Report

Current Biology

The Macronuclear Genome of *Stentor coeruleus* Reveals Tiny Introns in a Giant Cell

Highlights

- The introns of *Stentor coeruleus*, a giant ciliate, are 15–16 nt long
- The short introns of *Stentor* are the shortest spliceosomal introns yet reported
- *Stentor* uses a standard genetic code, unlike other characterized ciliates
- The ploidy of the *Stentor* macronucleus is proportional to the volume of the cell

Authors

Mark M. Slabodnick, J. Graham Ruby, Sarah B. Reiff, ..., Scott W. Roy, Wallace F. Marshall, Pranidhi Sood

Correspondence

scottwroy@gmail.com (S.W.R.), wallace.marshall@ucsf.edu (W.F.M.), psood1@gmail.com (P.S.)

In Brief

Stentor coeruleus is a giant single-celled organism that can regenerate after being cut in half. Slabodnick et al. describe the *Stentor* genome, a key tool for future experiments to understand regeneration in a single cell. The genome is unusual in that it contains extremely small introns.

Slabodnick et al., 2017, Current Biology *27*, 1–7
February 20, 2017
http://dx.doi.org/10.1016/j.cub.2016.12.057

CellPress

Please cite this article in press as: Slabodnick et al., The Macronuclear Genome of *Stentor coeruleus* Reveals Tiny Introns in a Giant Cell, Current Biology (2017), http://dx.doi.org/10.1016/j.cub.2016.12.057

Current Biology
Report

The Macronuclear Genome of *Stentor coeruleus* Reveals Tiny Introns in a Giant Cell

Mark M. Slabodnick,[1,8] J. Graham Ruby,[1,8] Sarah B. Reiff,[1,8] Estienne C. Swart,[2,8] Sager Gosai,[3] Sudhakaran Prabakaran,[4] Ewa Witkowska,[5] Graham E. Larue,[6] Susan Fisher,[5] Robert M. Freeman, Jr.,[4] Jeremy Gunawardena,[4] William Chu,[7] Naomi A. Stover,[7] Brian D. Gregory,[3] Mariusz Nowacki,[2] Joseph Derisi,[1] Scott W. Roy,[6,*] Wallace F. Marshall,[1,9,*] and Pranidhi Sood[1,*]

[1]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143, USA
[2]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland
[3]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[5]Department of Ob/Gyn, University of California, San Francisco, San Francisco, CA 94143, USA
[6]Department of Biology, San Francisco State University, San Francisco, CA 94132, USA
[7]Department of Biology, Bradley University, Peoria, IL 61625, USA
[8]Co-first author
[9]Lead Contact
*Correspondence: scottwroy@gmail.com (S.W.R.), wallace.marshall@ucsf.edu (W.F.M.), psood1@gmail.com (P.S.)
http://dx.doi.org/10.1016/j.cub.2016.12.057

SUMMARY

The giant, single-celled organism *Stentor coeruleus* has a long history as a model system for studying pattern formation and regeneration in single cells. *Stentor* [1, 2] is a heterotrichous ciliate distantly related to familiar ciliate models, such as *Tetrahymena* or *Paramecium*. The primary distinguishing feature of *Stentor* is its incredible size: a single cell is 1 mm long. Early developmental biologists, including T.H. Morgan [3], were attracted to the system because of its regenerative abilities—if large portions of a cell are surgically removed, the remnant reorganizes into a normal-looking but smaller cell with correct proportionality [2, 3]. These biologists were also drawn to *Stentor* because it exhibits a rich repertoire of behaviors, including light avoidance, mechanosensitive contraction, food selection, and even the ability to habituate to touch, a simple form of learning usually seen in higher organisms [4]. While early microsurgical approaches demonstrated a startling array of regenerative and morphogenetic processes in this single-celled organism, *Stentor* was never developed as a molecular model system. We report the sequencing of the *Stentor coeruleus* macronuclear genome and reveal key features of the genome. First, we find that *Stentor* uses the standard genetic code, suggesting that ciliate-specific genetic codes arose after *Stentor* branched from other ciliates. We also discover that ploidy correlates with *Stentor*'s cell size. Finally, in the *Stentor* genome, we discover the smallest spliceosomal introns reported for any species. The sequenced genome opens the door to molecular analysis of single-cell regeneration in *Stentor*.

RESULTS AND DISCUSSION

Shotgun Sequencing of *Stentor* Macronuclear Genome

As in all ciliates, the *Stentor* (Figures 1A and 1B) genome is organized into micronuclei and a macronucleus. Typically in ciliates, the micronucleus contains the diploid genome, is transcriptionally inert, and only functions during inheritance. The macronucleus contains a highly amplified genome derived from the micronuclear sequence and contains all genes functional during vegetative growth. We sequenced the macronuclear genome using the Nextera system for genomic library preparation and Illumina sequencing. The current assembly is based on 109.3 million paired-end reads, from which we generated a draft assembly of the *Stentor* genome using a combination of the SOAPdenovo [5] and PRICE [6] assemblers (see the Supplemental Experimental Procedures). The assembly was performed in close concert with experiments, and contigs were spot-checked using PCR to identify systematic mis-assembly problems. Our assembly included 9,198 contigs with a total length of 83 Mb and a contig N50 of 51 kb (Figure 1C). Of these contigs, 29 have telomeres on both ends and 465 have one on only one side (see the Supplemental Experimental Procedures for more details). Using three different approaches, we estimate that the SNP density ranges from one to four SNPs per 1,500 bases (see the Supplemental Experimental Procedures), suggesting that the genome exhibits low heterozygosity. The genome assembly and all associated raw data have been deposited in GenBank (BioProject PRJNA352242 and BioSample SAMN05968724). Additionally, the genome is available online at http://stentor.ciliate.org. As shown in Figure 1D, coverage is 50–100× in most regions. The mitochondrial genome is part of the assembly (contig 652).

The contig size distribution is consistent with prior biochemical analysis of isolated *Stentor* genomic DNA, in which it was estimated that 50% of the genome consisted of chromosomes in the 46–62 kb range [7]. We further investigated the distribution of chromosome sizes using a clamped homogeneous electric

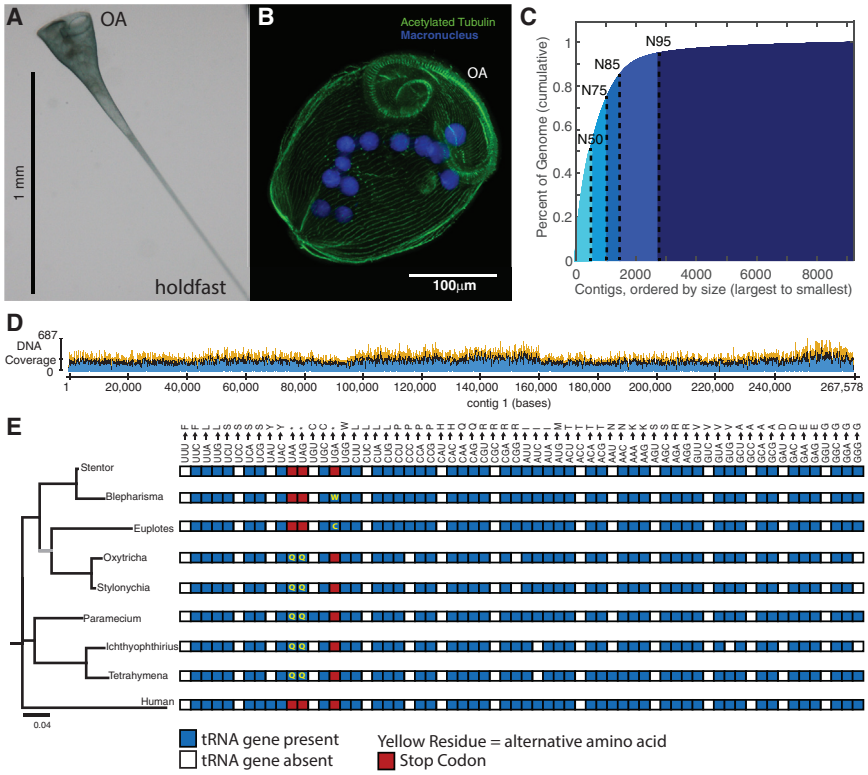


Figure 1. Shotgun Sequencing the *Stentor coeruleus* Macronuclear Genome

(A) Bright-field image of a live *Stentor* cell in its extended, feeding form. The oral apparatus is at the top of the image and the hold fast is at the bottom, as indicated.

(B) Fluorescence micrograph of a fixed and stained *Stentor* cell in its contracted form (cells contract upon fixation). The macronucleus is stained by DAPI. Cilia and the longitudinal bundles of microtubules, which run in parallel along the whole length of the cell, are marked by an antibody against acetylated tubulin. The cilia, which comprise the oral apparatus (OA), are indicated.

(C) Cumulative distribution depicting the N50 (50 kb) of the assembled *Stentor* genome. The largest percentage of the genome is accounted for by the longest contigs.

(D) Sequencing coverage for the first contig in the assembly.

(E) Left: Phylogenetic comparison of 18S RNA for ciliates using *Homo sapiens* as an outgroup. The tree was built using an HKY substitution model based on a ClustalW multiple sequence alignment. All bootstrap values are >90 with the exception of that marked in gray, which has a bootstrap value of 53. Right: a comparison of the genetic codes for ciliates and human. A blue box indicates the presence of a tRNA gene while white indicates its absence. Red boxes indicate codons used as termination signals, while yellow residues indicate alternative amino acid encodings. *Blepharisma* and *Stentor* both belong to the ciliate class Heterotrichea; *Euplotes*, *Oxytricha*, and *Stylonychia* represent class Spirotrichea; and *Paramecium*, *Tetrahymena*, and *Ichthyophthirius* represent class Oligohymenophorea.
See also Figure S1 and Table S1.

field (CHEF) gel (Figure S1D), and we found that the range of sizes is 20–250 kb, comparable to the range of sizes of contigs in the assembly (25–265 kb). It is important to note that contigs could lie outside the range we identified experimentally. The size of the *Stentor* genome has previously been estimated at ~92 Mb [7], similar to our estimate of 83 Mb. Considering the high level of alignment of cDNA reads with our genome (see below), we suspect that our assembly is mostly complete and that the biochemical estimates may have overestimated genome size slightly. The genome size is comparable to that of other ciliates (Table S1). The GC content of our assembly is 30%, comparable to the prior biochemical estimate of 32% [7].

Stentor Uses a Standard Genetic Code, Unlike Most Other Ciliates

Ciliates whose genomes have been sequenced to date all employ non-canonical genetic codes. For example, in *Tetrahymena* and *Paramecium*, the UAA and UAG stop codons encode glutamine. We searched for tRNA genes in the *Stentor* genome using tRNAScanSE [8] and found a full complement of genes encoding all necessary amino acids, but no glutamine tRNA genes with a UUA or CUA anticodon, nor any tryptophan tRNA gene with a UCA anticodon (as in the *Blepharisma* code). We performed mass spectrometry of peptides from total *Stentor* protein and mapped spectra to six-frame translations of the *Stentor* genome translated with four different genetic codes (the standard code, the so-called "ciliate code" used by most characterized ciliates, the *Blepharisma* code, and a less frequently observed ciliate code where UAA and UAG encode glutamate).

We found that ~135,165 open reading frames (ORFs) translated with the standard code had peptide support, compared to ~139,929 ORFs translated with the primary ciliate code, ~136,488 ORFs translated with the *Blepharisma* code, and ~139,076 with the UAR-glutamate ciliate code. Of the ORFs translated with the ciliate code, only 0.04% had peptide support for alternative codons. Of the ORFs translated with the *Blepharisma* code, only 0.02% had support for alternative codons. Of the ORFs translated with the UAR-glutamate ciliate code, only 0.07% had peptide support for alternative codons. In the majority of these cases (84%, 72%, and 93% of alternative codon-containing ORFs translated with the ciliate, *Blepharisma*, and UAR-glutamate tables, respectively), the conserved core of the protein was also identified using the standard table, and a search of the BLAST nr database produced hits that were of equivalent or less significant e value as the corresponding ORFs translated with the standard table, suggesting that translational read-through occurred at either UAA or UAG codons, resulting in a peptide extension. The remaining ORFs with peptide support for alternative codon usage lacked homology to annotated genes (see the Supplemental Experimental Procedures for more details about these cases).

