HHS Public Access

Author manuscript

Curr Biol. Author manuscript; available in PMC 2018 February 20.

Published in final edited form as:

Curr Biol. 2017 February 20; 27(4): 569–575. doi:10.1016/j.cub.2016.12.057.

The macronuclear genome of *Stentor coeruleus* reveals tiny introns in a giant cell

Mark M. Slabodnick[1,*], **J. Graham Ruby**[1,*], **Sarah B. Reiff**[1,*], **Estienne C. Swart**[2,*], **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,‡], 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

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* (Figure 1A,B [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 millimeter 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

‡co-corresponding authors: scottwroy@gmail.com. Wallace.marshall@ucsf.edu (Lead Contact). psood1@gmail.com.
*co-first authors with equal contributions

Author Contributions

M.M.S., S.B.R., S.G., B.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.

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*.

eTOC blurb

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.

Results and Discussion

Shotgun sequencing of *Stentor* macronuclear genome

As in all ciliates, the *Stentor* 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 Supplemental Methods). 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 9198 contigs with a total length of 83Mb and a contig N50 of 51kb (Figure 1C). Of these contigs, 29 have telomeres on both ends and 465 have one on only one side (see Supplemental Methods for more details). Using three different approaches, we estimate that the SNP density ranges from 1–4 SNPs per 1500 bases (see Supplemental Methods), suggesting that the genome exhibits low heterozygosity. The genome assembly and all associated raw data have been deposited in Genbank (BioProject PRJNA352242, BioSample SAMN05968724). Additionally, the genome is available online at http:/stentor.ciliate.org. As shown in Figure 1D, coverage is 50–100x 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 CHEF gel (Figure S1 D) and found that the range of sizes is 20kb – 250kb, comparable to the range of sizes of contigs in the assembly (250b–265kb). 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 approximately 92Mb [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 tRNA-ScanSE [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 6-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 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*, or 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 evalue 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 Supplemental Information 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, 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 CEGMA approach 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

the assembly is largely compete. 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 RNAseq data to train the Augustus gene prediction program [14], 34,506 gene models were generated. Of these models, 99% are supported by RNAseq reads and 33% have proteomic support. This gene number is comparable to that seen in other ciliates – for example the *Paramecium* genome encodes approximately 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 percent identity between reciprocal best BLAST hits, as well as their non-synonymous to synonymous rates of substitution (Figure S2D,E), 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 (Table S2). Of the 4,747 curated ortholog groups found in *Stentor*, all but 3 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 (Figures 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 10 orthology groups with the most *Stentor* genes contained kinases, and a sixth group was comprised mostly of protein phosphatase 2C orthologs. Using HMMER3 (hmmer.org) to find kinase domains in the *Stentor* gene models, we found that the *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 extremely short introns. 9,325 introns were predicted in gene models, and of those that we confirmed by Sanger sequencing, 94.5% were 15 nucleotides long, the rest were 16 nucleotides, 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 tiny introns have a long history in these ciliates (Figure S3A).

Whereas previously reported short introns lacked clear internal candidates for branchpoint sites[18–20], *Stentor* introns exhibit a strongly conserved A nucleotide, likely representing the branchpoint, near the 3′ end (6nt upstream for 15nt introns, 6–7nt for 16nt introns; asterisk in Figure 3A), suggesting 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 noncanonical T nucleotide, which is not complementary to the standard U2 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 15nt introns (84.8%) contained an in-frame stop codon (versus only 29.5% of 16nt introns). These stop codons largely reflect the fact that the consensus 15nt sequence contains stop codons in 2 out of 3 possible reading frames (brackets in Figure 3A); the 16nt consensus sequence has both stops in the same frame and thus only has stops in 1 out of 3 possible reading frames. It is thus unclear whether the presence of in-frame stops reflects a selection on stop codons or is simply a byproduct 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 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 16nt introns were somewhat less so (92.4%; $P = 4\times10^{-6}$ 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 frame shift, 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 17–18nt splicing events may represent splicing mistakes, since the 3′ AG lay directly downstream of an AG at the 15nt or 16nt position in (40.0% of cases), 78.8% of which are confirmed splice boundaries (although such cases may also represent functional alternative splicing).

