized to probes for either TBE1 transposons or other genes of interest. In addition, six BACs were completely sequenced on an Illumina GAIIx, using DNA prepared with the QIAGEN plasmid MIDI-prep kit (QIAGEN, Hilden, Germany) and ExoI treatment to reduce *E. coli* genomic DNA contamination, followed by standard Illumina fragmentation and library construction protocols. Inserts sizes of 400bp were used and each BAC library was constructed with a unique barcode for pooling. We also selected 6 BACs from the 96 end-sequenced BACs for complete PacBio sequencing, following the same procedures documented in "PacBio sequencing" section (Size selection was not performed).

Fosmid Library Construction and Clone Sequencing

DNA for fosmid library construction was prepared by sucrose fractionation of micronuclei as described above, followed by embedding in agarose plugs, as described for BAC library construction. Due to difficulties with in-plug restriction digestion during BAC library construction, we liberated large-size DNA molecules from plugs by phenol-chloroform extraction without electrophoresis. The material released in this way was fortuitously sheared to ~40kb fragments ideal for fosmid cloning with the pCC2fos Copy Control system (Epicenter). In total, 124,000 clones were gridded and spotted onto membranes.

For the pooled Illumina sequencing, 300 fosmids were randomly chosen and sequenced in indexed, overlapping subsets of 1, 6, 24, 60, 120 and 300 on an Illumina HiSeq 2000 with an insert size 300 bp. The 6-clone subset was contained in the 24-clone subset, and so on. In addition, 432 randomly selected fosmids (336 paired-end and an additional 96 single-end) were subjected to Sanger end-sequencing (DeWalch, Houston, TX and Macrogen, Rockvill, MD). 399 end sequences were sufficient sequence quality; of these, 218 were of bacterial origin. The remaining 181 end sequences were analyzed to infer genome composition as described in "Genome composition" section, below.

Use of BAC and Fosmid Sequences for Validation of the Assembly

We validated a small portion of the MIC assembly by comparing it to the assembled contigs derived from sequenced BACs and fosmids (listed in Table S1). Illumina reads for 6 BACs and 300 fosmids were assembled with IDBA (version 1.1.0) (Peng et al., 2012) using default parameters. We used HGAP assembler (Chin et al., 2013) to assemble PacBio reads sequenced from 6 additional BACs following procedures described in https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP-Whitelisting-Tutorial. Fosmid or BAC contigs were mapped to the MIC genome assembly with BLASTN (-word_size 100). We require a proper alignment to be longer than 90% of the contig length and the percent identity to be larger than 95%. In cases where a fosmid/BAC contig is not contained in one MIC contig but spans multiple MIC contigs, we require the final nucleotide of an alignment to be no more than 100 bp from either end of the reference or the query.

Among the 104 fosmid contigs (after exluding bacterial sequences) larger than 1kb (total bases 604.7kb), 96 (528.9 kb) map properly to the assembly, validating 523.5kb of sequence on 73 MIC contigs. Five other contigs (58.1kb) contain 13.9 kb of repetitive sequences that map improperly to many places in the assembly; however, sequence outside the repetitive regions map properly, validating 41.8kb on 8 MIC contigs. The remaining 3 fosmid contigs (12.8kb, 2.5kb, and 2.4kb) correspond entirely to repetitive sequence. It is also possible that the Illumina assembly of fosmids is inaccurate in repetitive regions.

Among the 14 BAC Illumina contigs larger than 1kb (total bases 109.7kb), 8 (35.2 kb) map properly to the assembly, validating 34.3kb of sequence on 8 MIC contigs. Four other contigs (61.6kb) contain 5.0 kb of repetitive sequences and the remaining nonrepeptive regions map properly, validating 55.7 kb of sequence on 6 MIC contigs. The remaining two BAC contigs (5.1kb and 7.8kb) correspond entirely to repetitive sequences.

PacBio sequencing of six BAC clones produced 7 long contigs larger than 20kb (total bases 397.2kb). They all contained long regions of satellite repeats (>40kb in 3 cases) that usually split contigs, and some also contain long regions of TBEs. This further documents the presence of large repetitive regions in the genome, as also inferred from fosmid inserts with both ends mapping to the same class of satellite repeats (described in "Genome Composition" section, below). The nonrepetitive regions all map properly to the MIC assembly, validating 138.4kb of sequence on 6 MIC contigs.

The limited number of fosmid and BAC end sequences (285 (107 pairs) and 119 (58 pairs) respectively, after eliminating bacterial sequences) were not as useful for validating the assembly, since, in most cases, at least one end maps to a repetitive sequence.

In total, BAC and fosmid sequences validated 793.7kb of the assembly.

Experimental Validation of a MIC Contig

We used PCR to validate a particularly complex 39.5kb MDS-rich region on a 54kb MIC contig (ctg7180000089708) (MDS-IES map available at http://trifallax.princeton.edu/cms/raw-data/genome/mic/Oxytricha_trifallax_micronuclear_genome_MDS_IES_maps.gff), using MIC-specific overlapping primer sets. The resulting products were approximately 1 to 4 kb in length, visualized by agarose gel electrophoresis. The end sequences of PCR products were mapped to the MIC assembly for validation, using Geneious (version

5.6) (Biomatters, http://www.geneious.com/). If needed, the PCR products were cloned into a TOPO vector (pCRTM2.1-TOPO® TA Vector, Invitrogen), transformed into One Shot® TOP10 Chemically Competent *E. coli* cells (Invitrogen), and amplified by colony PCR for direct Sanger end-sequencing. 37 kb of the 39.5 kb region was successfully validated, including 107 of 110 MDSs (67 of which are scrambled). Extensive paralogy at the 3′ end of this region precluded further validation by PCR.

Preprocessing of the MAC Genome Assembly before Mapping to the MIC Assembly

The MAC genome assembly was clustered using CD-HIT (Fu et al., 2012) at 95% similarity (-c 0.95 -aS 0.9 -uS 0.1) to remove redundant contigs. Contigs shorter than 200 bp were ignored. Repetitive contigs assembled from MIC contamination and bacterial contigs were also removed. For MDS annotation we only considered MAC contigs with one or both telomeres. This reduced the total number of MAC contigs for annotation from 22,450 to 18,405. MAC contigs were reoriented in the direction of the encoded gene, or the longest gene for multigene chromosomes. MAC contig names start with "Contig" and MIC contig names start with "ctg" or "deg."

piRNA Mapping

Non-redundant 27 nt Piwi (Otiwi1) co-immunoprecipitated sRNA (piRNA) reads (extracted from a conjugating time course between cells of JRB310 and JRB510 strains, Fang et al., 2012) were mapped to the MIC assembly using Bowtie2 (Langmead and Salzberg, 2012) with parameters "-N 1 -L 20." piRNA reads mapped with more than one mismatch were filtered out. 19,676,712 mapped reads were assigned to MDS (18,896,399) versus non-MDS (780,313) regions. Reads that mapped to non-MDS regions were subsequently mapped to the JRB510 MAC genome draft assembly (manuscript in preparation) with Bowtie2 and filtered as described above. 469,301 mapped to the JRB510 MAC assembly, suggesting that they belong to MIC regions that are either MAC-destined in JRB510 or were missed in the JRB310 MAC assembly.