Therefore, as shown in Figure 1E, we conclude that *Stentor* primarily uses the standard genetic code and does not exhibit the hallmark genetic code alterations seen in other ciliates,

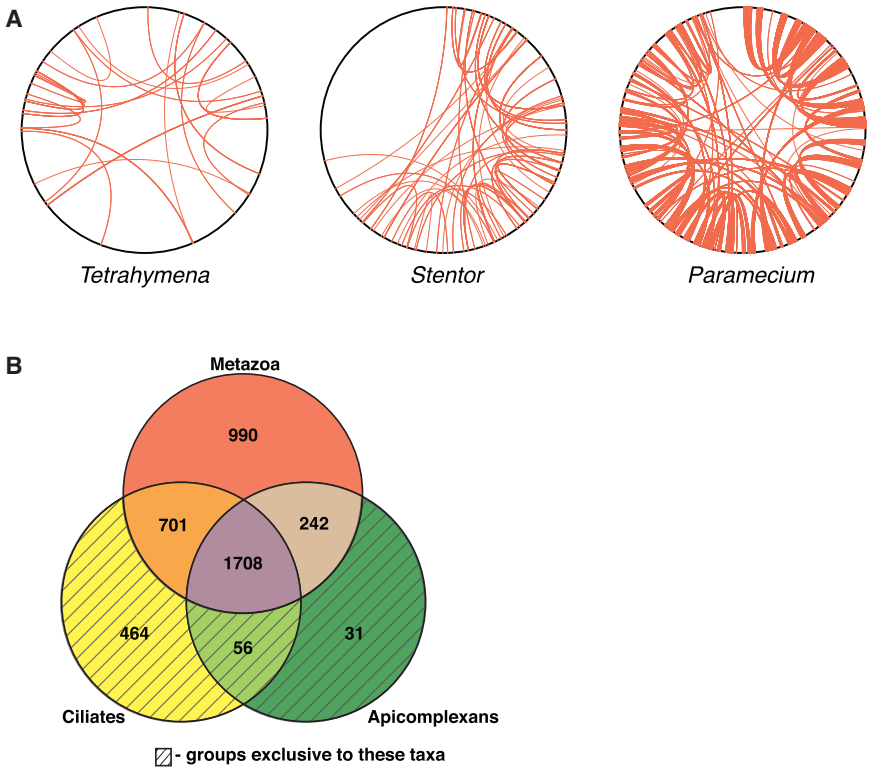


Figure 2. *Stentor* Gene Duplications and Orthology Groups

(A) Genome duplication events in the genomes of *Stentor coeruleus*, *Paramecium tetraurelia*, and *Tetrahymena thermophila.* For generation of coordinates on the perimeter of each circle, contigs/scaffolds were arranged from longest to shortest and then continuously numbered from one to the end of the assemblies. Red lines connect paralogous windows (see the Supplemental Experimental Procedures) between two scaffolds and indicate putative genome duplication events.

(B) Venn diagram showing numbers of orthologous gene groups in *Stentor* that are also found in other ciliates, apicomplexans, or metazoans. Shaded regions indicate gene groups that are exclusive to those taxa; for example, the ciliate-only region of the diagram represents gene groups that aren't found in any other taxa. An additional 555 curated groups are shared with other organisms but are not pictured in this diagram.

See also Figure S2 and Mendeley Data, http://dx.doi.org/10.17632/37gp2djcst.1.

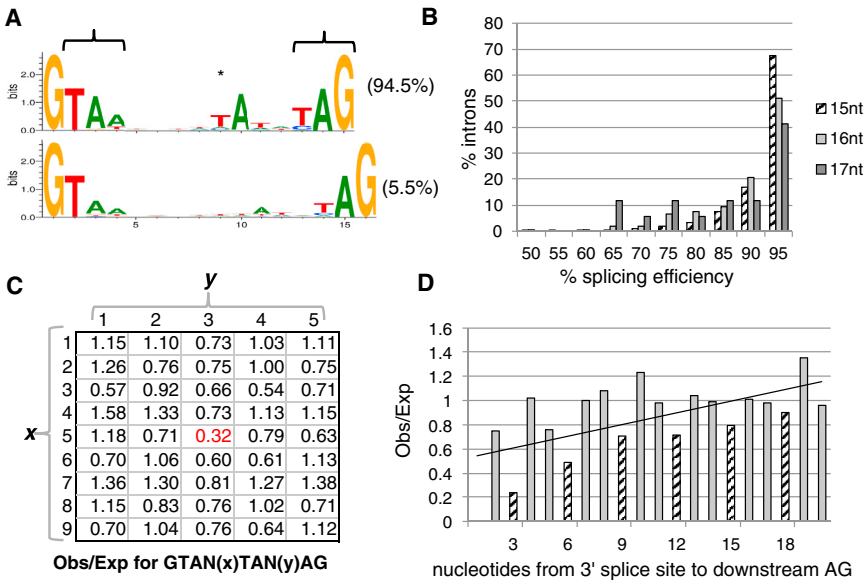
suggesting that *Stentor* branched from the ciliate common ancestor before genetic codes started to deviate. Figure S1E provides a sequence alignment of eukaryotic release factor (eRF1) for *Stentor*, although, with the availability of new sequence evidence [9], previous explanations linking mutations in the eRF1 to alterations in the genetic code [10] no longer appear to hold.

Gene Identification and Estimation of Gene Number

To estimate the completeness of the assembled genome, we used the Core Eukaryotic Genes Mapping Approach (CEGMA) to search for orthologs of highly conserved core eukaryotic genes [11], using search parameters previously employed for ciliate genomes [12]. Out of 248 genes in the standard core set, fully completed ciliate genomes, such as *Tetrahymena* or *Oxytricha*, typically contain 220–230. Our assembly contained orthologs of 243 of the 248 core eukaryotic genes, suggesting that the assembly is largely complete. The identification of a full complement of tRNA encoding genes further bolsters our assessment of completeness.

To identify *Stentor* genes, we combined de novo gene prediction with RNA sequencing. We sequenced 125 million cDNA reads, of which 97.25% mapped onto the genomic assembly using Bowtie2 [13], confirming a high level of completeness in the assembly. Using a set of 307 manually verified gene models combined with RNA sequencing (RNA-seq) data to train the Augustus gene prediction program [14], 34,506 gene models were generated. Of these models, 99% are supported by RNA-seq reads and 33% have proteomic support. This gene number is comparable to that seen in other ciliates; for example, the *Paramecium* genome encodes ~40,000 genes [15] and *Tetrahymena* encodes ~27,000 genes [16]. In *Paramecium*, the large number of genes is hypothesized to be the result of multiple whole-genome duplication events [15], whereas other mechanisms appear to drive the large number of genes found in the *Tetrahymena* genome [16, 17]. Although there is some evidence for duplication of a small number of genomic regions in *Stentor* (Figure 2A), the large number of genes in the genome cannot be explained by so few events (only 99 genes comprise the potential genome duplication events). Additionally, analysis of the percentage identity between reciprocal best BLAST hits, as well as their non-synonymous to synonymous rates of substitution (Figures S2D and S2E), indicate that, although genome duplication events might have shaped the *Stentor* genome to some extent, they played a greater role in shaping the *Paramecium* genome.

We matched our gene predictions to groups of orthologous genes in the OrthoMCL database, as well as to proteomes of other ciliates. Of *Stentor*'s 34,506 gene models, 21,602 were grouped into 7,676 ortholog groups shared with other species, including both curated ortholog groups in OrthoMCL and ciliate-specific ortholog groups (see Mendeley Data, http://dx.doi.org/10.17632/37gp2djcst.1). Of the 4,747 curated ortholog groups found in *Stentor*, all but three are shared with eukaryotes (Figure S2A); 464 *Stentor* gene groups were ciliate specific and 56 groups were alveolate specific (Figure 2B). Among this latter group were 31 gene groups previously thought to be specific to Apicomplexa, a sister phylum of ciliates. These ancestral alveolate genes may have been lost in other ciliate branches. A comparison of *Stentor* orthology groups to three other ciliates (Figure S2B) revealed 998 *Stentor* orthology groups shared with other organisms, but not present in the other ciliates. These groups may represent gene families lost in other ciliate classes since the branching of the Heterotrichidae. Half of the top ten orthology groups with the most *Stentor* genes contained kinases, and a sixth group was comprised mostly of protein phosphatase 2C orthologs. Using HMMER3 (http://hmmer.org) to find kinase domains in the *Stentor* gene models, we found that the



x \ y	1	2	3	4	5
1	1.15	1.10	0.73	1.03	1.11
2	1.26	0.76	0.75	1.00	0.75
3	0.57	0.92	0.66	0.54	0.71
4	1.58	1.33	0.73	1.13	1.15
5	1.18	0.71	0.32	0.79	0.63
6	0.70	1.06	0.60	0.61	1.13
7	1.36	1.30	0.81	1.27	1.38
8	1.15	0.83	0.76	1.02	0.71
9	0.70	1.04	0.76	0.64	1.12

Figure 3. Intron Sequences and Splicing in *Stentor*

(A) Nearly all identified introns in *Stentor* are 15 nt (94.5%, top) or 16 nt (5.5%, bottom), displaying an abbreviated 5′ splice site motif, atypical internal TA dinucleotide (asterisk), and potential stop codons (brackets). Weblogos were generated and normalized to neutral base frequencies in intergenic regions.

(B) Greater splicing efficiency of 15-nt introns. Graph shows a histogram of the distribution of introns in each size class (15–17 nt) showing a given level of splicing efficiency, defined as the number of spliced RNA-seq reads divided by the total number of spliced and unspliced reads for each intronic locus.

(C) Avoidance of intron-like motifs in protein-coding regions. The occurrence within protein-coding regions of intron-like motifs is shown, revealing stronger underrepresentation of intron-like GTAN(5)TAN(3)AG motifs (red) compared to similar motifs (other combinations of GTAN(1–9)TAN(1–5)AG). x indicates the number of the bases (N = ATCG) preceding the T before the branch point, and y indicates the number of bases following the branch point A (thus, the intronic motif is x = 5, y = 3).

(D) Avoidance of alternative 3′ splice sites. Downstream AG dinucleotides near the 3′ AG splice site are less common than expected, particularly for distances that do not induce a frameshift (multiples of three nucleotides, striped bars). The trend line is a linear fit to all data shown.

See also Figure S3 and Tables S2 and S4.

Stentor genome encodes a vast complement of kinases totaling over 2,000 kinase genes.

Stentor Introns Are Unusually Small

The most striking feature of the *Stentor* genome is its extremely short introns; 9,325 introns were predicted in gene models, and, of those that we confirmed by Sanger sequencing, 94.5% were 15 nt long, the rest were 16 nt, and all were of a canonical type (Figure 3A). These introns are shorter than those of the previous record holder, the *Bigelowiella natans* nucleomorph (with a mode of 19 nt), which possesses a reduced genome (284 genes) [18]. We also found that 15/16-nt introns are characteristic of other heterotrichous ciliates, as well as a ciliate from a sister class (Karyorelictea), suggesting that tiny introns have a long history in these ciliates (Figure S3A).

Whereas previously reported short introns lacked clear internal candidates for branch point sites [18–20], *Stentor* introns exhibit a strongly conserved A nucleotide, most likely representing the branch point, near the 3′ end (6 nt upstream for 15-nt introns, 6–7 nt for 16-nt introns; asterisk in Figure 3A), suggesting that these short introns could be spliced by a canonical two-step splicing reaction. There is evidence for splicing reactions for short introns with similarly spaced branch points and 3′ ends in other species [21–25]. Interestingly, for the vast majority of introns, this A was preceded by a non-canonical T nucleotide, which is not complementary to the standard U2 small nuclear RNA (snRNA). The *Stentor* U2 snRNA genes maintain the standard sequence found in other species and lack a complementary nucleotide. To our knowledge, this represents the first reported case in which a putative branch point motif, otherwise conserved, does not show the standard complementarity to the U2. The vast majority of 15-nt introns (84.8%) contained an in-frame stop codon (versus only 29.5% of 16-nt introns). These stop codons largely reflect the fact that the consensus 15-nt sequence contains stop codons in two of three possible reading frames (brackets in Figure 3A); the 16-nt consensus sequence has both stops in the same frame and thus only has stops in one of three possible reading frames. It is thus unclear whether the presence of in-frame stops reflects a selection on stop codons or is simply a by-product of the consensus sequence. These novel intron features do not seem to be associated with widespread intron creation, as the majority (71.4%) of introns in conserved regions are found at intron positions shared with one or more distantly related ciliates, suggesting that these atypical introns by and large evolved from more typical ones.

The near homogeneity of short intron lengths in this organism raises questions about the splicing mechanism and efficiency. RNA-seq data analysis indicated that introns were efficiently spliced (95.0% of reads spliced), but that 16-nt introns were somewhat less so (92.4%; $p = 4 \times 10^{-6}$ by randomization; Figure 3B). Several features suggest avoidance of off-target splicing may shape the transcriptome. First, within unspliced regions confirmed by RNA-seq, intron-like sequences (i.e., $GTAN_5TAN_3AG$; Figure 3C) were avoided, suggesting selection against off-target cryptic splicing. Second, AG nucleotides were less frequent downstream of confirmed 3′ splice sites, and those that were observed were more likely to produce a frameshift, suggesting a role of nonsense-mediated decay (NMD)—a process thought to be conserved in *Stentor* as it has orthologs of UPF1 and UPF2—in mitigating the deleterious effects of splicing mistakes (Figure 3D). Indeed, a substantial fraction of observed 17- to 18-nt splicing events may represent splicing mistakes, since the 3′ AG lay directly downstream of an AG at the 15-nt or 16-nt position in 40.0% of cases, 78.8% of which are confirmed splice boundaries (although such cases may also represent functional alternative splicing).