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%; *Paramecium* 20%).

Stentor's 3′UTRs are also small with a median length of 31 nucleotides, 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 Information.

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 S3). Intergenic lengths in *Oxytricha* are shorter than in *Stentor*, because the *Oxytricha* genome is comprised 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 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 Digital droplet PCR of seven different contigs in cells of varying sizes confirmed that *Stentor* is polyploid. For example, the rDNA locus-containing contig (Contig 2227), 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 approximately 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 2227) indicating 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 towards 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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 (MMS), an American Cancer Society postdoctoral fellowship (PS), the Herbert Boyer Junior Faculty Endowed Chair (WFM), a UCSF Resource Allocation Program New Directions grant (WFM), and by NIH grants R01 GM090305 (WFM) and R01 GM113602 (WFM).

References

1. Slabodnick MM, Marshall WF. Stentor coeruleus. Curr Biol. 2014; 24:R783–4. [PubMed: 25202864]
2. Tartar, V. The biology of *Stentor*. New York: Pergammon Press; 1961.
3. Morgan TH. Regeneration of proportionate structures in *Stentor*. The Biol Bull. 1901; 2:311–328.
4. Wood DC. Parametric studies of the response decrement produced by mechanical stimuli in the protozoan, *Stentor coeruleus*. J Neurobiol. 1969; 1:345–60. [PubMed: 5407046]
5. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20:265–72. [PubMed: 20019144]
6. Ruby JG, Bellare P, DeRisi JL. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda). 2013; 3:865–80. [PubMed: 23550143]
7. Pelvat B, De Haller G. Macronuclear DNA in *Stentor coeruleus*: a first approach to its characterization. Genetical Research. 1976; 27:277–89. [PubMed: 819329]
8. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997; 25:955–64. [PubMed: 9023104]
9. Swart EC, Serra V, Petroni G, Nowacki M. Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. Cell. 2016; 166:691–702. [PubMed: 27426948]
10. Lozupone CA, Knight RD, Landweber LF. The molecular basis of nuclear genetic code change in ciliates. Curr Biol. 2001; 11:65–74. [PubMed: 11231122]
11. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007; 23:1061–7. [PubMed: 17332020]
12. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, Khurana JS, Goldman AD, Nowacki M, Schotanus K, et al. The *Oxytricha trifallax* Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes. PLoS Biol. 2013; 11:e1001473. [PubMed: 23382650]
13. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012; 9:357–9. [PubMed: 22388286]
14. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008; 24:637–44. [PubMed: 18218656]
15. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aich N, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 2006; 444:171–8. [PubMed: 17086204]
16. Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, Hadjithomas M, Kirshnakumar V, Badger JH, Caler EV, et al. Structure of the somatic germline genome of Tetrahymena thermophila and relationship to the massively rearranged somatic genome. Elife. 2016; 5:e19090. [PubMed: 27892853]
17. Coyne RS, Hannick L, Shanmugam D, Hostetler JB, Brami D, Joardar VS, Johnson J, Radune D, Singh I, Badger JH, et al. 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. 2011; 12:R100. [PubMed: 22004680]
18. Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc Natl Acad Sci USA. 2006; 103:9566–71. [PubMed: 16760254]