MDS Sharing

The MAC genome assembly was aligned to itself using BLASTN. We selected pairs of MAC contigs with > 96% identify over a region larger than 200 bp and smaller than 80% of the length of the shortest contig in the pair. We then compared MDS patterns of their MIC precursors and examined whether their identical regions derive from the same MDSs on the same MIC contigs. To ensure that the shared MDSs are present in single copy in the MIC, we checked that the read coverage of shared MDSs is similar to that of unique MDSs.

Chromosome Fragmentation Sites

MIC sequences flanking terminal MDSs were first identified from the BLAST output produced from mapping the MAC genome against the MIC genome. MAC Illumina sequencing reads containing telomeres were mapped to MIC sequences flanking terminal MDSs with BLAT (Kent, 2002) (default parameters). The most frequently used telomere addition sites for each nanochromosome end were extracted from the mapping output and treated as chromosome fragmentation sites. The distance between adjacent telomere addition sites as well as the distance between telomere addition site and adjacent internal MDS ends were calculated and plotted. For each telomere addition site, 200bp flanking sequences were extracted to plot base frequency. Sequence logos for 20bp flanking sequences were produced with Weblogo (Crooks et al., 2004).

For alternatively fragmented sites, because the MAC assembly may have mis-incorporated telomeric sequences, we extracted MIC precursor sequences (excluding IESs) corresponding to MAC regions with alternative fragmentation sites (Swart et al., 2013). We then mapped all MAC telomeric reads to these sequences and extracted the most frequently used telomere addition sites.

Previous studies identified sequences that carry the specific signal for chromosome breakage and telomere addition in *Tetrahymena* (Yao et al., 1990) (in MIC-limited sequences flanking MAC precursors) and *Euplotes* (Klobutcher et al., 1998) (mostly in MAC subtelomeric regions but sometimes in MIC-limited sequences). A chromosome breakage signal similar to that of *Euplotes* may be present in *Stylonychia* (Jönsson et al., 2001), a genus more closely related to *Oxytricha*. We searched for *cis*-acting sequences associated with telomere addition sites in *Oxytricha*. We first examined cases where terminal MDSs are adjacent to each other. For 10,468 chromosome fragmentation sites < 5 bp away from the terminal MDS, the base composition flanking the fragmentation site reflects a complementary pattern that was observed in subtelomeric regions of MAC chromosomes (Cavalcanti et al., 2004; Swart et al., 2013), with a purine bias downstream of telomeres and a pyrimidine bias upstream (Figure S3A). When two terminal MDSs are close to each other, the micronuclear region downstream of the fragmentation site for one nanochromosome locus is also the macronuclear-destined subtelomeric region of another nanochromosome. To exclude overlapping signals, we examined 3,215 chromosome fragmentation sites that are > 100bp away from adjacent terminal MDSs. Notably, while the regions upstream of fragmentation sites reflect the typical pyrimidine bias upstream of telomeres, there is still a dramatic purine bias in the MIC-limited regions downstream of a fragmentation site (Figure S3B). We also examined 2,875 chromosome fragmentation sites next to internal MDSs. These cases display a strong purine-to-pyrimidine bias downstream (Figure S3C). While motif finding algorithms failed to identify any significant motifs in regions flanking fragmentation sites, it is possible that the sharp switch from pyrimidine-richness to purine-richness facilitates chromosome fragmentation via a signal in DNA or chromatin structure or by enhancing binding or recruitment of proteins involved in fragmentation.

Approximately 10% of *Oxytricha*'s nanochromosomes are alternatively fragmented (Swart et al., 2013). We analyzed base composition flanking 3,195 alternatively fragmented sites and found a much weaker complementary sequence bias than non-alternatively

fragmented sites (Figure S3D). This supports the hypothesis that fragmentation at these sites is less frequent because of a weaker signal (Williams et al., 2002).

Analysis of Repeat Content and Transposable Elements

RepeatModeler (Smit and Hubley, 2008–2010) was used with default parameters to identify de novo repetitive DNA regions in the MIC genome. This program uses two de novo repeat finding programs RECON (Bao and Eddy, 2002) and RepeatScout (Price et al., 2005). RepeatModeler generated a library of 398 consensus sequences. We then filtered out from this library repeats shorter than 100 bp or having significant hits to known proteins in the NCBI nr protein database other than proteins known as belonging to transposable elements (TEs).

The 301 remaining consensus sequences were annotated by a TBLASTX search against RepBase (Jurka et al., 2005), BLASTX to the NCBI nr database (BLAST+ (Camacho et al., 2009), parameters: -query_gencode 6 -evalue 1e-7) and a HMMER (version 3.0) (Finn et al., 2011) search against the Pfam-A profile HMM database (version 26.0) (Punta et al., 2012). The consensus sequences were also analyzed by Transposon-PSI (http://transposonpsi.sourceforge.net), a program that performs TBLASTN searches using a set of position specific scoring matrices (PSSMs) specific for different TE families. 31 consensus sequences were classified as TBE transposons; 3 as *Helitrons,* and 4 as LINE elements.

Another 3 consensus sequences contain DDE_Tnp_IS1595 protein domains. Two contain long terminal inverted repeats (261 and 116 bp) identified by the einverted program from the EMBOSS program suite (Rice et al., 2000) and have weak BLASTX hits to the ISBun2 insertion sequence from the IS1595 family (E-value = 3e-6 and 7e-5, respectively) in ISfinder (Siguier et al., 2006). Thus we classified them as insertion sequences (IS). The third consensus sequence is short (873 bp), lacking terminal inverted repeats and ISfinder matches. These 3 consensus sequences use the *Oxytricha* genetic code and do not match bacterial sequences, so they are not likely contaminants.

We also identified the two most abundant classes of satellite repeats: a 380 bp satellite repeat and a 170 bp satellite repeat, both of which have previously been characterized by hybridization (Dawson et al., 1984) but not sequencing. The remaining 252 consensus sequences were not categorized.

The number of TE occurrences and the percent of genome coverage were assessed by masking the genome assembly using RepeatMasker (Smit et al., 1996–2010) with the 301 consensus sequences. We parsed the RepeatMasker output with several criteria: if the repeat query was shorter than 500 bp, the length of the matched region had to be longer than 50 bp to be counted; if the repeat query was longer than 500 bp, the length of the matched region had to be longer than 100 bp to be counted. RepeatMasker masked 35.9% of the genome assembly (Table S8).