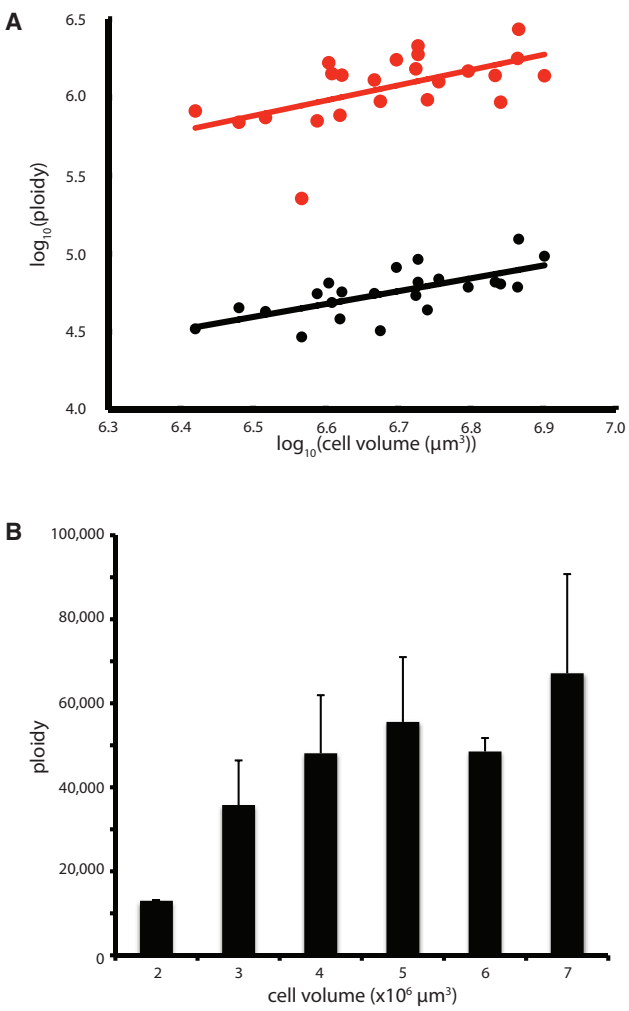


Figure 4. Macronuclear Ploidy Scales with Cell Volume

(A) Scaling of two contigs with cell volume. Graph depicts the $\log_{10}$ of contig copy number versus the $\log_{10}$ of cell volume, based on droplet digital PCR of individual cells. The copy number of rDNA-containing contig (red) and a large contig that does not contain rDNA (black) are shown. Each point represents a single cell. Ploidy data used two different y axis scales because the average ploidy is ~20 times greater for the contig containing the rDNA locus. Lines represent best-fit power law relation.

(B) Average ploidy for five contigs spanning a size range of 42,000–230,000 bp, not including the rDNA contig. Error bars indicate SD.

See also Table S3.

The introns of *Stentor* and the other heterotrichs we analyzed are the shortest spliceosomal introns ever reported. By contrast, average intron sizes in *Tetrahymena* and *Paramecium* are 165 nt [26]) and 25 nt [19, 27]), respectively. We do not know why heterotrich genomes have such short introns, but it suggests that there may be evolutionary pressure to minimize the length of transcripts in the macronuclear genome or to reduce regulation through splicing. This idea is supported by the fact that, in *Stentor*, the majority of genes are single-exon genes (82%), whereas in other ciliates this proportion is smaller (*Tetrahymena*, 32%; *Ichthyophthirius*, 22%; *Oxytricha*, 36%; and *Paramecium*, 20%).

Stentor's 3′ UTRs are also small with a median length of 31 nt, similar in length to other heterotrichs (median 24–26 nt [9]). Further details of 3′ UTR size distribution, poly(A) tail position, and UTR-specific regulatory elements are given in the Supplemental Experimental Procedures.

As expected from the short introns and UTR sequences, the proportion of coding sequence per gene is higher in *Stentor* than in other ciliates (Figure S3B). Intergenic lengths in the *Stentor* genome are similar in length to *Paramecium*'s (Table S2). Intergenic lengths in *Oxytricha* are shorter than in *Stentor*, because the *Oxytricha* genome is composed of nanochromosomes, most of which contain only one gene. The compactness of introns and UTRs, but not intergenic regions, raises the question of whether the *Stentor* genome has been under pressure to have short transcripts for protein-coding genes.

Stentor Genome Copy Number Is Proportional to Cell Size

One of the most striking features of *Stentor* is the huge size of its cells. Cell size frequently correlates with genome size [28–31]. Even within a single species, increased cell size is often accompanied by increased DNA content via polyploidization [32–35]. In some cases, polyploidization may be sufficient to drive the expansion of cell volume [36].

Given that the *Stentor* macronuclear genome is comparable in size with other, smaller, ciliates, we hypothesized that the large size of *Stentor coeruleus* might be accompanied by a higher ploidy. Droplet digital PCR of seven different contigs in cells of varying sizes confirmed that *Stentor* is polyploid. For example, the rDNA locus-containing contig (contig 2,227) is present at an average of 1.1 million copies per cell. Six other contigs examined had an average copy number of 60,000, indicating that the rDNA-containing contig is present at ~20 times higher copy number than other contigs. Similar enrichment of rDNA-containing DNA occurs in other ciliates [12]. In *Tetrahymena,* the rDNA copy number is at least 200 times more than that of other contigs [37]. A log-log scaling plot (Figure 4A) shows that copy number scales with cell volume with a best-fit slope of 0.91 (for contig 2) and 0.98 (for contig 2,227), indicating that ploidy is proportional to cell volume.

Figure 4B plots average ploidy for six non-rDNA contigs as a function of cell size, indicating a trend toward increased copy number in larger cells.

Scaling of ploidy with cell size agrees with observations that macronuclear DNA synthesis occurs throughout interphase in *Stentor* [38] and suggests DNA content may determine cell size in *Stentor* or vice versa.

ACCESSION NUMBERS

The accession number for the genome assembly and all associated raw data reported in this paper are GenBank: PRJNA352242 and SAMN05968724. The accession number for the orthologous gene groups in *Stentor* reported in this paper is Mendeley Data: http://dx.doi.org/10.17632/37gp2djcst.1.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2016.12.057.

AUTHOR CONTRIBUTIONS

M.M.S., S.B.R., S.G., B.D.G., S.P., and P.S. designed and performed experiments. P.S., M.M.S., J.G.R., S.B.R., E.C.S., M.N., S.P., R.M.F., J.G., G.E.L., E.W., J.D., S.F., W.C., N.A.S., W.F.M., and S.W.R. analyzed data. P.S., M.M.S., S.B.R., E.C.S., S.W.R., and W.F.M. wrote the paper.

ACKNOWLEDGMENTS

We thank Joel Rosenbaum, Denis Diener, Hiten Madhani, Bruce Alberts, and Michael Lynch for helpful discussions. This work was supported by an ARCS Graduate Fellowship (M.M.S.), an American Cancer Society postdoctoral fellowship (P.S.), the Herbert Boyer Junior Faculty Endowed Chair (W.F.M.), a UCSF Resource Allocation Program New Directions grant (W.F.M.), and by NIH grants R01 GM090305 (W.F.M.) and R01 GM113602 (W.F.M.).

Received: June 30, 2016
Revised: November 17, 2016
Accepted: December 28, 2016
Published: February 9, 2017

REFERENCES

1. Slabodnick, M.M., and Marshall, W.F. (2014). Stentor coeruleus. Curr. Biol. *24*, R783–R784.
2. Tartar, V. (1961). The Biology of *Stentor* (Pergammon Press).
3. Morgan, T.H. (1901). Regeneration of proportionate structures in *Stentor*. Biol. Bull. *2*, 311–328.
4. Wood, D.C. (1969). Parametric studies of the response decrement produced by mechanical stimuli in the protozoan, *Stentor coeruleus*. J. Neurobiol. *1*, 345–360.
5. 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.
6. Ruby, J.G., Bellare, P., and Derisi, J.L. (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda) *3*, 865–880.
7. Pelvat, B., and de Haller, G. (1976). Macronuclear DNA in *Stentor coeruleus*: a first approach to its characterization. Genet. Res. *27*, 277–289.
8. 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.
9. Swart, E.C., Serra, V., Petroni, G., and Nowacki, M. (2016). Genetic codes with no dedicated stop codon: context-dependent translation termination. Cell *166*, 691–702.
10. Lozupone, C.A., Knight, R.D., and Landweber, L.F. (2001). The molecular basis of nuclear genetic code change in ciliates. Curr. Biol. *11*, 65–74.
11. Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067.
12. 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.
13. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
14. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637–644.
15. 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.
16. Hamilton, E.P., Kapusta, A., Huvos, P.E., Bidwell, S.L., Zafar, N., Tang, H., Hadjithomas, M., Krishnakumar, V., Badger, J.H., Caler, E.V., et al. (2016). Structure of the germline genome of Tetrahymena thermophila and relationship to the massively rearranged somatic genome. eLife *5*, e19090.
17. Coyne, R.S., Hannick, L., Shanmugam, D., Hostetler, J.B., Brami, D., Joardar, V.S., Johnson, J., Radune, D., Singh, I., Badger, J.H., et al. (2011). Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. Genome Biol. *12*, R100.
18. Gilson, P.R., Su, V., Slamovits, C.H., Reith, M.E., Keeling, P.J., and McFadden, G.I. (2006). Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc. Natl. Acad. Sci. USA *103*, 9566–9571.
19. Russell, C.B., Fraga, D., and Hinrichsen, R.D. (1994). Extremely short 20-33 nucleotide introns are the standard length in *Paramecium tetraurelia*. Nucleic Acids Res. *22*, 1221–1225.
20. Ogino, K., Tsuneki, K., and Furuya, H. (2010). Unique genome of dicyemid mesozoan: highly shortened spliceosomal introns in conservative exon/intron structure. Gene *449*, 70–76.
21. Irimia, M., and Roy, S.W. (2008). Evolutionary convergence on highly-conserved 3′ intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. PLoS Genet. *4*, e1000148.
22. Lee, R.C.H., Gill, E.E., Roy, S.W., and Fast, N.M. (2010). Constrained intron structures in a microsporidian. Mol. Biol. Evol. *27*, 1979–1982.
23. Vanácová, S., Yan, W., Carlton, J.M., and Johnson, P.J. (2005). Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. Proc. Natl. Acad. Sci. USA *102*, 4430–4435.
24. Nixon, J.E.J., Wang, A., Morrison, H.G., McArthur, A.G., Sogin, M.L., Loftus, B.J., and Samuelson, J. (2002). A spliceosomal intron in *Giardia lamblia*. Proc. Natl. Acad. Sci. USA *99*, 3701–3705.
25. Xu, F., Jerlström-Hultqvist, J., Einarsson, E., Astvaldsson, A., Svärd, S.G., and Andersson, J.O. (2014). The genome of *Spironucleus salmonicida* highlights a fish pathogen adapted to fluctuating environments. PLoS Genet. *10*, e1004053.
26. 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.
27. Jaillon, O., Bouhouche, K., Gout, J.-F., Aury, J.-M., Noel, B., Saudemont, B., Nowacki, M., Serrano, V., Porcel, B.M., Ségurens, B., et al. (2008). Translational control of intron splicing in eukaryotes. Nature *451*, 359–362.
28. Price, H.J., Sparrow, A.H., and Nauman, A.F. (1973). Correlations between nuclear volume, cell volume and DNA content in meristematic cells of herbaceous angiosperms. Experientia *29*, 1028–1029.
29. Olmo, E. (1983). Nucleotype and cell size in vertebrates: a review. Basic Appl. Histochem. *27*, 227–256.
30. Shuter, B.J., Thomas, J.E., Taylor, W.D., and Zimmerman, A.M. (1983). Phenotypic correlates of genomic DNA content in unicellular eukaryotes and other cells. Am. Nat. *122*, 26–44.
31. Mueller, R.L. (2015). Genome biology and the evolution of cell-size diversity. Cold Spring Harb. Perspect. Biol. *7*, a019125.
32. Winkelmann, M., Pfitzer, P., and Schneider, W. (1987). Significance of polyploidy in megakaryocytes and other cells in health and tumor disease. Klin. Wochenschr. *65*, 1115–1131.
33. Biesterfeld, S., Gerres, K., Fischer-Wein, G., and Böcking, A. (1994). Polyploidy in non-neoplastic tissues. J. Clin. Pathol. *47*, 38–42.
34. Anatskaya, O.V., and Vinogradov, A.E. (2010). Somatic polyploidy promotes cell function under stress and energy depletion: evidence from tissue-specific mammal transcriptome. Funct. Integr. Genomics *10*, 433–446.

35. Gillooly, J.F., Hein, A., and Damiani, R. (2015). Nuclear DNA content varies with cell size across human cell types. Cold Spring Harb. Perspect. Biol. *7*, a019091.

36. Losick, V.P., Fox, D.T., and Spradling, A.C. (2013). Polyploidization and cell fusion contribute to wound healing in the adult *Drosophila* epithelium. Curr. Biol. *23*, 2224–2232.

37. Yao, M.C., Kimmel, A.R., and Gorovsky, M.A. (1974). A small number of cistrons for ribosomal RNA in the germinal nucleus of a eukaryote, *Tetrahymena pyriformis*. Proc. Natl. Acad. Sci. USA *71*, 3082–3086.

38. De Terra, N. (1967). Macronuclear DNA synthesis in *Stentor*: regulation by a cytoplasmic initiator. Proc. Natl. Acad. Sci. USA *57*, 607–614.