19. Russell CB, Fraga D, Hinrichsen RD. Extremely short 20–33 nucleotide introns are the standard length in *Paramecium tetraurelia*. Nucleic Acids Res. 1994; 22:1221–5. [PubMed: 8165136]
20. Ogino K, Tsuneki K, Furuya H. Unique genome of dicyemid mesozoan: highly shortened spliceosomal introns in conservative exon/intron structure. Gene. 2010; 449:70–6. [PubMed: 19747532]
21. Irimia M, Roy SW. Evolutionary convergence on highly-conserved 3′ intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. PLoS Genet. 2008; 4:e1000148. [PubMed: 18688272]
22. Lee RCH, Gill EE, Roy SW, Fast NM. Constrained intron structures in a microsporidian. Mol Biol Evol. 2010; 27:1979–82. [PubMed: 20360213]
23. Vanacova S, Yan W, Carlton JM, Johnson PJ. Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. Proc Natl Acad Sci USA. 2005; 102:4430–5. [PubMed: 15764705]
24. Nixon JEJ, Wang A, Morrison HG, McArthur AG, Sogin ML, Loftus BJ, Samuelson J. A spliceosomal intron in *Giardia lamblia*. Proc Natl Acad Sci USA. 2002; 99:3701–5. [PubMed: 11854456]
25. Xu F, Jerlström-Hultqvist J, Einarsson E, Astvaldsson A, Svärd SG, Andersson JO. The genome of *Spironucleus salmonicida* highlights a fish pathogen adapted to fluctuating environments. PLoS Genet. 2014; 10:e1004053. [PubMed: 24516394]
26. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, et al. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 2006; 4:e286. [PubMed: 16933976]
27. Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, Saudemont B, Nowacki M, Serrano V, Porcel BM, Segurens B, et al. Translational control of intron splicing in eukaryotes. Nature. 2008; 451:359–62. [PubMed: 18202663]
28. Price HJ, Sparrow AH, Nauman AF. Correlations between nuclear volume, cell volume and DNA content in meristematic cells of herbaceous angiosperms. Experientia. 1973; 29:1028–9.
29. Olmo E. Nucleotype and cell size in vertebrates: a review. Basic Appl Histochem. 1983; 27:227–56. [PubMed: 6360135]
30. Shuter BJ, Thomas JE, Taylor WD, Zimmerman AM. Phenotypic Correlates of Genomic DNA Content in Unicellular Eukaryotes and Other Cells. American Naturalist. 1983; 122:26–44.
31. Mueller RL. Genome Biology and the Evolution of Cell-Size Diversity. Cold Spring Harb Perspect Biol. 2015; 7:a019125. [PubMed: 26254312]
32. Winkelmann M, Pfitzer P, Schneider W. Significance of polyploidy in megakaryocytes and other cells in health and tumor disease. Klin Wochenschr. 1987; 65:1115–31. [PubMed: 3323647]
33. Biesterfeld S, Gerres K, Fischer-Wein G, Böcking A. Polyploidy in non-neoplastic tissues. J Clin Pathol. 1994; 47:38–42. [PubMed: 8132807]
34. Anatskaya OV, Vinogradov AE. Somatic polyploidy promotes cell function under stress and energy depletion: evidence from tissue-specific mammal transcriptome. Funct Integr Genomics. 2010; 10:433–46. [PubMed: 20625914]
35. Gillooly JF, Hein A, Damiani R. Nuclear DNA Content Varies with Cell Size across Human Cell Types. Cold Spring Harb Perspect Biol. 2015; 7:a019091. [PubMed: 26134319]
36. Losick VP, Fox DT, Spradling AC. Polyploidization and Cell Fusion Contribute to Wound Healing in the Adult *Drosophila* Epithelium. Curr Biol. 2013; 23:2224–32. [PubMed: 24184101]
37. Yao MC, Kimmel AR, Gorovsky MA. A small number of cistrons for ribosomal RNA in the germinal nucleus of a eukaryote, *Tetrahymena pyriformis*. Proc Natl Acad Sci USA. 1974; 71:3082–6. [PubMed: 4606151]
38. de Terra N. Macronuclear DNA synthesis in *Stentor*: Regulation by a cytoplasmic initiator. Proc Natl Acad Sci USA. 1967; 57:607–14. [PubMed: 16591507]

Highlights

- The introns of *Stentor coeruleus*, a giant ciliate, are 15–16 nucleotides 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.