To further classify TBE elements, we took the 42kDa transposases coding sequences from (Nowacki et al., 2009) and queried them against the MIC genome assembly. Clustering 812 putative 42kDa transposase coding sequences in the MIC with USEARCH (Edgar, 2010) produced three families, corresponding to TBE1, 2 and 3. We extracted 22kDa and 57kDa genes encoded close to these 42kDa transposases genes and queried their consensus sequences against the MIC assembly. The 888 22kDa genes clustered into three families corresponding to TBE1, 2 and 3, while the 732 57kDa genes clustered into four families, corresponding to TBE1, two subfamilies within TBE2, and TBE3. Therefore, all TBE elements in the MIC genome can be classified into the three original families (with 2 subfamilies within TBE2). The three TBE families (TBE1, TBE2, and TBE3) have distinct terminal inverted repeats (the first 17 bp sequence is always the *Oxytricha* telomeric repeat $CA_4C_4A_4C_4$).

We also looked for LINE elements and *Helitrons* in the *Paramecium* MAC genome assembly, as well as the *Tetrahymena* MAC and MIC genome assemblies (GenBank accession: GCA_000261185.1) using Transposon-PSI (http://transposonpsi.sourceforge.net).

Genome Composition

We inferred the percentage of TBE, satellite repeats, other repetitive elements, and MIC-limited noncoding, non-repetitive regions in the annotated assembly from the RepeatMasker output. We also used three other approaches to assess genome composition: Illumina reads, error-corrected PacBio reads and fosmid end sequences (Figure S1; Table S8).

For Illumina reads, 12 million reads were randomly sampled and mapped with BLAT (-minIdentity = 98) to the MAC genome and with BLASTN to MIC genome sequences masked by RepeatMasker as repetitive elements. The percentage of reads that mapped to each category was calculated and plotted. For mapping against the MAC genome, we counted the length of covered nucleotides on Illumina reads to avoid including IESs. Fosmid end sequences (one end randomly selected from each fosmid to avoid double sampling) were mapped with BLASTN to the MAC genome and to MIC repetitive regions, and analyzed in a similar way to Illumina reads to infer genome composition. We randomly sampled 145,545 error-corrected PacBio reads and analyzed them with RepeatMasker (Smit et al., 1996–2010) in a similar way to the assembly. These reads were also mapped with BLASTN to the MAC genome to assess the percentage of nucleotides corresponding to MDS regions aligned with at least 98% identity.

For the portion of MDSs in the MIC genome, the genome assembly, Illumina reads and PacBio reads all give an estimate of ~10% (11.1%, 9.2% and 9.9% respectively), whereas 5.1% was estimated by the fosmid end mapping approach. These numbers fall in the estimated range given by the original re-association kinetics study (2.4% −18%) (Lauth et al., 1976), though most literature cites this paper and describes the MAC-destined portion of the MIC genome as 5%. In terms of repeat content, the MIC genome assembly consists of 15.6% TBEs, 3.3% satellite repeats and 17.0% other repetitive elements (overall 35.9% repetitive). Illumina read mapping suggests that the MIC genome consists of 15.0% TBEs, 17.9% satellite repeat and 14.3% other repetitive elements (overall 47.3%

repetitive), while PacBio read analysis suggests that the MIC genome contains 15.2% TBEs, 10.3% satellite repeats and 15.2% other repetitive elements (overall 40.7% repetitive). Mapping of fosmid end sequences, on the other hand, a much smaller sample, suggests that the MIC genome comprises 8.4% TBEs, 64.6% satellite repeats and 4.5% other repetitive elements (overall 77.5% repetitive). As a result, the percentage of remaining MIC-limited noncoding non-repetitive regions differs noticeably (53% for the assembly, 43.5% for Illumina reads, 49.4% for PacBio reads and 17.4% for fosmid ends); however, these methods have their own biases, which we discuss below.

The assembly is biased toward sequences that are more easily sequenced and assembled; hence it is more depleted of repetitive elements, which will tend to collapse. The MDS containing portion of the MIC genome is gene rich and also repeat-poor, which would favor assembly, potentially inflating the percentage of MDSs in the assembly.

The fosmid-end Sanger sequencing approach is subject to cloning biases. AT-rich regions are recalcitrant to cloning (Eichinger et al., 2005). In addition, the fosmids were constructed by natural breakage of long MIC DNA. Therefore, the fosmid ends may not represent a random set of sequences from the genome, but rather regions that are prone to breakage, such as fragile sites. Furthermore, the sample size of the fosmid ends is small (178 end sequences, total 129kb sequenced, which is just ~0.026% the size of the assembly). The sequenced fosmid ends do not include any of the identified *Helitrons*, ISs and LINE elements and they also have poor coverage of uncategorized repetitive elements (Table S8), suggesting too small a sample size to contribute to this assessment. Without a more comprehensive genome sampling, it is difficult to conclude that 181 fosmid end sequences capture the true MIC genome composition. Notably, however, 42 fosmids contain the same identical sequence repeat at both ends of their 40 kb insert (21 fosmid contains the 380 bp repeat and 21 contain the 170 bp repeat). Though these fosmids could not be completely assembled, it suggests the possibility that some 40 kb genomic regions could be complete concatamers of such repeats, suggesting the presence of large, repeat-dense regions of the genome.

The sampled Illumina reads (1.2Gb) and PacBio reads (1.1Gb) offer better coverage of the genome than the limited number of fosmid end sequences. Illumina sequencing can suffer from PCR biases against regions with high (>56%) or low GC content (<11%) (Aird et al., 2011), but the GC content of the MIC genome falls in a range that is not typically strongly biased against (lowest GC: IES regions ~18%; highest GC: 170bp satellite repeat 52%). PacBio is reported to be the least biased among current sequencing technologies (Ross et al., 2013), free from amplification bias and with the least GC bias. However, because PacBio reads were error-corrected using Illumina reads, any bias in Illumina sequencing may have influenced corrected PacBio reads.

With these biases in mind, we infer from the four methods that approximately half or more of the MIC genome is comprised of repetitive elements and that the macronuclear-destined portion is as much as 10%.

We also assessed the fraction of IESs in the MIC genome. We calculated the MDS to IES ratio for 11,361 MAC chromosomal loci whose MDS-IES structures are completely mapped and whose IESs do not contain MDSs for other MAC chromosomes. This approach identified 31.5 Mb of MDSs and 13.6 Mb of IESs for this portion of the data. Therefore, we estimate that IESs occupy approximately 43% as much space in the genome as MDSs. This implicates ~4% of the MIC genome as IES, if MDSs represent ~10%.

Noncoding RNA Annotation

We searched the MIC assembly for ribosomal RNAs by performing BLAST searches against the small and large subunit ribosomal RNA sequences downloaded from SILVA rRNA database (Pruesse et al., 2007). We used tRNAscan (Lowe and Eddy, 1997) with default parameters to search for tRNAs. We also searched for other noncoding RNAs by running Infernal (version 1.1) (Nawrocki et al., 2009) against Rfam (version 11.0) (Burge et al., 2013). These searches did not identify any significant MIC-limited ncRNAs.