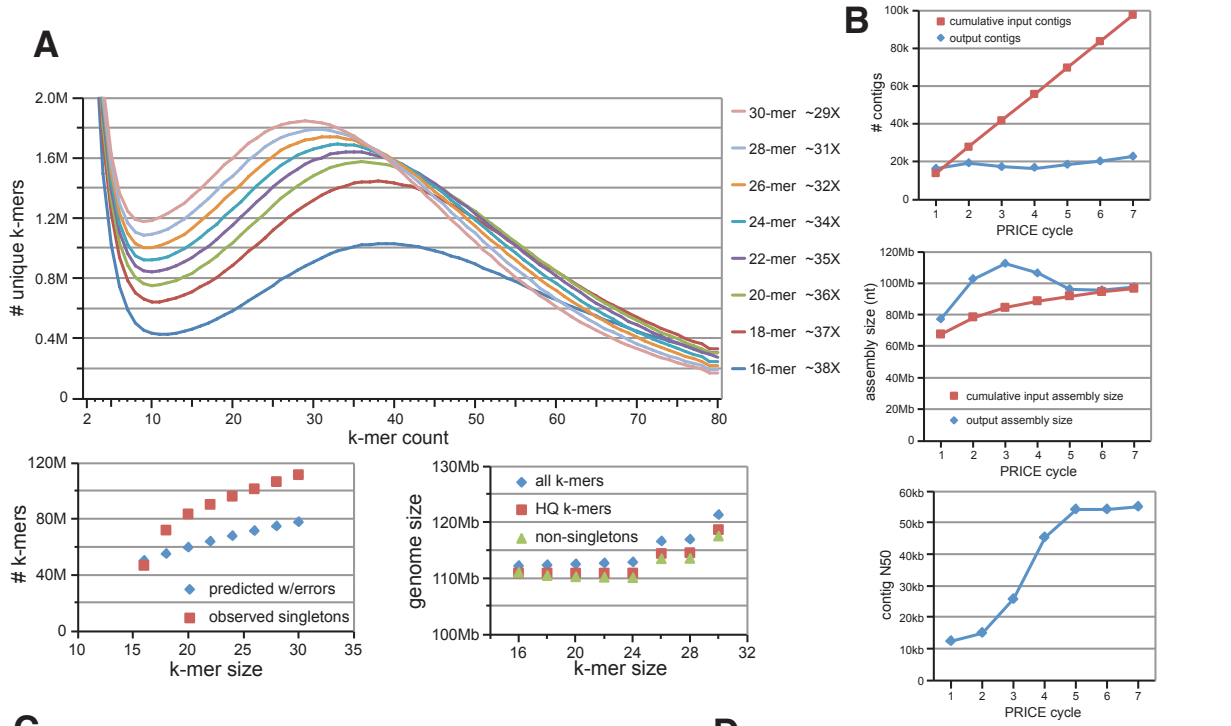
Current Biology, Volume *27*

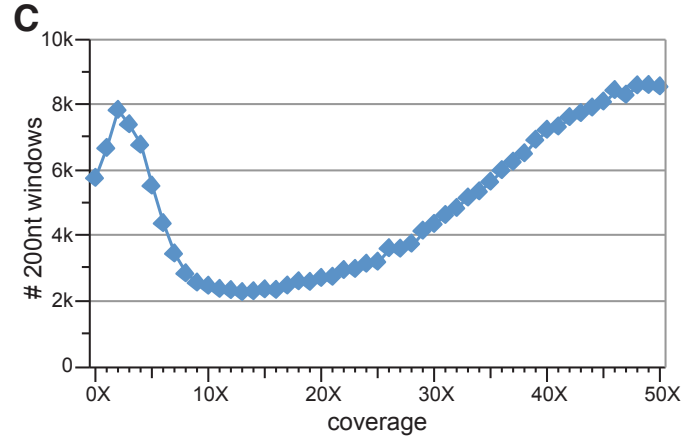
Supplemental Information

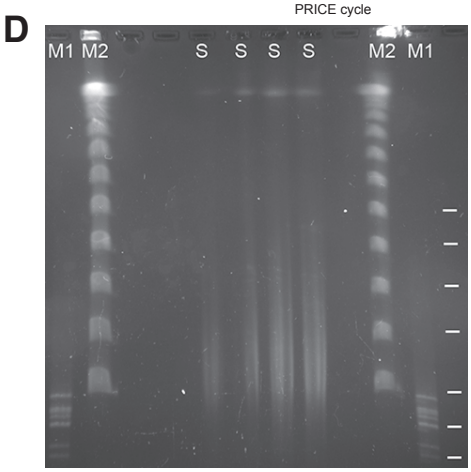
The Macronuclear Genome of *Stentor coeruleus* Reveals Tiny Introns in a Giant Cell

Mark M. Slabodnick, J. Graham Ruby, Sarah B. Reiff, Estienne C. Swart, Sager Gosai, Sudhakaran Prabakaran, Ewa Witkowska, Graham E. Larue, Susan Fisher, Robert M. Freeman, Jr., Jeremy Gunawardena, William Chu, Naomi A. Stover, Brian D. Gregory, Mariusz Nowacki, Joseph Derisi, Scott W. Roy, Wallace F. Marshall, and Pranidhi Sood

Supplemental Figures







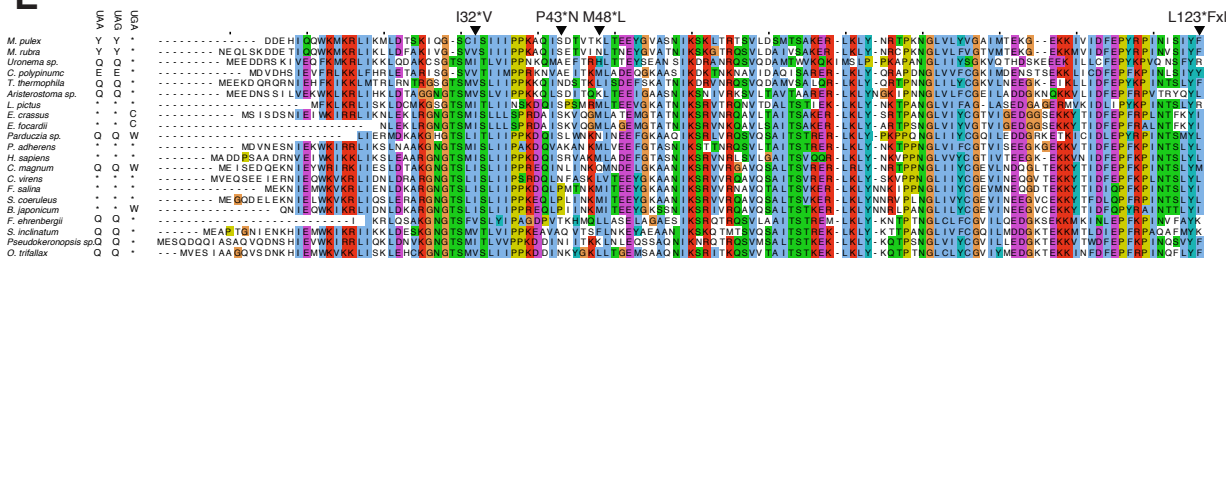


Figure S1. Features of the de novo assembly of the *Stentor* genome. (**A**) Estimating sequencing coverage based on k-mer counting. The distribution of k-mer frequencies was used to determine a typical level of sequencing coverage based on the modal value of the right peak of the distribution. (**B**) Progress of the initial PRICE extension of the SOAPdenovo assembly through 7 cycles of contig extension. Contig number, total assembly size, and contig N50 are plotted as a function of cycle number, along with the total number/size of SOAP contigs that had cumulatively been added to the assembly by the beginning of each cycle. (**C**) Coverage distribution of non-overlapping 200nt windows across an intermediate genome assembly; based on this plot, 10X was used as a cut-off value for eliminating low-coverage sequences that were likely to derive from mis-assembly or from contaminants (see Methods). (**D**) CHEF gel of Stentor DNA samples. M1, lambda DNA monocut ladder; M2, lambda PFG ladder; S, Stentor DNA samples in agarose plugs. White tick marks, from top to bottom - 242.5 kb, 194 kb, 145.5 kb, 97 kb, 48.5 kb, 24 kb, 15 kb. (**E**) N-terminal alignment of eRF1 based upon Supplemental Figure S7 from [S1]. Indicated are amino acid changes previously proposed to result in alteration of the recognition of stop codons [S2], which no longer appear to consistently explain alterations observed in ciliates as additional sequences have been added to this alignment. This figure supplements the data presented in Figure 1 of the main text.

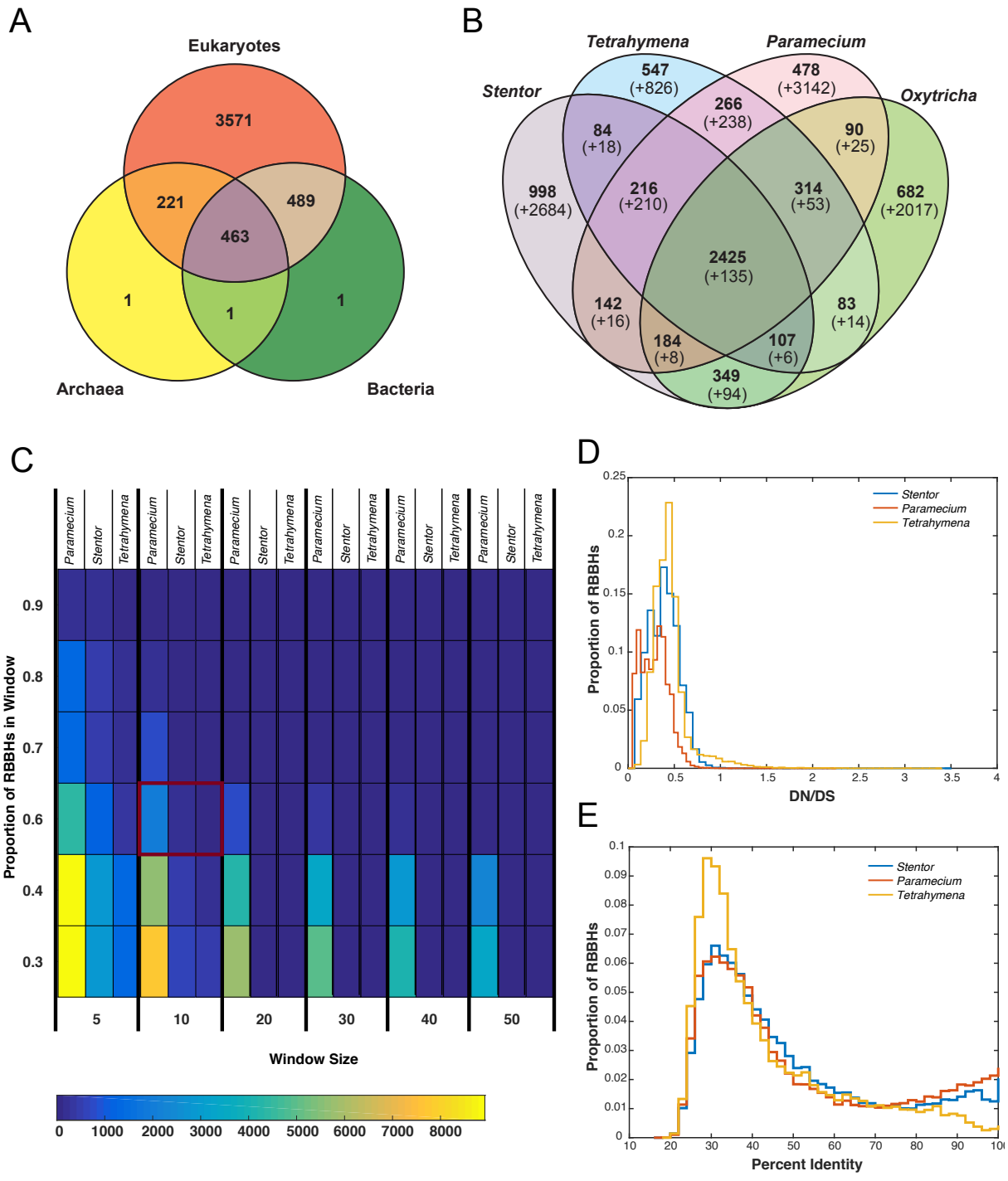
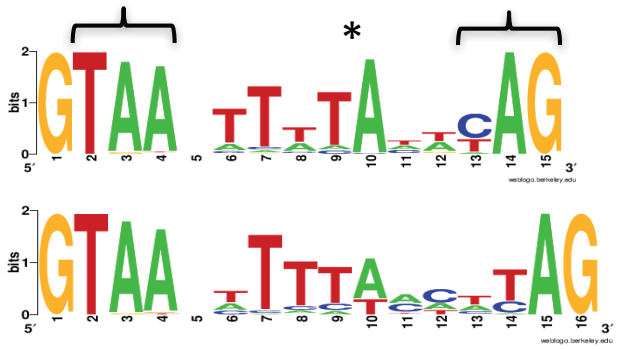

A
Eukaryotes
3571
221
489
463
1
1
1
Archaea
Bacteria
B
Tetrahymena
Paramecium
Stentor
Oxytricha
547
(+826)
478
(+3142)
266
(+238)
84
(+18)
90
(+25)
998
(+2684)
216
(+210)
314
(+53)
682
(+2017)
2425
(+135)
142
(+16)
83
(+14)
184
(+8)
107
(+6)
349
(+94)
C
Paramecium
Stentor
Tetrahymena
Proportion of RBBHs in Window
0.9
0.8
0.7
0.6
0.4
0.3
5
10
20
30
40
50
Window Size
0
1000
2000
3000
4000
5000
6000
7000
8000
D
Stentor
Paramecium
Tetrahymena
Proportion of RBBHs
DN/DS
E
Stentor
Paramecium
Tetrahymena
Proportion of RBBHs
Percent Identity

Figure S2. Orthology grouping of *Stentor* genes. (**A**) Venn diagram showing phyletic grouping of *Stentor* gene groups with the three domains of life. Includes curated orthology groups from OrthoMCL only. (**B**) Venn diagram showing shared orthology of gene groups from *Stentor* and three other ciliates. Bold numbers represent gene groups that are also found outside ciliates; numbers in parentheses represent gene groups that are exclusive to ciliates. (**C**) Effect of variation in criteria used for detecting duplicated syntenic blocks, depicted here for *Paramecium*, *Tetrahymena* and *Stentor*. Window size is the number of successive RBBHs analyzed. Red outline indicates the detection criteria used to generate Figure 2A of the main text. (**D**) Non-synonymous to synonymous substitution rates among *Paramecium*, *Tetrahymena* and *Stentor* for RBBHs. (**E**) Percent identity between RBBHs identified in the *Paramecium*, *Tetrahymena* and *Stentor* genomes. This figure supplements data presented in Figure 2 of the main text.