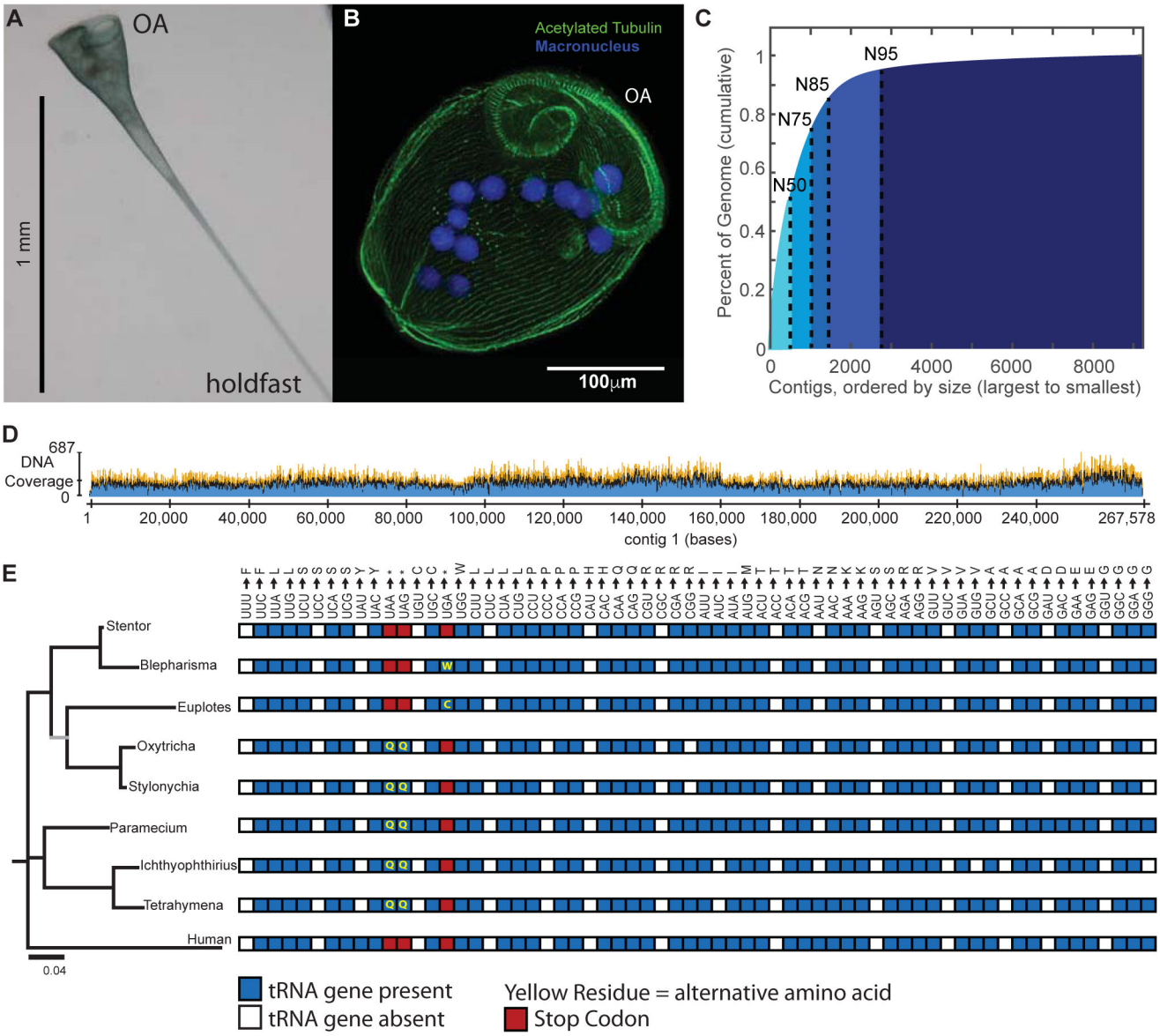

Figure 1. Shotgun sequencing the *Stentor coeruleus* macronuclear genome

(**A**) Brightfield 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 are indicated by OA. (**C**) Cumulative distribution depicting the N50 (50kb) 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**) Phylogenetic comparison of 18S RNA (left) 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; *Paramecium*, *Tetrahymena* and *Ichthyophthirius* represent class Oligohymenophorea. See also Figure S1 and Table S1.