Micronuclear Gene Prediction

RNA-seq data from (Swart et al., 2013) were previously obtained from cells collected at various time points (vegetative "fed" stage for JRB 310 cells; 0, 10, 20, 40, 60 hr post mixing of starved, mating-compatible strains JRB310 and JRB510 for the conjugating time series) and processed as described in (Swart et al., 2013). We eliminated all reads that mapped to the *Oxytricha* macronuclear genome (Swart et al., 2013) (GenBank accession: AMCR00000000), the *Oxytricha* mitochondrial genome (Swart et al., 2012) (GenBank accession: JN383843) and the *Chlamydomonas reinhardtii* genome (Merchant et al., 2007) (version 4.0 - http://genome.jgi-psf.org/Chlre4/download/Chlre4_genomic_scaffolds.fasta.gz). The remaining reads were assembled using SOAPdenovo-Trans (Li et al., 2010) using a k-mer size of 25. We also assembled the RNA-seq reads with Trinity (Grabherr et al., 2011) using default parameters. The assembled transcripts by SOAPdenovo-Trans and Trinity were aligned to the genome and these transcript alignments were further assembled using the PASA pipeline (Haas et al., 2003) to produce more complete transcript structures.

Gene predictions were produced by AUGUSTUS (version 2.5.5) (Stanke and Morgenstern, 2005) using transcripts provided by PASA as hints for predictions. Hints were generated from the PASA output using a custom Python script. We used the *Oxytricha* model that was previously trained for MAC genome annotation (Swart et al., 2013). By examining PASA transcript alignments, we found relatively few intron-spliced genes in the MIC genome. To avoid over-prediction of introns, we changed 4 parameters in the extrinsic evidence configuration file to restrict intron predictions to only those that were supported by hints, namely 'intronpart malus' was reduced from 1 to 0.7, 'intron malus' was reduced from 0.34 to 0.1, 'intronpart bonus' was increased from 1 to 1e20 and 'intron bonus' was increased from 1e5 to 1e40. AUGUSTUS was run with–UTR = on and–alternatives- from-evidence = true. Gene models predicted by AUGUSTUS were filtered to only include models supported by hints. Any predicted genes supported by fewer than

5 RNA-seq reads were filtered out. Predicted coding sequences were clustered using CD-HIT (Fu et al., 2012) at 95% identity (-c 0.95 -aS 0.9 -uS 0.1) to remove redundancy.

DNA sequencing reads that derive from the macronuclear genome (Swart et al., 2013), as well as macronuclear sequencing of the JRB510 strain, were mapped to predicted coding sequences with BLAT (Kent, 2002) and filtered for > 94% identity. Any coding sequences that are more than 80% covered by MAC reads with more than 1x coverage were removed to filter out genes that are not MIC-specific.

Predicted genes associated with repetitive regions were identified by comparing their sequences to repeat consensus sequences generated by RepeatModeller and comparing their genomic locations with RepeatMasker output.

All predicted gene models were functionally annotated using BLASTP against the NCBI nr database (BLAST+ (Camacho et al., 2009), parameters: -query_gencode 6 -evalue 1e-7) and a HMMER (version 3.0) (Finn et al., 2011) search with default parameters against the Pfam-A profile HMM database (version 26.0) (Punta et al., 2012). The thresholds for considering domains were set as independent E-value $\leq$ 0.001 and conditional E-value $\leq$ 0.1 for at least one domain match in potentially repeated domains.

Protein sequences of predicted MIC and MAC genes were mapped to the UniRef90 database (ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref90/uniref90.fasta.gz) using BLASTP (BLAST+ (Camacho et al., 2009), parameters: -query_gencode 6 -evalue 1e-7) and then mapped to GO terms using the gene association file (ftp://ftp.geneontology.org/pub/go/gene-associations/gene_association.goa_uniprot) through AutoFACT (Koski et al., 2005). GO term enrichment (Fisher's Test) was performed with the goatools Python module (https://github.com/tanghaibao/goatools/) using MIC genes as the test set and all *Oxytricha* genes (MAC and MIC) as the reference-set (p value cutoff of 0.05).

RNA-seq reads from (Swart et al., 2013) were mapped to predicted coding sequences and the MAC genome using BLAT (Kent, 2002) and filtered with pslCDNAfilter (Karolchik et al., 2004) (minimum indentity > 0.92, min coverage > 0.8, max alignment = 3) and normalized with DESeq (Anders and Huber, 2010) (the default method), as described in (Swart et al., 2013). To produce the expression heat maps, for each MIC-encoded gene we calculated 100,000 × normalized RNA-seq counts divided by the coding sequence length. Expression patterns were clustered by the complete linkage clustering method and visualized as heat maps.

Proteomic Sample Preparation and Analysis by Mass Spectrometry

Exconjugant cells from a 40 hr post-mixing time point (grown and mixed as described in Fang et al. (2012) were harvested and washed on 10 μM Nytex filters as described in "Nuclei Isolation, DNA Extraction, and Illumina Sequencing" and their developing macronuclei were isolated by lysis and sucrose sedimentation through a 10%–40% discontinuous gradient as described above and in Lauth et al. (1976). While conceptually similar to the MIC purification to isolate DNA, here we retained only the pellet after the 250 g, 10 min spin in a Sorvall SH3000. This pellet is enriched in developing macronuclei (anlagen) as well as parental (old) macronuclei undergoing DNA degradation. The nuclear pellet was resuspended in a modified TMS buffer (as described in Zahler and Prescott, 1988): 0.01M Tris-Cl pH 8.0, 0.1M MgCl, 0.24M sucrose. The proteins from these resuspended nuclei were released upon snap-freezing in liquid nitrogen and ethanol, followed by a room-temperature thaw and pipetting 30-40 times. These raw lysates were cleared by centrifugation at 17,900 g for 2 min at 4°C and the supernatant was kept on ice and used for proteomic analysis.