A


bits
5′
3′
weblogo.berkeley.edu

(95.2%)

bits
5′
3′
weblogo.berkeley.edu

(2.8%)

B

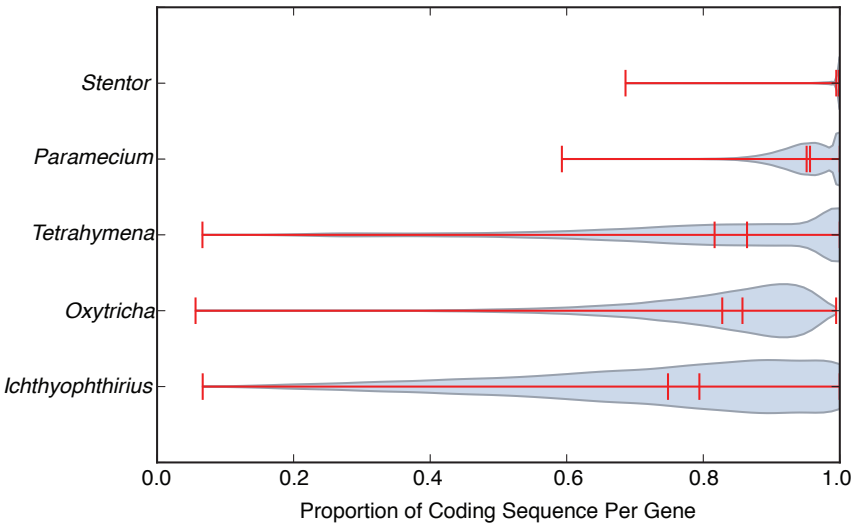

Stentor
Paramecium
Tetrahymena
Oxytricha
Ichthyophthirius
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of Coding Sequence Per Gene

Figure S3. (**A) Intron size in *Condylostoma***. As with *Stentor*, nearly all identified introns in the related heterotrich *Condylostoma magnum* [S1] are 15nts (95.2%, top) or 16nts (2.8%, bottom). *Condylostoma* introns also display the same unusual features as *Stentor* introns including an abbreviated 5' splice site motif, atypical internal TA dinucleotide (asterisk), and potential stop codons (brackets). **(B) Fraction of coding sequence per gene for select ciliates**. We compared the fraction of coding sequence per gene in *Stentor* with a selection of ciliates for which there are well-annotated gene models. For each of these ciliates, we show a violin plot which describes the distribution of this fraction. Whiskers indicate minimum and maximum values as well as both the mean and median of the distributions. Of note is that most of the genes found in the *Stentor* genome are entirely coding, so that the bulk of the distribution is at 1. The mean proportion of coding sequence per gene in the *Stentor* genome is 0.995. This figure supplements data presented in Figure 3 of the main text.

Table S1: Comparison of *Stentor coeruleus* macronuclear genome size with that of other ciliates. Related to Figure 1.

	Genome Size (MB)	Gene Number
Stentor coeruleus [this study]	83	34,506
Paramecium tetraurelia [S22]	72	40,000
Tetrahymena thermophila [S30]	105	27,000
Ichthyophthirius multifiliis [S31]	49	8,100
Oxytricha trifallax [S17]	50	~18,400
Stylonychia lemnae [S26]	50	15,102
Euplotes octocarinatus [S29]	89	-

Table S2. Intergenic Lengths in *Stentor* and other ciliate genomes. Related to Figure 3.

	Mean Intergenic Length (bases)	Standard Deviation Intergenic Length (bases)
Stentor coeruleus	2158.5	2733.0
Paramecium tetraurelia	2070.9	3049.6
Tetrahymena thermophila	4986.1	7393.0
Ichthyophthirius multifiliis	7465.2	8145.3
Oxytricha trifallax	559.8	856.3

Table S3. Table of correlation coefficients for raw ddPCR data. P-values calculated from Vassar online statistical calculator (http://vassarstats.net). Related to Figure 4.

Contig number	Size (bases)	correlation coefficient	n	P value
2	233043	0.68	23	1.94E-04
18	164593	0.80	15	1.82E-04
558	46105	0.56	23	2.63E-03
1255	20125	0.46	8	1.26E-01
1700	10437	0.59	8	6.30E-02
2224	4287	0.40	8	1.65E-01
2227	4280	0.56	23	3.00E-03

Table S4. Table of Selenoprotein gene models. Related to Figure 3.

gene	annotation	contig	start	end	strand
SteCoe_g40903	glutathione peroxidase	SteCoe_contig_4	5629	6135	+
SteCoe_g40904	glutathione peroxidase	SteCoe_contig_122	51947	52444	-
SteCoe_g6726	glutathione peroxidase	SteCoe_contig_94	61704	62234	-
SteCoe_g21264	glutathione peroxidase	SteCoe_contig_501	1252	1758	-
SteCoe_g16643	glutathione peroxidase	SteCoe_contig_334	13296	13826	+
SteCoe_g40905	glutathione peroxidase	SteCoe_contig_1253	9497	10018	-
SteCoe_g26857	glutathione peroxidase	SteCoe_contig_763	17484	17933	+
SteCoe_g4842	thioredoxin reductase	SteCoe_contig_62	62790	64337	+

Supplemental Experimental Procedures

Strains, media and growth conditions

Stentor coeruleus cells were obtained commercially (Carolina Biological Supply, Burlington, NC) but subsequently maintained in culture within the lab by growing in the dark at 20°C in Modified *Stentor* Medium (MSM), 0.75 mM Na_2CO_3, 0.15 mM $KHCO_3$, 0.15 mM $NaNO_3$, 0.15 mM KH_2PO_4, 0.15 mM $MgSO_4$, 0.5 mM $CaCl_2$, 1.47 mM NaCl modified from the original recipes described by Tartar and De Terra. This medium provides no nutrients and must be supplemented with living prey. We use *Chlamydomonas reinhardtii* grown separately in TAP medium and washed in MSM before being added to the *Stentor* cultures for feeding, but also include boiled wheat seeds in the cultures to promote additional microbial growth and give the *Stentor* fibrous material on which to anchor. 300mL *Stentor* cultures are given $3x10^7$ *Chlamydomonas* cells two or three times per week and grown with four wheat seeds. *Stentor* cultures are available upon request.

Stentor Imaging

To image cells for Figure 1, cells were starved for 24 hours and washed in sterilized media 2-3 times. Cells were fixed in ice cold Methanol for 10 minutes at -20° C and then incubated at room temperature in a 1:1 Tris-buffered saline (TBS):Methanol solution for 5 minutes. Here, we use a standard formulation for TBS: 0.05 M Tris and 0.15 M sodium chloride, pH 7.6, made in water. Following another room temperature incubation in TBS for 10 minutes, cells were blocked in a mixture of 2% BSA, 1xTBS, 0.1% Triton X-100 and 0.1% Sodium Azide for one hour at room temperature. Finally, cells were incubated with primary antibodies diluted in blocking mixture for one hour at room temperature. Cells were washed 3 times in TBS and then incubated with secondary antibody for 1 hour in the dark at room temperature. Cells were washed and mounted for visualization using a Deltavision Spectris deconvolution microscope. Brightfield images of *Stentor* cells were taken on a Zeiss AxioZoom microscope.

For volume estimation in Figure 4, living cells were imaged in the contracted state so that they became ellipsoidal, and volume was calculated from the axes of the ellipse of a cross section, assuming radial symmetry.

Genomic DNA isolation

300 cells from a clonal population (that was not inbred initially) were manually isolated and washed 3x in fresh MSM and incubated without additional food for 48 hours. After starvation, cells were again washed 3x in MSM and isolated in minimal media. Genomic DNA was isolated using the DNeasy Blood and Tissue kit (Qiagen), following

the suspension cell protocol, and eluting in 75 µL yielding 3µg. Whole cell DNA was isolated, and thus should contain DNA from both the macronucleus as well as the micronucleus, however we expect that the vast majority of the reads will be derived from the macronuclear genome for two reasons. First, when inspected by DAPI staining there are no micronuclei visible in our Carolina strains. Secondly, even if there are micronuclei present, the DNA content of the macronucleus is present at a copy number of approximately 50,000-100,000 (see Figure 4), while micronuclear genomes are present at single or a few copies, so that any micronuclear contamination would be present at levels less than a ten thousandth of the macronuclear DNA. We suspect that the cells in our cultures do contain micronuclei as we observe rare events of mating and are currently developing methods to better identify micronuclei using various approaches including immunofluorescence. Additionally, samples were checked for contamination with DNA from *Chlamydomonas reinhardtii* (the food source on which the cells had been grown) using PCR amplification of the *Chlamydomonas* mating type locus, but no bands were detected, confirming that our starvation and washing protocol eliminated the majority of the food cells.

Genomic DNA library preparation and sequencing

Paired-end sequencing libraries were prepared from 100 ng of genomic DNA using Nextera DNA sample preparation kit, Rev. A October 2011 (Illumina) following the manufacturers protocol. Libraries were sequenced using the HiSeq 2000 with Illumina's HiSeq paired-end cluster generation kit and HiSeq sequencing kit for 2 x 100 bp reads. Illumina paired-end sequencing on two lanes of a flow cell yielded a total of 629,226,200 paired-end 100nt reads (314,613,100 pairs total; lane 5: 156,618,004 pairs; lane 6: 157,995,096 pairs), with a median insert size of 130 nt. For each lane, sequences were 3´ truncated if a 90%-identity ungapped alignment was found to the beginning of the Illumina adapter sequence: CTGTCTCTTATACA. Partial matches at the 3´ end of each read were allowed, removing the overlapping portion of the putatively adapter-derived sequence. Paired-end reads were culled if either read of the pair was shortened to <97nt. This left 109,273,816 pairs (lane 5: 55,215,100 pairs; lane 6: 54,058,716 pairs).

Clamped homogeneous electric fields (CHEF) methods

To create agarose plugs, we used a modified version of the protocol from [S3]omitting zymolyase treatment. Briefly, *Stentor* were collected and excess media removed until the cells reached a concentration of ~100 cells per 50 uL. Cells were gently mixed with an equal volume of 1.25% low-melt agarose solution (50C), and pipetted into plug molds (100 uL each). Plugs were allowed to solidify, then incubated with Proteinase K overnight at 55C. The following day, the cells were washed four times for an hour each, then stored at 4C.

For pulsed field gel electrophoresis, we used the CHEF-DR II System (Bio-Rad). Briefly, we ran agarose plugs of *Stentor* in a 1% gel made from Pulsed Field Certified Agarose (Bio-Rad) in 0.5X TBE. Lambda DNA MonoCut Mix and Lambda PFG Ladder (New England Biolabs) were used as size standards. The gel was run at 6 V/cm2 with a 5s initial switch time and a 30s final switch time, for 20h at 14C. Afterwards gel was stained with ethidium bromide.

K-mer counting analysis

K-mer count histograms were generated using the filtered read data from above and the k-mer-counting software Jellyfish (v1.1.6) [S4]. Histograms for each even k-mer size over the range 16-30 were generated, and the mode of the right peak was estimated to determine the average fold coverage (**Fig S1A**). That number was used to divide the total number of k-mers in the dataset to yield an estimate of ~110Mb across k-mer sizes of 16-24nt, with the estimate for larger k-mers possibly amplified by genetic polymorphism or sequencing errors (**Fig S1A**). Two attempts were made to address sequencing errors: first, the calculation was performed removing all singleton k-mers from the count of total k-mers, presuming those to be the result of sequencing errors. Second, the quality scores of all nucleotides from the input dataset were used to estimate the number of k-mers in the datasets deriving from miscalled nucleotides. For each possible nucleotide score allowed by the fastq file format, the number of nucleotides with that score was determined. The average number of k-mers overlapping a nucleotide was estimated as k * (R - k + 1) / R, where k = k-mer size and R = read length. The number of nucleotides with a given score was multiplied by the probability of a nucleotide with that score being incorrect (for these files, 10 ^ ((64 - Q) / 10) where Q is the quality score. Those products were summed across scores, and the resulting tally multiplied by the number of k-mers expected to be affected by a mis-called nucleotide in order to arrive at an estimate of the number of erroneous k-mers. Neither error-inclusion method significantly altered the k-mer-counting size estimate (**Fig S1A**). We used this estimate to determine cutoffs for eliminating low-coverage sequences as described below.

Initial contig synthesis using SOAPdenovo

Initial scaffolds were generated using SOAPdenovo v1.05 [S5] with both lanes of adapter-filtered data described above, using the following command: "./SOAPdenovo-63mer all –K 63 –p 80 –R –s [config file] –o [output file prefix]". The config file made the following specifications for both data sets: "pair_num_cutoff=3, avg_ins=250, asm_flags=3, reverse_seq=0, map_len=90, rank=1, rd_len_cutoff=100". SOAPdenovo assembly was followed by running of the SOAPdenovo GapCloser program [S5] on the scaffold output from SOAPdenovo and the config file from that run. By using the scaffold output from SOAPdenovo, instead of just the contig output, larger inserts could be used to generate scaffolds. The output file was re-formatted from fasta to priceq format as described in the next paragraph. The 201,835 scaffolds, totaling 111,219,579nt in length, were fed into the PRICE assembler (v0.18, http://derisilab.ucsf.edu/software/price/) [S6] without any read data for a collapse of redundant sequences using the following command: "./PriceTI -icf [priceq-formatted SOAP output file] 1 1 1 -TPI 95 -nc 1 -a 20 -o [fasta-format output file]". That collapse yielded 139,937 "contigs" (collapsed scaffolds that may have become contiguous in the process) of 101,837,818 nt total length. The PRICE-collapsed scaffolds were cut at any stretch of 2 or more consecutive uncalled nucleotides (N's), with terminal N's trimmed from the split-apart contigs, and contigs <100nt removed, yielding 140,051 contigs totaling 101,833,627nt in length. That fasta file was reformatted to priceq format as described below.