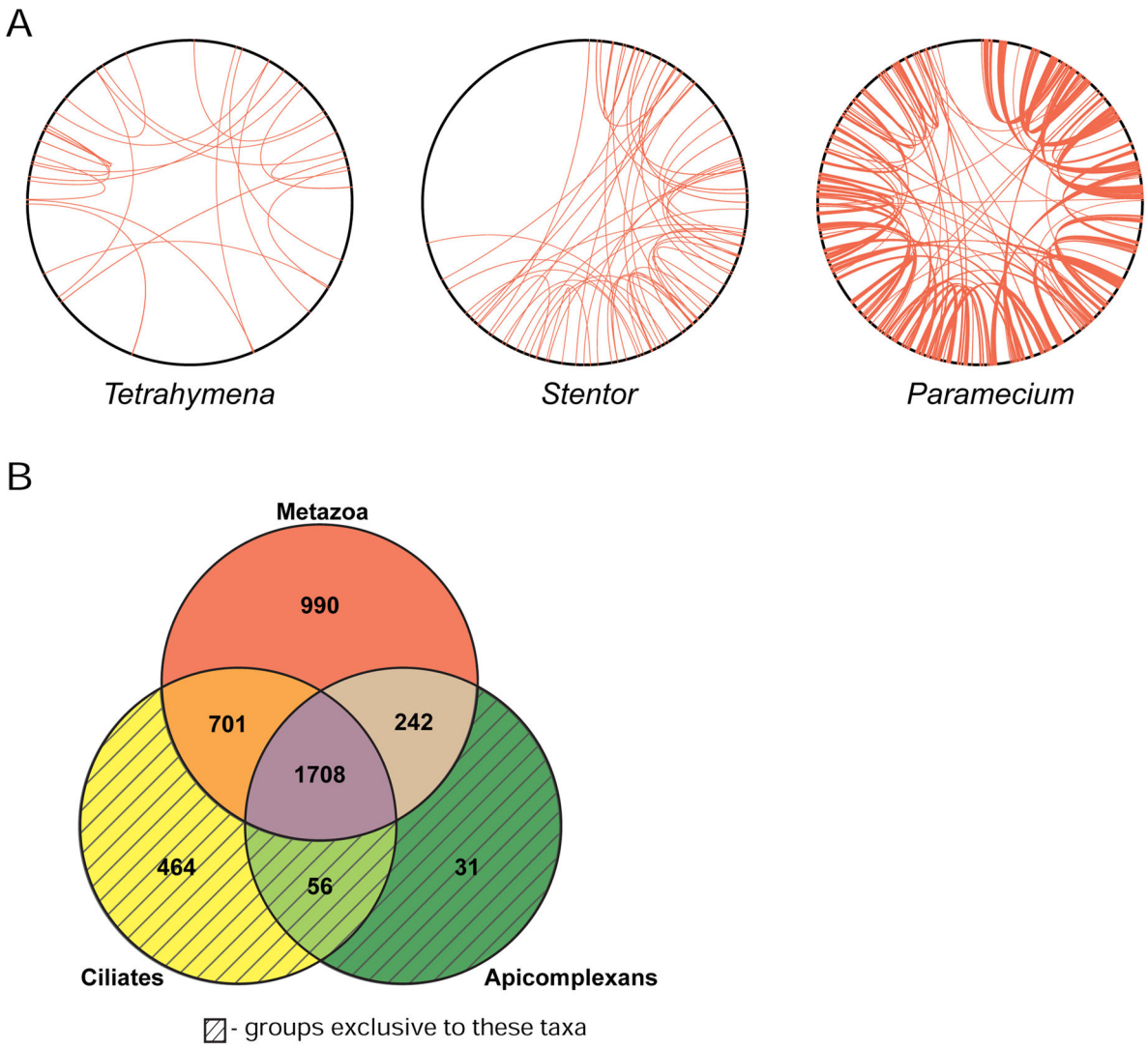

Figure 2. *Stentor* gene duplications and orthology groups

(**A**) Genome duplication events in the genomes of *Stentor coeruleus*, *Paramecium tetraurelia,* and *Tetrahymena thermophila.* To generate coordinates on the perimeter of each circle, contigs/scaffolds were arranged from longest to shortest and then continuously numbered from 1 to the end of the assemblies. Red lines connect paralogous windows (see methods) 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 not pictured in this diagram. See also Figure S2 and Table S2.

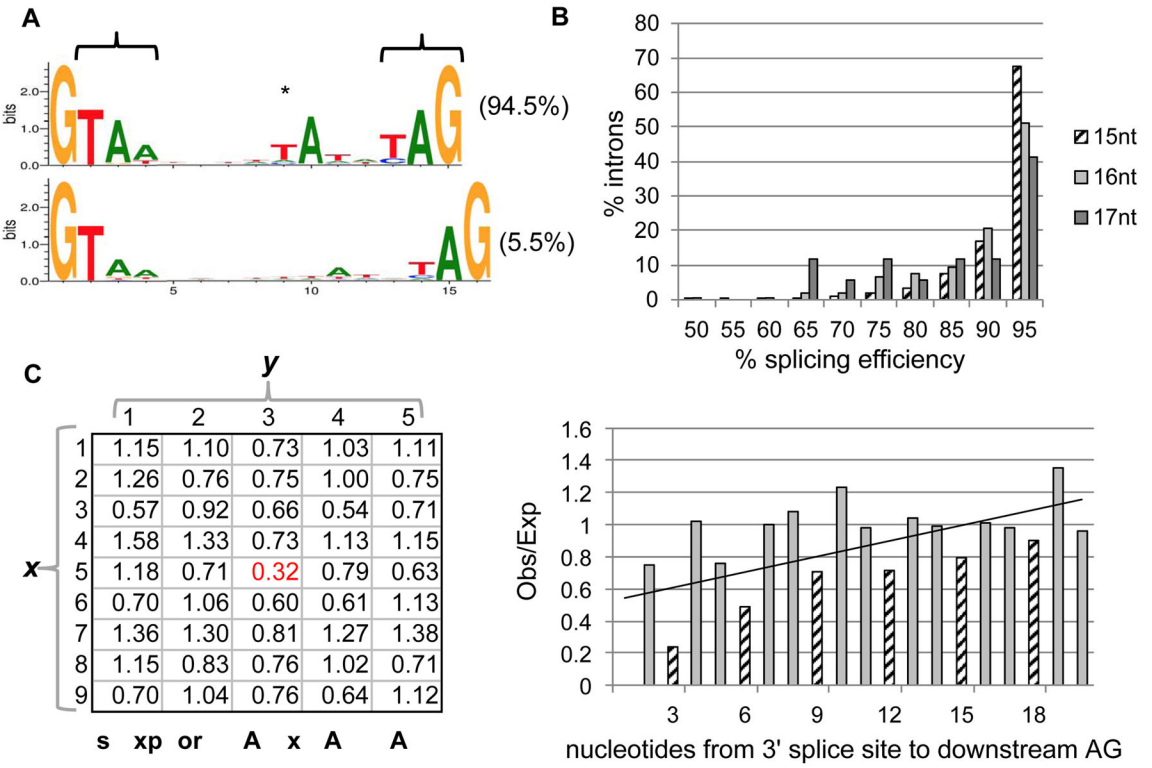

Figure 3. Intron sequences and splicing in *Stentor*

(**A**) Nearly all identified introns in *Stentor* are 15nts (94.5%, top) or 16nts (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 15nt introns. Graph shows a histogram of the distribution of introns in each size class (15–17nts) 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. 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 branchpoint and y indicates the number of bases following the branchpoint 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 trendline is a linear fit to all data shown. See also Figure S3 and Tables S3 and S5.