Nuclear lysate was resolved by SDS-PAGE on a 4%–20% gradient Tris-glycine pre-cast gel from Thermo Scientific® until the dye front had migrated approximately 1 cm into the gel. Following a water wash, the protein-containing region of the gel lane was excised, sliced into 16 uniform slices, and each slice was further sliced and diced into 1 mm^3 cubes, and dehydrated in acetonitrile. Gel pieces were washed extensively, subjected to thiol reduction by Tris (2-carboxyethyl)phosphine hydrochloride) (TCEP), iodoacetamide thiol alkylation, and overnight trypsin digestion in gel, following the methods of (Shevchenko et al., 2006). Peptides were eluted from the gel pieces, concentrated, and then desalted using StageTip micro-scale reversed-phase chromatography (Rappsilber et al., 2003). Peptides from each gel slice were then subjected individually to reversed-phase nano-LC-MS and MS/MS performed on an Easy nLC 1000 Ultra nano-flow capillary ultra-high performance liquid chromatography (UPLC) system (ThermoFisher Scientific) coupled to a VelosPro-orbitrap Elite hybrid mass spectrometer (ThermoFisher Scientific) outfitted with a Flex ion source (Proxeon). Sample concentration and washing were performed online using a trapping capillary column (100 μm x ca. 40 mm, packed with 3 μm, 100 Å Magic AQ C18 resin (Michrom)) at a flow rate of 4 μL/minute with 3X the sample injection volume. Separation was achieved using an extended-length analytical capillary column (75 μm x ca. 45 cm, packed with 3 μm, 100 Å Magic AQ C18 resin (Michrom)), and a linear gradient from 5% to 35% solutions A and B (solution A: 100% water/ 0.1% formic acid; solution B: 100% acetonitrile/ 0.1% formic acid) applied over 180 min at a flow rate of approximately 300 nL/ minute. Nano-electrospray ionization was accomplished on the Orbi Elite platform at 2.4 kV using a capillary temperature of 275°C. Full-scan mass spectra were acquired in the Orbitrap in positive-ion mode over the *m/z* range of 335–1800 at a resolving power of 120,000. MS/MS spectra were simultaneously acquired using collision-induced dissociation (CID) in the VelosPro linear trap for the fifteen most abundant multiply charged species in the full-scan spectrum having signal intensities of > 1000 NL. All spectra were acquired in profile mode. Dynamic exclusion was set such that MS/MS was acquired only once for each species over a period of 120 s.

The resulting LC-MS/MS data were subjected to database searching and analysis using the framework of ProteomeDiscoverer software (v. 1.4, ThermoFisher). The search was conducted using a workflow employing a Mascot search engine server module (v. 2.4, Matrix Science) to search against an *Oxytricha* database consisting of protein sequences predicted to be encoded by the MAC concatenated with those of the MIC. Mascot search parameters included an initial mass error of 6 ppm for the precursor and 1.2 Da for the fragment ion species, $\leq$ 2 missed trypsin cleavages, carbamidomethylation of cysteines as a fixed modification,

with methionine oxidation and *N*-terminal protein acetylation as variable modifications. Aggregate search results were subjected to statistical analysis and filtering to $\leq$ 1% FDR on the peptide level using the Percolator support vector machine module (Käll et al., 2007), which employs semi-supervised machine learning to dynamically discriminate between correct and incorrect spectral matches, based on parameter optimization to minimize reversed-database decoy matches. Further posttranslationally modified peptides were gleaned from the data by searches conducted against the same database and allowing for either lysine acetylation or serine, threonine, or tyrosine phosphorylation as variable modifications. Additionally, modification of MIC-limited histones was determined through spectral matches to a MIC plus MAC histone database requiring no greater than 4 ppm error on the precursor m/z measurement and allowing for $\leq$ 8 missed trypsin cleavages, fixed cysteine carbamidomethylation, and variable methionine oxidation, *N*-terminal protein acetylation, acetylation and mono-, di-, and trimethylation of lysine, and mono- and di-methylation of arginine as variable modifications. All spectral matches were further filtered to only those of high confidence, which had a Mascot score greater than or equal to 30, and a precursor species with a charge state of +2 to +4 within the mass range of 600-4,000 Da. Fragmentation spectral assignments were subject to manual inspection and validation using ProteomeDiscoverer and the original tandem mass spectra acquired in profile mode, using Xcalibur software (ThermoFisher).

Additional Analysis of MDSs, IESs, and Pointers and Comparison to Paramecium

MDSs, IESs and pointers are three different categories of sequences whose definitions depend on the somatic genome sequence. Previously inferred properties of these sequences were based on a very small sample (Cavalcanti et al., 2005; Prescott, 1994; Prescott and DuBois, 1996). Here we investigate their sequence features on a genome-wide scale.

Among previously known pointers in *Oxytricha*, the shortest are 2 bp (Prescott and DuBois, 1996). We searched the MIC genome for the possible presence of shorter (1 or 0 bp) pointers and found a limited number of possible cases (882 and 330, respectively). These could result from assembly errors in either the MAC or MIC genomes, be part of longer pointers containing mismatches, or be explained by the occurrence of 0 bp MDSs (comprised of two tandem pointers with no intervening MAC sequence) that were missed in the annotation due to their short length. Therefore, we excluded 1 or 0 bp pointers from this study.

Nonscrambled pointers longer than 3 bp have a strong preference for purines (especially A) at the 5′ end and pyrimidines (especially T) at the 3′ end (Figure S2F). The bias becomes weaker toward the center of IESs. For scrambled pointers longer than 3 bp (Figure S2G), this bias is less prominent and only detectable at the first position at both ends. The sequence pattern could be a signal that facilitates recognition of nonscrambled pointers during genome rearrangement, since nonscrambled pointers are shorter and less GC-rich than scrambled pointers, and thus have lower sequence complexity.

The ends of all nonscrambled IESs in *Oxytricha* show a complementary base composition bias (Figure S2L). This pattern is also observed in *Oxytricha* scrambled IESs (Figure S2M) and *Paramecium* IESs (Figure S2N), though the base composition profiles differ. Note that *Parameium* IESs are all nonscrambled and use TA as a universal pointer, whereas different pointers are used in *Oxytricha* IESs. The complementary base biases might facilitate recognition by proteins involved in IES deletion (Arnaiz et al., 2012).

MDS sequences flanking pointers show a biased base composition. There is a strong increase in GC content at the two MDS sites immediately flanking pointers (a preference for C at the site upstream of the pointer and G at the site downstream of the pointer). This is true for both nonscrambled (Figure S2O) and scrambled pointers (Figure S2P). This sequence pattern could be a signal that helps mark the pointer-MDS boundary. A GC bias is also present in *Paramecium* (Figure S2Q) with an increase in G in the two sites upstream of the pointer and an increase in C in the two sites downstream of the pointer.

Paramecium MAC genome sequences containing IESs were downloaded from *Paramecium*DB (Arnaiz and Sperling, 2011): http://paramecium.cgm.cnrs-gif.fr/download/fasta/archive/genes_with_IES/Ptetraurelia_genes_with_IES_v1.66.fasta.bz2. For IESs, lowercase sequences were extracted excluding the TA pointer at both ends. For MDS-pointer junctions, 30 bp (all upper case) flanking lowercase sequences were extracted.

Assessment of Macronuclear Contamination

We mapped 80 million MIC reads to 107,787 sampled nonscrambled MDS-IES junctions and identified gapped alignments as potentially contaminating MAC reads. Among 934,571 reads that mapped across MDS-IES junctions, only 5,482 (0.587%) mapped with gaps corresponding to the IES length. Therefore, the level of MAC contamination is low.

Assessment of IES Excision Efficiency

We mapped 68 million MAC sequencing reads (Swart et al., 2013) to 107,787 sampled nonscrambled MDS-IES junctions. 10,582,764 out of 10,591,570 mapped with gaps, suggesting removed IESs. Thus the IES excision efficiency appears high (99.9%), unlike in *Paramecium*, where thousands of IESs were identified in MAC reads with ~13X macronuclear sequencing coverage (Duret et al., 2008). The retained IESs could also derive from trace amounts of micronuclear contamination in the MAC sequencing reads.