Fasta-format contig files were converted to priceq-format, a format specifically designed for compatibility with the PRICE assembler (though not required for its functionality) using BLAT [S7] to align reads to the SOAPdenovo contigs. Reads were aligned requiring 90% identity across the entire read. Coverage of reads across each nucleotide of each contig was then transformed into a priceq scores using the formula specified for that file format (http://derisilab.ucsf.edu/software/price/) [S6]. Reads overlapping two consecutive nucleotides contributed a count to the phosphodiester score between those nucleotides. The resulting tallies were written out in priceq format and used for further assembly steps.

Manual examination of the resulting contigs revealed a large number of contigs that began with repeats of the 8nt sequence "CCCTAACA". This repetitive motif only occurred at the 5′ ends of contigs, with repeats of the complementary sequence "TGTTAGGG" appearing repeatedly at the 3′ ends of contigs. We presumed that this sequence derives from telomeric sequence, and filtered it as a repetitive sequence from extending contigs during the PRICE assembly described below.

Contig extension and collapse using PRICE

The contigs generated by SOAPdenovo were extended and collapsed using the PRICE assembler (v0.18, http://derisilab.ucsf.edu/software/price/) [S6], which was developed for assembly of metagenomic datasets. Metagenomic assembly is appropriate because our sequencing library constructions used total DNA isolated from *Stentor* cells, which includes DNA from the macronuclear genome but also from mitochondria and the micronucleus. We note, however, that although our library used total DNA, the vast majority of DNA in the cell is present in the macronucleus, so much so that micronuclei are not even detectable when whole fixed cells are stained with DNA dyes (**Figure 1B**). Whereas the micronucleus is diploid and thus contains two copies of the genome, we have found (**Figure 4**) that the macronucleus is on the order of 100,000-ploid. We thus expect that any micronuclear DNA present in our sample would only constitute a miniscule fraction of the total library. Nevertheless, PRICE provides an additional layer of robustness because of its established ability to separately assemble genomes from metagenomic mixtures. We executed PRICE using the following command: "./PriceTI -icf [seed priceq file] 10 1 1 -fpp [filtered read files, lane 5] 250 97 -fpp [filtered read files, lane 6] 250 97 -badf [telomere fasta file] 80 -lenf 100 0 -TPI 95 -targetF 95 0 -rnf 95 -nc 10 -a 20 -o [output fasta and priceq-format files]". The "telomere fasta file" contained a single sequence entry: the telomere repeat 8-mer "CCCTAACA" (see above) repeated consecutively 30 times. Although that job was set to run for ten cycles of extension, it was terminated after seven cycles, in each cycle importing an additional 14,005 contigs from the input contig file. The resulting output included 23,016 contigs with a total length of 97.7Mb and a contig N50 of 55.2kb (**Fig S1B**).

Error correction using PRICE modules

Several strategies were implemented to address assembly artifacts of unknown origin that were evident from manual inspection of the assembly, and confirmed by PCR to be errors. Scripts to address each area available for download from http://derisilab.ucsf.edu/software/price/accessories/. First, short stretches of sequence were found to be tandemly duplicated at many genomic loci, generally with two copies of near-perfect identity separated by a single "N". Such tandem duplications were collapsed using the script "correctShortTandem.py", which for each specified repetitive region performs a gapped self-against-self alignment and, if an alignment is found meeting

minimum percent identity and length requirements, collapses the tandem duplication into a single copy. This script was run many times in succession, as multiple tandem duplications could at most be reduced in copy number by half. A minimum percent identity of 90% was specified for collapsing redundant sequences.

Following the collapse of tandem repeats, BLAT [S7] was used as above to generate coverage maps of the genome assembly. Reads mapping to multiple genomic loci had their counts normalized across all the loci to which they could be mapped with an equal (highest) score. The coverage distribution for genomic loci, defined here as non-overlapping 200nt bins, was bimodal (**Fig S1C**). Given the observed distribution, 200nt blocks of sequence with <10X coverage were removed from the genome assembly using "correctLowCovRegions.py" (http://derisilab.ucsf.edu/software/price/accessories/), changing the contig count/total assembly size/contig N50 to 23,670 contigs / 92.4Mb / 40.4kb. Those contigs were provided to PRICE for a no-read-input cycle to collapse redundant contigs (v1.0.1, "./PriceTI -icf [contig fasta file] 1 1 10 -nc 1 -MPI 97 -TPI 25000 -o [fasta/priceq output file]"). That yielded a contig count/total assembly size/contig N50 of 22,187 contigs / 89.4Mb / 41.2kb.

An additional cycle of PRICE with reads was launched, this time using adapter-trimmed reads from above that were further filtered for high-quality read pairs only using PriceSeqFilter (v1.0.1 "-rqf 95 .99"). That cycle, run using PRICE v1.0.2, used the following command: ".PriceTI -icf [input contig file] 1 1 10 -MPI 97 -TPI 25000 -fpp [quality-filtered lane 5 reads] 250 97 -fpp [quality-filtered lane 6 reads] 250 97 -badf [telomere repeat file] 80 -lenf 200 0 -targetF 98 0 -nc 5 -mol 30 -o [fasta/priceq output file]". Though specified for five cycles, that job was terminated after a single cycle, yielding a contig count/total assembly size/contig N50 of 19,940 contigs / 95.5Mb / 48.9kb. Tandem-repeat correction was repeated, not significantly altering the assembly size statistics.

Using coverage maps again generated with BLAT, low- and high-coverage regions of the assembly were split away from the rest of the genome using the script "correctLargeRepeats.py" (http://derisilab.ucsf.edu/software/price/accessories/). High-coverage regions were defined using a 5nt-resolution coverage map of the genome generated with all quality-filtered reads ("-mc 600 1200 -hl 40 -ll 2000 -he .72 .11 .17 -le .19 .09 .72"), as were low-coverage regions ("-mc 10 15 -hl 200000 -ll 200000 -he .6 .2 .2 -le .2 .2 .6 -low"). The high-coverage and average-coverage genomic blocks were re-collapsed using a price no-read, collapse-only, single-cycle run (v1.0.3; "./PriceTI -icf [high-coverage file] 1 1 10 -icf [medium-coverage file] 1 1 10 -MPI 95 -TPI 100000 -nc 1 -o [output fasta file]"). That yielded a contig count/total assembly size/contig N50 of 15,384 contigs / 88.5Mb / 47.4kb.

The alignment of contigs to NT revealed some with high-identity matches to sequences from *Janthinobacterium agaricidamnosum* NBRC 102515 (taxid 1349767). Contigs were isolated from the larger assembly if they shared more than 50% of their sequence with the *J.agaricidamnosum* genome (HG322949; unpublished direct submission) at >80% identity when aligned by Blastn [S8], or with annotated proteins from that species >100 amino acids in length that could be aligned by blastx with a proteome-specific expect value of <1e-5 and >80% identity. Contigs thus isolated were subjected to further cycles of PRICE extension and consolidation. A more limited number of contigs were aligned and assigned to *E.coli* (CU928161.2; direct submission). No satisfactory matches were found to the *Chlamydomonas reinhardtii* genome (confirming our PCR results described above), nor to the *Triticum urartu* (wheat) genome, suggesting that during our starvation and washing procedure prior to DNA isolation, the food sources (Chlamydomonas cells and wheat seeds) were largely removed. Matches were found to the *Triticum aestivum* genome assembly (assembly ID GCA_000334095.1), but those were found to have equally good matches to the genome of bacteriophage S13. Additional blast searches revealed a limited number of contigs with high-quality matches to common laboratory plasmids; these were presumed to derive from laboratory contaminants and are presented separately.

Analysis of bacterial contaminating sequences by PCR

Because our initial assembly contained a contig consisting of bacterial sequences, we tested whether this contig represented bacterial contamination using a PCR approach. DNA samples were prepared from whole cells. Single cells were washed 3x in MSM, isolated in 10 μL, and then incubated in 9 μL 2x PCR buffer and 1 μL proteinase K (New England Biolabs, Ipswich, MA) for 1 hr at 55°C. The proteinase K was heat inactivated at 95°C for 10 min and the resulting solution was used as a DNA template for PCR reactions using the following primers:

Target	Forward Primer	Reverse Primer
Chlamy P2 (Plus mating type)	GCTGGCATTCCTGTATCCTTGACGC	GCGGCGTAACATAAAGGAG GGTCG

Chlamy M3 (Minus mating type)	CGACGACTTGGCATCGACAGGTGG	CTCGGCCAGAACCTTTCATAGGGTGG
Janthinobacterium sequence	GCAAGCATTATCTGGCGGTG	TCGAGCAGCGATTCCTGATC
Stentor β-tubulin	ATGAGAGAAATTGTTCACGTACAAGGC	GGAGTAGTGAGCTTAAGAGTTCTGAAGC

PCR using the *Janthinobacterium* specific primers of bacterially contaminated MSM media (produced by adding wheat seeds to MSM media and growing in open air), revealed distinct bands that differed in size from the predicted product from the assembled *Janthinobacterium* sequence in our assembly. In all cases, Sanger sequencing of these amplification products identified sequences homologous to several different soil bacteria (*Mesorhizobium* and *Thuaera)*. In no case did these amplification products exactly match the *Janthinobacterium* sequence. These same primers failed to amplify any bacterial sequences from *Stentor* cells that had been carefully washed from their growth media. We therefore conclude that the *Janthinobacterium* reads that were assembled in our genome assembly represent contamination from the growth media during our initial sample preparation, and not an endosymbiotic bacterium within the *Stentor* cells themselves. The contig was therefore removed from the final assembly.

Analysis of SNP density

Since our genome was not sequenced from an inbred population of cells, we sought to assess the heterozygosity of the genome by measuring the SNP density. To this end, we employed three different approaches. The first approach is a reference-free approach to identify SNPs, DiscoSNP++ [S9], which we run with default settings, including those for mapping back to the genome. Using VCFtools [S10], we analyzed the SNP density [parameter: --SNPdensity 1500] and identified 1.4 SNPs in windows of 1500 bases. The next two approaches were referenced based approaches. First, we used a combination of samtools mpileup [parameters: -uf] [S11-S13]and bcftools [parameters: bcftools call -c -v -o b, followed by vcfutils.pl varFilter -D100] (http://samtools.github.io/bcftools/call-m.pdf). Using vcftools as above, we identified 1.2 SNPs in windows of 1500 bases. The final approach we used was based upon the Genome Analysis Tool Kit (GATK) [S14,S15], following published best practice protocols [S16]. Using VCFtools, we found 4.1 SNPs in windows of 1500 bases. SNP density was consistent across all contigs. In particular the SNP density on the contig containing the rDNA locus was comparable to the genome as a whole.

Detection of telomeres

In order to identify contigs that are capped on one or both sides by telomeric sequences, we created a library of reads containing telomeric sequences. Following the approach of [S17] we selected all paired reads matching the regular expression CCCTAACA[CAN]*, masking all matches with a single N. We restricted future anlaysis to all pairs where both reads were >= 30 bp long (259,312 pairs). We mapped all reads to the genome assembly using gmapper version 2.2.3 in paired mode with the following flags: -p col-bw -h 80 -I 0,30000 -N 16. We then searched for contigs with at least 10 reads mapping to either end.

RNA Isolation and cDNA synthesis

RNA was isolated from *S. coeruleus* cultured cells growing vegetatively using the RNeasy kit (Qiagen, Venlo, Netherlands), following the manufacturer's instructions. RNA for mRNA-seq libraries was isolated from two samples of 1000 cells each, and RNA used for traditional cDNA synthesis or for RACE was isolated from 500 cell samples. cDNA synthesis was performed using the SuperScript III First Strand Synthesis System (Life Technologies, Carlsbad, CA), following the manufacturer's instructions and priming with oligo-dT.

Sanger sequencing and RACE

For Sanger sequencing of cDNA and genomic DNA regions, we selected gene models generated by either CEGMA (http://korflab.ucdavis.edu/datasets/cegma/) or MAKER (which we initially used to predict genes based on homology to related ciliates) (http://www.yandell-lab.org/software/maker.html) that were predicted to have at least one intron. The following primers were designed to contain predicted start and stop codons of these putative genes:

Primer Name	Sequence
SteCoe_contig_916-1F	5’- ATG GAG TAT CTG GAA ACT TTA CC –3’
SteCoe_contig_916-1R	5’- TTA ACT ATC TAT TTC CAT AGG GAC TTC -3’
SteCoe_contig_1057-1F	5’- ATG GCA GCA ATC GGG GTA AG -3’
SteCoe_contig_1057-1R	5’- GAC ATA GCA AGC GAA AGG GC -3’
SteCoe_contig_64-1F	5’- ATG AGT GGA GCT GGA ACA GG -3’
SteCoe_contig_64-1R	5’- CTA CTC ACC ACG TTC TTC TCT TTC -3’
SteCoe_contig_212-1F	5’- ATG TCG GGC CAT TAT TCC TC -3’
SteCoe_contig_212-1R	5’- CTA ATA TCT TCT CGG GCT ACG AC -3’
SteCoe_contig_78-1F	5’- ATG GAA AGC AGA AGA CTC C -3’
SteCoe_contig_78-1R	5’- CAT TAA ATT ACC TAA GCT GAT GAT AG -3’
SteCoe_contig_270-2F	5’- ATG ACT ACA CCT GCA AGA AGA AG -3’
SteCoe_contig_270-2R	5’- TTA ACT ATT GCA CCA GGA GTC TTC -3’
SteCoe_contig_172-1F	5’- ATG GAC TAT GTA GAA GTG GTC G -3’
SteCoe_contig_172-1R	5’- CTA ATT CTC CTG ATC ACT CC -3’
SteCoe_contig_3-1F	5’- ATG GCA CAG TTC TCA AGA TAT G -3’
SteCoe_contig_3-1R	5’- CTA TCT ATC AAC TTC CAT ATC TTC ATC -3’

These primers were then used to perform PCR from cDNA or genomic DNA with Phusion polymerase (New England Biolabs, Ipswich, MA). PCR products were then cloned into plasmids using the Zero Blunt TOPO PCR Cloning kit (Life Technologies, Carlsbad, CA).