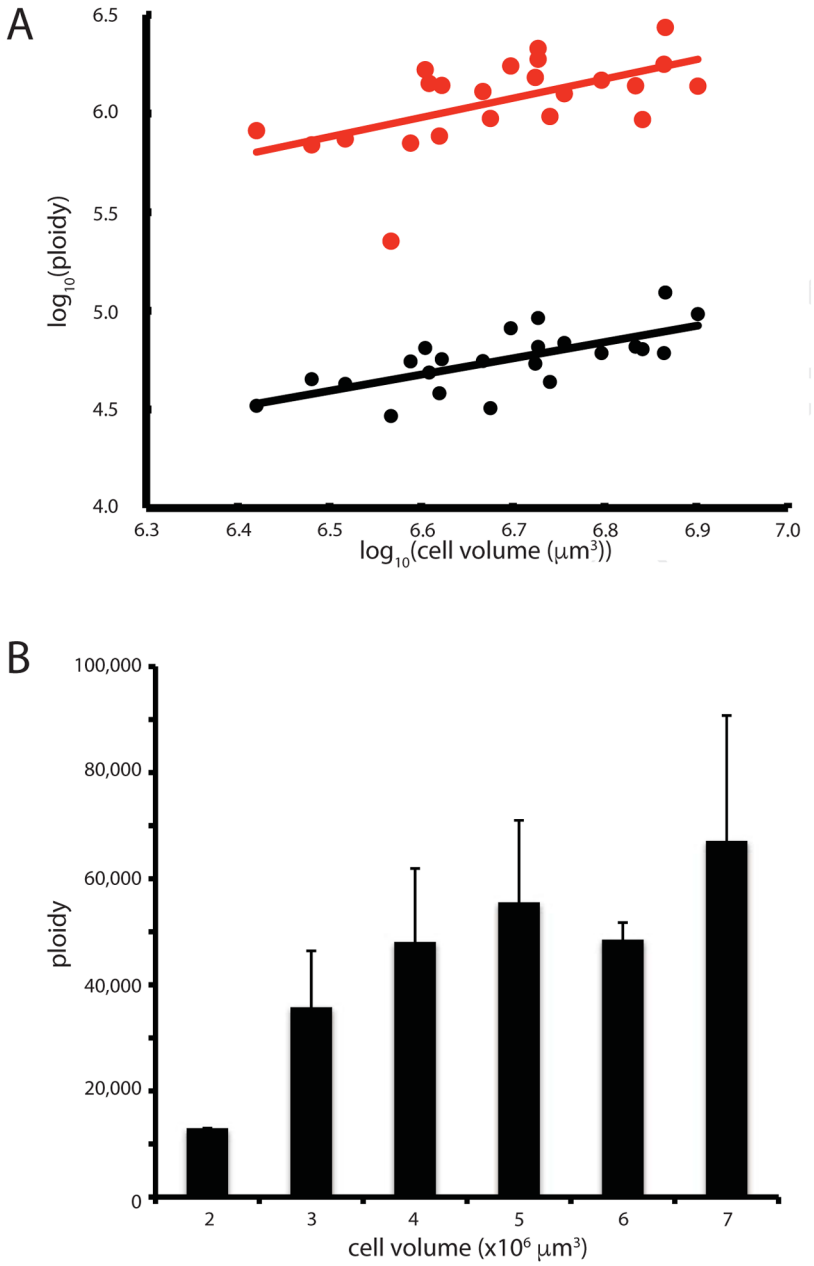

Figure 4. Macronuclear ploidy scales with cell volume

(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. (Red) copy number of rDNA-containing contig and (Black) a large contig that does not contain rDNA. Each point represents a single cell. Ploidy data used two different y axis scales because the average ploidy is approximately 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 standard deviation. See also Table S4.