SUPPLEMENTAL REFERENCES

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. *12*, R18.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Arnaiz, O., and Sperling, L. (2011). *Paramecium*DB in 2011: new tools and new data for functional and comparative genomics of the model ciliate *Paramecium tetraurelia*. Nucleic Acids Res. *39*, D632–D636.

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

Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P., and Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. *41*, D226–D232.

Cavalcanti, A.R.O., Dunn, D.M., Weiss, R., Herrick, G., Landweber, L.F., and Doak, T.G. (2004). Sequence features of *Oxytricha trifallax* (class *Spirotrichea*) macronuclear telomeric and subtelomeric sequences. Protist *155*, 311–322.

Cavalcanti, A.R.O., Clarke, T.H., and Landweber, L.F. (2005). MDS_IES_DB: a database of macronuclear and micronuclear genes in spirotrichous ciliates. Nucleic Acids Res. *33*, D396–D398.

Chaisson, M.J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics *13*, 238.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569.

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

Duret, L., Cohen, J., Jubin, C., Dessen, P., Goût, J.-F., Mousset, S., Aury, J.-M., Jaillon, O., Noël, B., Arnaiz, O., et al. (2008). Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: a somatic view of the germline. Genome Res. *18*, 585–596.

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

Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., et al. (2005). The genome of the social amoeba *Dictyostelium discoideum*. Nature *435*, 43–57.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37.

Jönsson, F., Steinbrück, G., and Lipps, H.J. (2001). Both subtelomeric regions are required and sufficient for specific DNA fragmentation during macronuclear development in Stylonychia lemnae. Genome Biol. *2*, RESEARCH0005.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods *4*, 923–925.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, D493–D496.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Klobutcher, L.A., Gygax, S.E., Podoloff, J.D., Vermeesch, J.R., Price, C.M., Tebeau, C.M., and Jahn, C.L. (1998). Conserved DNA sequences adjacent to chromosome fragmentation and telomere addition sites in *Euplotes crassus*. Nucleic Acids Res. *26*, 4230–4240.

Koski, L.B., Gray, M.W., Lang, B.F., and Burger, G. (2005). AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics *6*, 151.

Lander, E.S., and Waterman, M.S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics *2*, 231–239.

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

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

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

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., et al. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science *318*, 245–250.

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics *25*, 1335–1337.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420–1428.

Prescott, D.M., and DuBois, M.L. (1996). Internal eliminated segments (IESs) of *Oxytrichidae*. J. Eukaryot. Microbiol. *43*, 432–441.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21* (*Suppl 1*), i351–i358.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glöckner, F.O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. *35*, 7188–7196.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. Nucleic Acids Res. *40*, D290–D301.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal. Chem. *75*, 663–670.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol. *14*, R51.

Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V., and Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat. Protoc. *1*, 2856–2860.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. *34*, D32–D36.

Simpson, J.T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. Genome Res. *22*, 549–556.

Smit, A., and Hubley, R. (2008–2010). RepeatModeler Open-1.0. http://www.repeatmasker.org.

Smit, A., Hubley, R., and Green, P. (1996–2010). RepeatMasker Open-3.0. http://www.repeatmasker.org/.

Swart, E.C., Nowacki, M., Shum, J., Stiles, H., Higgins, B.P., Doak, T.G., Schotanus, K., Magrini, V.J., Minx, P., Mardis, E.R., and Landweber, L.F. (2012). The *Oxytricha trifallax* mitochondrial genome. Genome Biol. Evol. *4*, 136–154.

Williams, K.R., Doak, T.G., and Herrick, G. (2002). Telomere formation on macronuclear chromosomes of *Oxytricha trifallax* and *O. fallax*: alternatively processed regions have multiple telomere addition sites. BMC Genet. *3*, 16.

Yao, M.C., Yao, C.H., and Monks, B. (1990). The controlling sequence for site-specific chromosome breakage in *Tetrahymena*. Cell *63*, 763–772.

Zahler, A.M., and Prescott, D.M. (1988). Telomere terminal transferase activity in the hypotrichous ciliate *Oxytricha nova* and a model for replication of the ends of linear DNA molecules. Nucleic Acids Res. *16* (14B), 6953–6972.

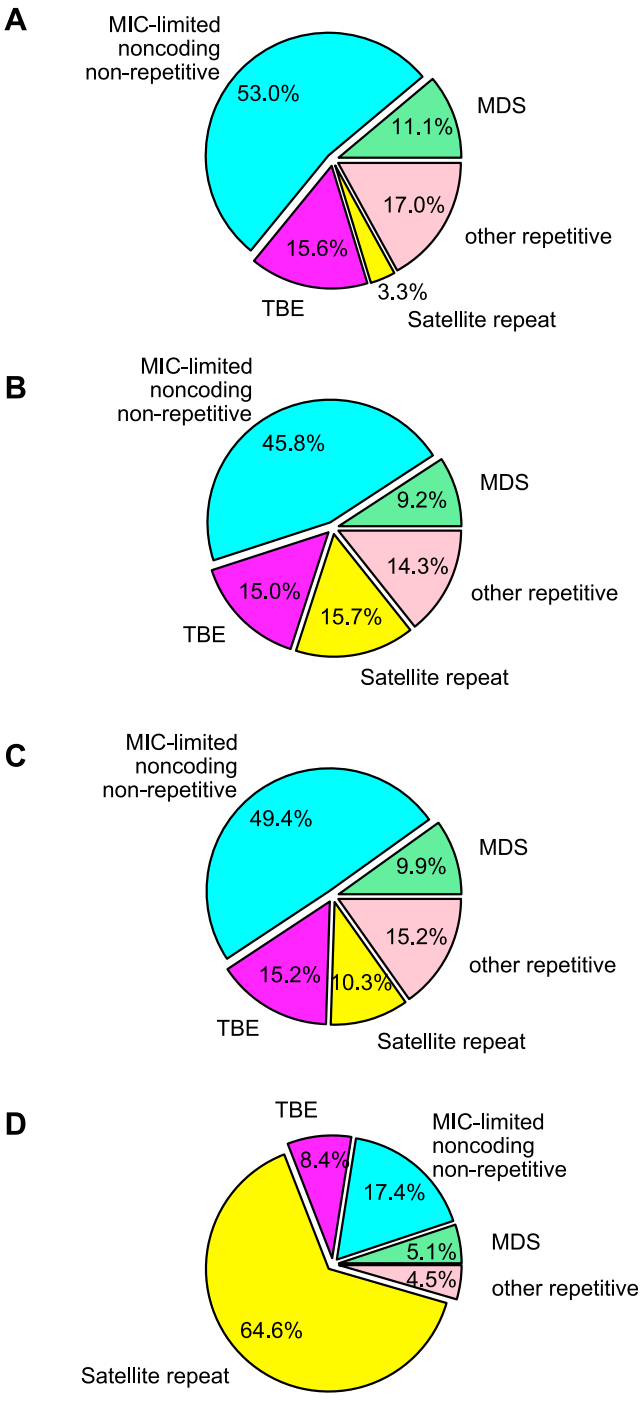

Figure S1. Comparison of Genome Composition Inferred from Four Data Sources, Related to Figure 1

Data sources include (A) genome assembly, (B) Illumina reads, (C) error-corrected PacBio reads, and (D) fosmid end sequences. The "other repetitive" category includes the LINES, *Helitrons*, ISs, and uncategorized repetitive elements.