RACE was performed using the SMARTer cDNA amplification kit (Clontech, Mountain View, CA), following the manufacturer’s instructions. The following gene-specific primers were used:

Primer Name	Sequence
SteCoe_contig_754-3’RACE1	5’- GGA AGA AGA AGA TAA TGG GCA GGG C -3’
SteCoe_contig_754-5’RACE1	5’- CCA GTC TTG TAA GAA ACC CAA CGA GGC -3’
SteCoe_contig_218-3’RACE1	5’- GAT TCG CCG ACA ATA CCT ACA CTG AGA G -3’
SteCoe_contig_218-5’RACE1	5’- GTT GAG ATT TCT GCT GTG ATG CTA CCG G -3’
SteCoe_contig_270-3’RACE1	5’- GCA ATG GAC CG CGT TTG GGA GC -3’

SteCoe_contig_270-5’RACE1	5’- AAC CCA TCC TTA TCA CAC ATG CAG CC -3’
SteCoe_contig_295-3’RACE1	5’- GGG ATT GTT GGT GCC CAA GTC CCT GTT G -3’
SteCoe_contig_295-5’RACE1	5’- CAG CTC TTT AGC ATC AGG CAC AGG GTC -3’
SteCoe_contig_884-3’RACE1	5’- GAA GAA GCG AGA CGA AGA ATT GCC CGA C -3’
SteCoe_contig_884-5’RACE1	5’- CCA GCA TGA ATA GCC GTA CTC GGA AAC C -3’
SteCoe_contig_1210-3’RACE1	5’- CAG AGC CAA TCT CAT CAT GGA GCC -3’
SteCoe_contig_1210-5’RACE1	5’- CCC TGC TCT ACC TGC TCT TCC TAT CC -3’
SteCoe_contig_127-3’RACE1	5’- CCT CCT GCT TCG TGA AGG AAC TGA CAC -3’
SteCoe_contig_127-5’RACE1	5’- GAA TCT GCG TCC TCT GCC TCT TCC -3’
SteCoe_contig_5-3’RACE1	5’- CTG GGA TAC AGC AGG TCA AGA ACG G -3’
SteCoe_contig_5-5’RACE1	5’- CAG CAC TTC CCT TCA CCT TTC TTA TCC G -3'
SteCoe_contig_2266-3’RACE1	5’- CCT GGT AGT TGC TGC GAC TGA CGG C -3’
SteCoe_contig_2266-5’RACE1	5’- GTT CCT GCT CCT GCT CCA TCC TCC -3’
SteCoe_contig_1522-3’RACE1	5’- TGA GCG AGG TAT CAC CGT AAG AGC CC -3’
SteCoe_contig_1522-5’RACE1	5’- CGT CCC TGA GTC AAC CCT AAC ATT TCC -3’
SteCoe_contig_323-3’RACE1	5’- CGT GAC TCT CGG GTC TTT CTT ATC GGT G -3’
SteCoe_contig_323-5’RACE1	5’- TGC GTT TAC CAC ATT GAC AGC CCT TG -3’
SteCoe_contig_282-3’RACE1	5’- GGG AGT CAA TGG CAG GAG GTA ACT TTG -3’
SteCoe_contig_282-5’RACE1	5’- CCA GGA GGT CCA CAA TAG CAC ACA AGA G -3’
SteCoe_contig_79-3’RACE1	5’- CCA GTT GGT GCT GAC CTG TTT GTG ATT G -3’
SteCoe_contig_79-5’RACE1	5’- CTG GCT CTT CAA CCA TGC TCT TGA TAG C -3’
SteCoe_contig_698-3’RACE1	5’- CAA TCA AAC ACA CCA GCA ACC CTT CG -3’
SteCoe_contig_698-5’RACE1	5’- ATC ACA GGT CGG TCC CCA AAT CAC AG -3’
SteCoe_contig_275-3’RACE1	5’- CAG AAG AGT TTG GAA GCG GTT GGG -3’
SteCoe_contig_275-5’RACE1	5’- GAC CCT GCT GTG ACT TGC CAG ATT TC -3’
SteCoe_contig_1628-3’RACE1	5’- CAG CTT ATG GGC CAA GTG ACA ATC CGCC -3’
SteCoe_contig_1628-5’RACE1	5’- CGA CAA CCG ACC CAT CAG GAA GTT C -3’
SteCoe_contig_1010-3’RACE1	5’- GAA CTT GAT CCT CGT ATG GTT GCC G -3’
SteCoe_contig_1010-5’RACE1	5’- GGT CCT TCC CCA TTG GCT TCT CTT AG -3’
SteCoe_contig_1024-3’RACE1	5’- GAT CTT GGG ATT GGA GGA GCA GAA CAG -3’
SteCoe_contig_1024-5’RACE1	5’- CGC TTG ATT GAA CTT TGA CGC TGG GTG C -3’
SteCoe_contig_184-3’RACE1	5’- CCT CCC CTC CCA ACA CCC GCA AG -3’
SteCoe_contig_184-5’RACE1	5’- CTT GCG GGT GTT GGG AGG GGA GG -3’
SteCoe_contig_939-3’RACE1	5’- GAT TAT CTC GCT GGT ACA CAA CAA TTC C -3’
SteCoe_contig_939-5’RACE1	5’- GCT TAG AAG TCT CTG TAA TTT CCC CTC C -3’

RACE reactions were then analyzed on an agarose gel, and major products excised and TA-cloned into pCR2.1 with the TOPO TA Cloning kit (Life Technologies, Carlsbad, CA). Cloned RACE fragments, cDNA regions, and genomic DNA regions were all Sanger sequenced at Elim Biopharm (Hayward, CA) using M13 F and M13 R primers.

mRNA-seq Library Preparation and Sequencing

5 ug total RNA was used to create a strand-specific mRNA-seq library, as previously described [S32]. Library quality was tested on a 2100 Bioanalyzer (Agilent, Santa Clara, CA). QC analysis was then performed by running a small amount of the libraries on an Illumina GAIIx to get 5-10M 50-bp single end reads from each library. After these reads were analyzed a larger run was performed on the Illumina HiSeq 2500 at the Center for Advanced Technology at UCSF in rapid run mode to get 100 bp single-end reads.

RNA-seq Analysis

RNA-seq reads were trimmed with Cutadapt (https://code.google.com/p/cutadapt/) to remove any adapter read-through at the 3' ends of reads, and then with Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) to remove 5' adapter sequence, before quality filtering with FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Trimmed and filtered reads were then mapped to the genome assembly using Bowtie 2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). We ran Bowtie 2 in local mode so that reads at exon-exon junctions would be more accurately mapped.

Assessment of assembly by Identification of Core Eukaryotic Genes

We used CEGMA (v 2.5) to analyze the *Stentor* genome for core eukaryotic genes, using the default parameters. From this analysis, we found 202 of the 248 core eukaryotic genes defined by CEGMA. Following the approach of Swart et al, we reduced the restrictions of the CEGMA search in order to find evidence for the remaining genes. Briefly, using hmmscan (HMMER 3.1b1/May 2013; http://hmmer.org/), we searched the Pfam-A HMM profiles in order to assign each EuKaryotic Orthologous Groups (KOG) to the best Pfam domain. A domain was assigned to a KOG if it was the best domain assignment for the majority of KOG members and had a domain, full-sequence E-value < 1e-3. We then used these Pfam domains to search all detected ORFs in the *Stentor* genome using hmmscan with a domain, full-sequence e-value < 1e-3. From this search we identified 29 additional core eukaryotic genes. Finally, for the remaining KOGs, we searched the EggNOG database for updated HMMs [S18]. We repeated the scan of all detected ORFs and then verified hits with a BLAST search. Using this approach, we detected 12 additional COGs. In total, the sequenced *Stentor* genome contains strong evidence for 243/248 core eukaryotic genes (indicating that 98% of core genes are present in the assembly). The missing five include KOG2719 (a metalloprotease), KOG1523 (an actin-related protein member of the Arp2/3 complex), KOG2311 (NAD/FAD-utilizing protein involved in translation), KOG1712 (adenine phosphoribosyl transferases), and KOG2653 (6-phosphogluconate dehydrogenase).

Gene Predictions

We generated a set of 307 hand curated gene models that were verified by Sanger sequencing of cDNA and/or RNA-seq data. We note that all introns verified by sequencing were either 15 or 16 bases. Of these 70% were single exon genes. The rest included at least 1 15-16 base intron (22% included 1 such intron, 6% included 2 introns, and 1% included at least 3 introns). We used Augustus (3.0.2) [S19,S20] to perform gene predictions, only training with gene models which were <70% identical at the protein level as according to Augustus documentation. In order to detect the appropriate minimum intron length, we altered the source code (filename: extrinsic.cc) so that that the minimum intron length possible is 9 bases (the default is 39). Additionally, we altered the parameters of the signal models to be the minimum length possible (filename: intronmodel.cc, types.cc) (personal communication, Mario Stanke). After recompiling Augustus, we used the instructions found here as a guide (http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/training.html) and trained Augustus for *Stentor*'s genes using half of our hand curated gene models, testing on the other half. Under these conditions, 90.6% of the testing set was predicted exactly at the gene level, 88.6% of the exons were predicted exactly and 91.2% of the predicted exons were exactly as in the test set.

In order to generate hints for introns for gene predictions, we used Tophat2 (v2.0.11) [S21] to align RNAseq reads from vegetative cells to the repeat masked *Stentor* genome (in order to ensure small introns are detected, we used the following flags: tophat2 -i 9 -I 101 --min-segment-intron 9 --min-coverage-intron 9 --max-segment-intron 101 --max-coverage-intron 101). Since we did not have a significant number of sequenced UTRs in our gene models, we did not include these in our gene predictions. In addition to intron hints, we used the above Tophat output to generate exon hints following the instructions here: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.GSNAP. We then ran Augustus with the following flags, including both intron and exon hints: --alternatives-from-evidence=true --hintsfile=hints.gff --allow_hinted_splicesites=atac.

Upon further Sanger sequencing of a subset of predicted gene models, we found that the only verifiable gene models were those that were single exons or contained introns of 15 or 16 bases. Introns of different length most often appeared to arise due to genome mis-assembly (i.e. the gene model contained an N). Additionally, we only found support for gene models that included an GTR-AG (R = A or G) splicing signal. Furthermore, we found that some predicted genes models were in fact falsely joined gene models. We wrote a Perl script to filter any gene models that met these criteria from the final set we used for all downstream analysis. In the case of falsely joined gene models, we simply split the models based on the presence of stop and start codons.

In an effort to predict the lengths of UTRs genome wide, we combined both our RNAseq data and predicted gene models. Searching upstream/downstream of a predicted gene 's start/stop codons, we marked the start and stop coordinates of the UTRs as the first/last points at which reads aligned to the genome adjacent to a gene but outside the ORF.

Analysis of extent of genome duplication based on synteny

We sought to explain whether the apparent expansion of genes in the *Stentor* genome was due to genome duplication or gene duplication events. To this end, following the methods of Aury et al [S22], we found all reciprocal best blast hits (RBBHs) among the translated gene models. Using blastp, we compared all translated gene models against each other, retaining those hits with an e-value <1e-5. A pair of genes is considered to be a best blast hit if the reciprocal search has the same e-value. A gene can have no more than 5 RBBH pairs. Restricting our search for genome duplication events to the universe of RBBHs, we slid a window containing 10 RBBHs across each contig. If 60% of the RBBHs within a window paired with a window on another contig, this was considered a paralogous block. Contiguous blocks were merged if they paired with a common contig. We repeated the analysis for the *Paramecium tetraurelia* genome (version 99.13) and the *Tetrahymena thermophila* genome (June 2014 version). Results were visualized using Circos version 0.69 [S23]. Since the initial conditions we used were optimized for the *Paramecium tetraurelia* genome, we extended this analysis by varying the window sizes and proportion of RBBHs required to define a syntenic region (Supplemental Figure S2, panel C). We note that not all contigs were long enough to contain the minimum of 10 RBBHs to be used for this analysis. These short contigs are still included in Figure 2A, and account for the lack of syntenic blocks in the upper left half of the Circos plot. These contigs account for 29% of the assembly.

Phylogenetic analysis

We used rnammer (v. 1.2) [S24] to identify the genomic region that corresponds to *Stentor*'s ribosomal RNA and found that contig_2227 contained the 18S and 8S ribosomal subunits. This was confirmed by blastn to the nt database. For 18S rRNA comparisons, we downloaded the following sequences from NCBI: *Oxytricha trifallax*--FJ545743, *Stylonychia lemnae*--AM086653, *Euplotes crassus*--AJ305255.1, *Tetrahymena thermophila*—M10932, *Ichthyophthirius multifiliis*—IMU17354; *Paramecium tetraurelia*—AB252009, *Blepharisma japonicum*--AM713185.1 and an outgroup, human NR_003286.2. Using Geneious, we performed a multiple sequence alignment with clustalW using the default parameters. We then used the Geneious tree builder to build a neighbor-joining tree using an HKY substitution model with human as the root of the tree.