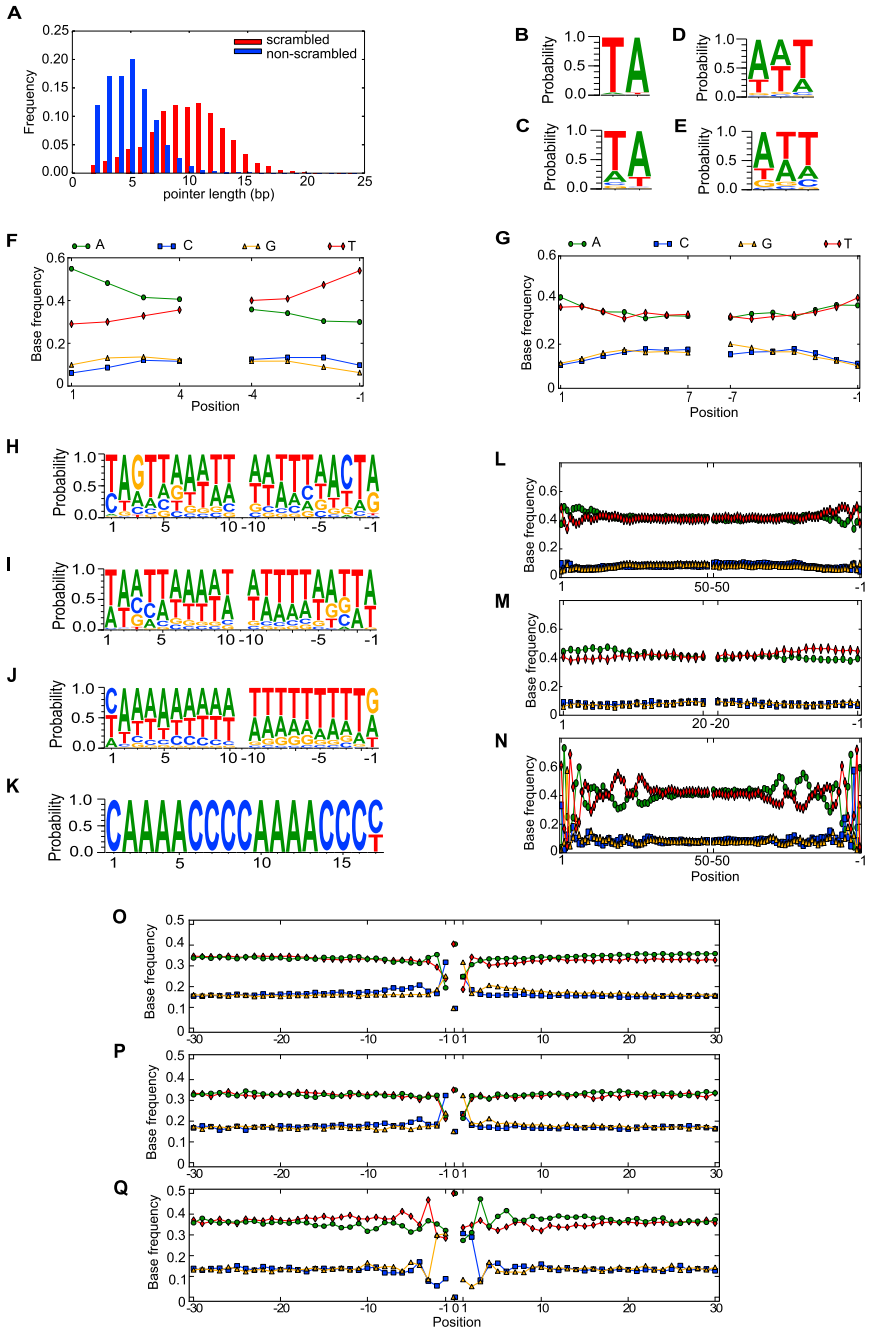

Figure S2. Pointers, MDSs, and IESs Have Distinct Sequence Properties, Related to Figure 2

(A) Length distribution of 157,648 pointers flanking nonscrambled MDSs (blue) versus 10,728 pointers flanking scrambled MDSs (red). Only pointers longer than 1 bp are shown.

(B) Sequence logo of all 2 bp pointers flanking nonscrambled MDSs (n = 17,430).

(C) 2 bp pointers flanking scrambled MDSs (n = 105).

(D) 3 bp pointers flanking nonscrambled MDSs (n = 24,741).

(E) 3 bp pointers flanking scrambled MDSs (n = 124).

(F) Base composition at the ends of all nonscrambled pointers > 3 bp (n = 115,477). The coordinate for each site is calculated from its nearest end. Positive coordinates represent distance from the 5′ end of the pointer, and negative coordinates represent distance from the 3′ end.

(G) Ends of scrambled pointers > 3 bp (n = 10,499).

(H) Ends of *Paramecium* IESs (n = 44,928) downloaded from *Paramecium*DB (Arnaiz and Sperling, 2011) (TA pointer not included). The coordinate for each site is calculated from its nearest end. Positive coordinates represent distance from the IES 5′ end (first ten bp), and negative coordinates represent distance from the 3′ end (last ten bp).

(legend continued on next page)

(I) Base composition at ends of nonscrambled *Oxytricha* IESs flanked by a TA pointer (n = 16,184) (TA pointer not included).
(J) Ends of nonscrambled *Oxytricha* IESs with an ANT pointer (n = 13,162) (ANT pointer not included).
(K) Base composition of the first 17 sites of TBE terminal inverted repeats found in the MIC assembly (n = 9,884).
(L) Base composition of first and last 50bp of nonscrambled IESs (n = 157,228), excluding the pointer.
(M) Base composition of first and last 20 bp of scrambled IESs (n = 9,327).
(N) Base composition of first and last 50 bp of *Paramecium* IESs (n = 44,928).
(O) 60 bp surrounding the location of nonscrambled pointers in the MAC (n = 156,756). All pointers are compressed to one position 0 (with the average base frequency shown at 0). Negative numbers represent 30 bp upstream of the pointer locations in the MAC and positive numbers represent 30 bp downstream of the pointers.
(P) 60 bp surrounding scrambled pointers in the MAC (n = 10,703). (Q) 60 bp surrounding pointers in the *Paramecium* MAC (n = 44,928).