Genetic code analysis using MS Data

In order to identify the tRNA genes encoded by *Stentor*'s genome, we searched the genome using tRNAscan-SE with the default settings (v. 1.23) [S25]. Similarly, we searched the genomes of *Oxytricha trifallax*, *Paramecium tetraurelia*, *Ichthyophthirius multifiliis* and *Tetrahymena thermophila*. Evidence for the remaining ciliates genetic code was gathered from the literature – *Blepharisma* [S2], *Euplotes* [S2], and *Stylonychia* [S26]. For human, the genetic code was obtained from the Genomic tRNA database (http://gtrnadb.ucsc.edu).

To further understand the genetic code usage by *Stentor* we used proteomic information obtained by mass spectrometry analysis of total cellular protein. Whole-cell protein samples were prepared by snap freezing cells,

lysing in buffer in the presence of protease inhibitors (Roche), precipitating the protein with 10% TCA and then removing the lipid by acetone extraction. Proteins were resuspended, digested with trypsin and analyzed using a Q-Exactive hybrid quadrupole mass spectrometer (Thermo Scientific). We then used the mass spectra to test which genetic code gives the most consistent predictions for peptides based on the assembled genome. We translated the genome in 6 frames using four genetic code tables (standard, ciliate *Blepharisma*, and UAR-glutamate ciliate). Peptides identified through Mass-spec were aligned to these translations using ProteinPilot (SCIEX) version 5.0 with the Paragon Algorithm. The AB Sciex search engine ProteinPilot™ v. 5.0 (Revision Number 4769) with the Paragon™ Method algorithm 5.0.0.0, 4767 (Shilov et al, 2007) was employed for peptide and ORF identification. A total of 51306 non-empty spectra were searched utilizing the following parameter settings: Identification for "Sample Type"; Iodoacetamide for "Cys Alkylation"; Trypsin for "Digestion"; Orbi-FT MS (1-3 ppm)/LTQ MSMS for "Instrument"; None for "Species"; Thorough ID for "Search Effort"; Biological modifications Variants: Evolutionary for "ID Focus"; Yes for "FDR Analysis"; No for "User Modified Parameter Files"; and 2 for "Competitor Error Margin (ProtScore)". "Detected Protein Threshold" was set to 0.05 (10.0%).

Using the most confident peptide alignments determined by the Paragon algorithm (peptide identification threshold greater than 95% confident), we searched for cases where codons encoded alternative amino acids as in the genetic code employed by many model ciliates (UAR encodes glutamine), another employed by *Blepharisma* (UGA encodes tryptophan) and a third employed by few ciliates (UAR encodes glutamic acid). We wrote a custom script to identify open reading frames (ORFs) in the *Stentor* genome using these genetic codes in all six frames, defining an ORF as occurring between two stop codons. We then used a custom script to find cases where a mass spec peptide was found in an ORF. The vast majority of alternative codons used corresponded to read-through events. For every predicted gene model, we defined a read-through event as the extrapolated peptides that would occur if read-through occurred at the stop codon (i.e., a stop codon was translated to a Q, E, or W if it were a UAA/UAG or UGA). We then searched for evidence of support of the read through peptides using mass spec data and found that these events accounted for the majority of occurrences of alternative codon uses. If the alternative codons were found in an open reading frame, we BLASTed the translation of the ORF using the alternative encodings as well as the standard encodings. When the BLAST hit for a standard ORF was better or the same as that of the alternative encoding, this was considered good evidence for a read-through event. Otherwise, the alternative encoding for an ORF was examined more closely and verified using BLAST as well as manual inspection of the mass spectra for the underlying peptides.

In addition to the dominant class of alternative stop codon-containing ORFs described in the main text (those that matched better with the standard code, suggesting translational read-through), a smaller fraction of the ORFs (13% Ciliate; 14% *Blepharisma* table; 6% UAR-glutamate), showed a better BLAST hit when translated using the alternative codon table than did the corresponding ORF translated with the standard table. But all of these cases, homology was only found to predicted or poorly annotated proteins (e-value < 1e-5, percent identity > 20), suggesting the ORFs in question may not correspond to actual protein encoding genes. The remaining ORFs translated with alternative codon tables (1.6% Ciliate; 6.9% *Blepharisma*; 1% UAR-glutamate) did not have a corresponding standard ORF at all, and in all cases these ORFs did not show strong homology to the BLAST database, again suggesting that many of them may be spurious ORFs that do not correspond to protein coding genes. Finally, two ORFs translated by the Ciliate and *Blepharisma* tables, each of which lacked a corresponding standard ORF, exhibited strong homology to a *Tetrahymena* small nuclear ribonucleoprotein.

Calculation of Intergenic Lengths

To calculate the intergenic lengths in the Stentor genome and to avoid any bias that might arise from genes that were not predicted by Augustus, we included open reading frames (ORFs) of a minimal length of 450 nt. We found these ORFs using getorf [S27].

Estimating Ploidy by Droplet Digital PCR

Single cell DNA samples were prepared as described above, and 2 μL of the sample were used as the DNA template in the ddPCR reaction using the following primers:

Contig	Forward Primer	Reverse Primer	Probe

SteCoe_contig_2	AAAGATGGCCAAGTGCAAAG	TCGTTCTAATCCTGCCATATCC	AGTCCAGATCCTACAATTGGAGTATGT
SteCoe_contig_18	TGTACTGCTCAAAGGTACACTAAG	CATTGATGCAGCTTGAAGATAAGG	CACCTTCAGACGATTGCTCATTCATTGC
SteCoe_contig_43	ACCTTCTTCCACATCACAATCT	AGAGATCATGGGAGGTTATAGGA	ACCCATCATCCAACATCCTCCTCTCT
SteCoe_contig_558	CCTACTCGGCCCATCAAATC	TCAGAAGCTAGCTCAGGATACA	TGCACAGACCAAATCCCATTGTCTCT
SteCoe_contig_2227	CCTACCGATTTCGAGTGATGAG	CCTTGTTACGACTTCTCCTTCC	TACTCAACTTCCCAACGCCGAAGC
pPR-T4P Plasmid	CTACATACCTCGCTCTGCTAATC	GCGCCTTATCCGGTAACTATC	AAGACACGACTTATCGCCACTGGC

Dual labeled probes were ordered with either 5'-FAM or 5'-HEX as the fluorescent indicator and ZEN-Iowa Black quenchers (Integrated DNA Technologies, Inc., Coralville, IA). ddPCR reactions were prepared using the 2x ddPCR Supermix (Bio-Rad) with target amplification primers (900 nM) and probes (500 nM) on the QX100 ddPCR system (Bio-Rad). Droplet generation, PCR, and droplet detection were performed following the QX100 system protocols (Bio-Rad). Briefly, 25 μL PCR samples were loaded onto 8-well cartridges with 65 μL of droplet generation oil and placed on the droplet generator (Bio-Rad). Droplets were then loaded onto 96-well PCR plates, heat-sealed, and PCR was performed on a standard thermal cycler. Plates were then transferred to the QX100 droplet reader (Bio-Rad) and analysis was performed using QuantaSoft (Bio-Rad). In order to determine the ploidy of a single cell, the "copies-per-microliter" value was multiplied by 250 to account for both the 25 μL PCR volume and initial sample volume of 20 μL. In this analysis, cells started off at varying sizes and were not surgically manipulated. In order to correct for variation in reaction loading volumes, each ddPCR reaction was performed with two probe sets, one specific for a given contig, and one specific for contig_558 which was used as a standard. The measured value for contig_558 should be identical across all ddPCR reactions for an individual cell, but in reality this number varied slightly among reactions. To normalize all of the separate ddPCR reactions for a given cell and account for this variation in the standard, the average measured values for contig_558 were used to normalize the measured values for the other contigs, according to the relation (Measured_ploidy / normalized_ploidy) = (Measured_558 / Average_558). Reproducibility of the measurement was assessed by analyzing two different, non-overlapping probe sets on two contigs (contigs 2 and 2227). This was done for eight cells, and the correlation coefficients between the two probes on each contig were 0.99 and 0.99, respectively. Because of the high correlation observed, single probes were used for the analysis reported in the text.

3'UTR Lengths

We performed 3' RACE on a selected group of genes, and Sanger sequenced the RACE products. Next, we took RNAseq reads with polyA tails and mapped them to our set of *Stentor* protein coding regions with Bowtie2. In both cases we found the 3' UTR length by measuring the distance between the stop codon and the polyA tail.

In a given gene, the poly(A) tail often initiated at slightly different sites across different reads. Some poly(A) tails initiated directly after the stop codon, effectively indicating the absence of a 3' UTR. 82% of 3' UTRs examined in RNAseq data were less than 50 bp. Such short UTRs would be too short to encode a SECIS element in the case of a selenoprotein. As *Stentor* does encode a UGA-selenocysteine tRNA, we looked for examples of putative selenoprotein-coding genes. We found 7 homologs of glutathione peroxidase and a homolog of thioredoxin reductase that all appear to contain an in-frame UGA codon (**Supplemental Table S4**). All eight of these genes

possess 3' UTRs in the 83-139 bp range and appear to encode the stem-loop SECIS elements required for selenocysteine incorporation.

Orthology Grouping of Stentor Gene Models

The predicted proteomes of *Stentor coeruleus, Tetrahymena thermophila, Paramecium tetraurelia*, and *Oxytricha trifallax* were analyzed by OrthoMCL (http://www.orthomcl.org/) [S28] for assignment into curated ortholog groups, which uses all-to-all BLASTP searches followed by Markov clustering. While *Tetrahymena* is already present in the OrthoMCL database, this is a previous version of the gene predictions, and we found that the current predicted proteome contained some proteins that didn't match anything in the database. After this initial step, genes from the four ciliates that didn't match any of the curated groups were then pooled together and reanalyzed in OrthoMCL to predict ciliate-specific co-orthologs. *Tetrahymena*-specific ortholog groups from the curated database were also added to the ciliate-specific count.

To find kinase domains, profile HMMs for all the kinase family domains in Kinbase (http://kinbase.com/kinbase/) were downloaded. The *Stentor* predicted proteins were searched with these profiles using HMMER3 (http://hmmer.org) with an e-value cutoff of 0.05. Hits were confirmed with BLASTP against against a custom database of all the kinase domain sequences found in kinbase (http://kinase.com/web/current/).

Supplemental References

S1 Swart, E.C., Serra, V., Petroni, G., and Nowacki, M. (2016). Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. Cell 166:691–702.

S2 Lozupone, C.A., Knight, R.D., and Landweber, L.F. (2001). The molecular basis of nuclear genetic code change in ciliates. Current Biology 11, 65–74.

S3 Farmer, S., Leung, W.-K., and Tsubouchi, H. (2011). Characterization of Meiotic Recombination Initiation Sites Using Pulsed-Field Gel Electrophoresis. DNA Recombination, v745, 33–45.

S4 Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–70.

S5 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, 265–72.

S6 Ruby, J.G., Bellare, P., and DeRisi, J.L. (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda) 3, 865–80.

S7 Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res. 12, 656–64.

S8 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–10.

S9 Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., Lemaitre, C,. and Peterlongo, P. (2014). Reference-free detection of isolated SNPs. Nucleic Acids Res. 43,e11.

S10 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156–8.

S11 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–9.

S12 Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–93.

S13 Li, H. (2011). Improving SNP discovery by base alignment quality. Bioinformatics 27, 1157–8.

S14 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altschuler, D.,Gabriel, S., Daly, M., et. al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297–303.

S15 DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491–8.

S16 Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1

S17 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.

S18 Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44, D286–93.

S19 Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl 2:ii215–25.

S20 Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24, 637–44.

S21 Kim, D., Pertea, G., Trapnell, C., Pimentel, H., and Kelley, R. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36.

S22 Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Segurens, B., Daubin, V, Anthouard, V., Alach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444, 171–8.

S23 Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. 19, 1639–45.

S24 Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35, 3100–8.

S25 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, 955–64.

S26 Aeschlimann, S.H., Jönsson, F., Postberg, J., Stover, N.A., Petera, R.L., Lipps, H.-J., Nowacki, M., and Swart, E.C. (2014). The draft assembly of the radically organized *Stylonychia lemnae* macronuclear genome. Genome Biol. Evol. 6, 1707–23.

S27 Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16, 276–7.

S28 Chen, F., Mackey, A.J., Stoeckert, C.J., and Roos, D.S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34, D363–8.

S29 Wang, R., Xiong, J., Wang, W., Miao, W., and Liang, A. (2016). High frequency of +1 programmed ribosomal frameshifting in *Euplotes octocarinatus*. Sci. Rep. 6, 21139.

S30 Hamilton, E.P., Kapusta, A., Huvos, P.E., Bidwell, S.L., Zafar, N., Tang, H., Hadjithomas, M., Kirshnakumar, V., Badger, J.H., Caler, E.V., et al. (2016). Structure of the somatic germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. Elife 5, e19090

S31 Coyne, R.S., Hannick, L., Shanmugam, D., Hostetler, J.B., Brami, D., Joardar, V.S., Johnson, J., Radune, D., Singh, I., Badger, J.H., et al. (2011). Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. Genome Biol 12, R100.

S32 Elliott, R., Li, F., Dragomir, I., Chua, M.M., Gregory, B.D., and Weiss, S.R. (2013). Analysis of the host transcriptome from demyelinating spinal cord of murine coronavirus-infected mice. PLoS One 8, e75346.