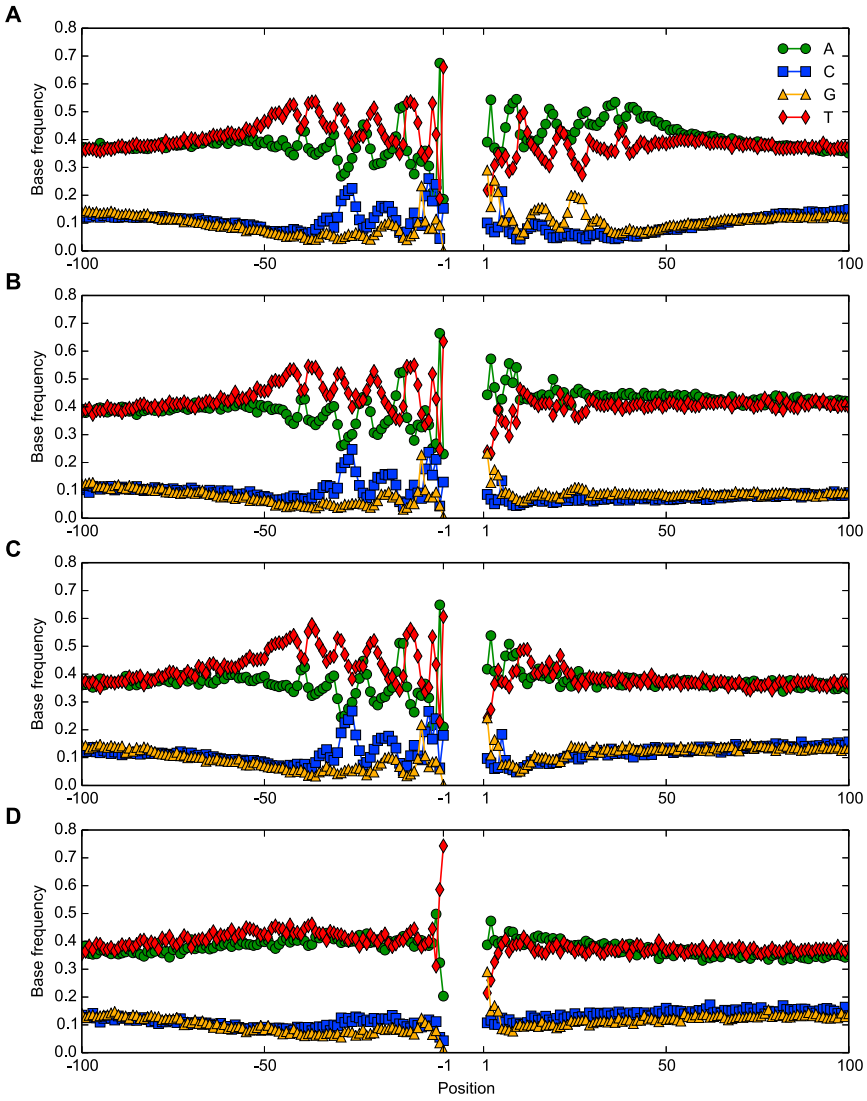

Figure S3. Chromosome Fragmentation Occurs at Regions with a Sharp Transition from Pyrimidine-Rich to Purine-Rich Sequences, Related to Figure 4

Base composition flanking telomere addition sites. The −1 position is the telomere addition site. Negative coordinates represent sites upstream of telomere addition sites; positive coordinates represent sites downstream of telomere addition sites.

(A) Base frequency of 100 bp flanking telomere addition sites < 5bp away from the neighboring terminal MDSs (n = 10,468).

(B) Base frequency of 100 bp flanking telomere addition sites > 100 bp away from the neighboring terminal MDSs (n = 3215).

(C) Base frequency of 100 bp flanking the telomere addition sites which are adjacent to an internal MDS (n = 2875).

(D) Base frequency of 100 bp flanking alternative fragmentation sites (n = 3195).

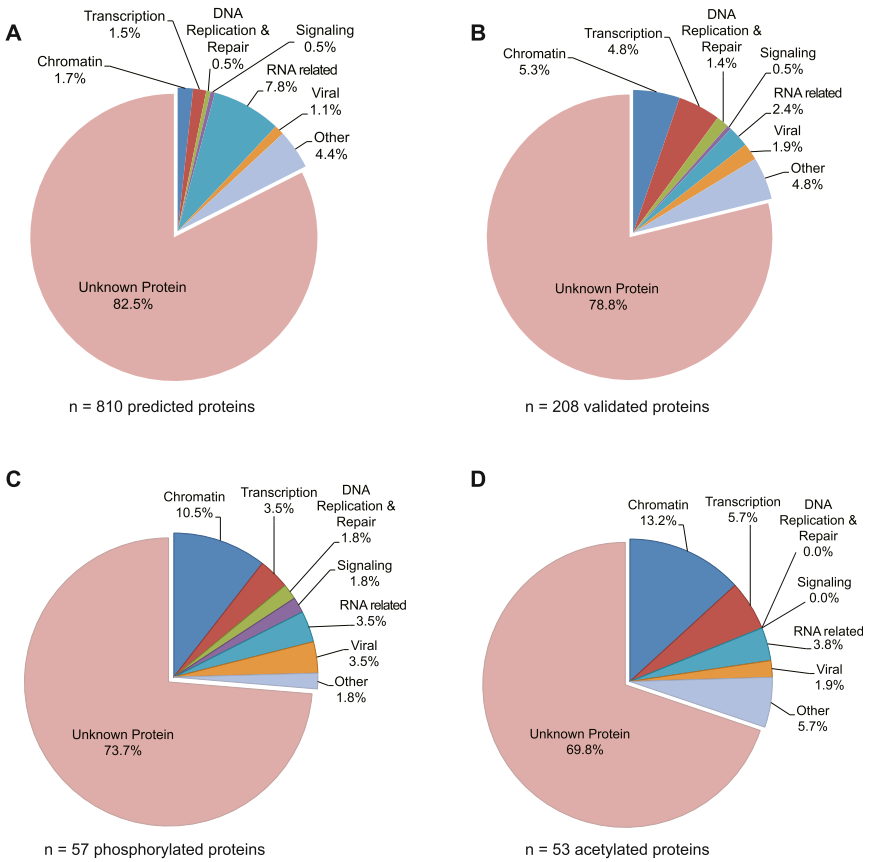

Figure S4. Grouping of MIC-Limited Proteins by Protein Domains, Related to Figure 5

(A) Predicted MIC-limited genes.

(B) 208 MIC-limited genes validated by mass spectrometry (MS) - based proteomic analysis of 40hr nuclear lysate.

(C) Phosphorylated genes identified in 40hr MS data.

(D) Acetylated proteins identified in 40hr MS data. All groupings based on Pfam annotations of predicted domains.