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

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

1Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143, USA
2Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland
3Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
4Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
5Department of Ob/Gyn, University of California, San Francisco, San Francisco, CA 94143, USA
6Department of Biology, San Francisco State University, San Francisco, CA 94132, USA
7Department of Biology, Bradley University, Peoria, IL 61625, USA
8Co-first author
9Lead 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 *Paramaecium*. 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 macronucleus 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

**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**
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 \( \sim 92 \text{ Mb} \) \([7]\), 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 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 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.

<table>
<thead>
<tr>
<th>ORFs Translated with</th>
<th>Peptide Support for Alternative Codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Code</td>
<td>0.04%</td>
</tr>
<tr>
<td>Primary Ciliate Code</td>
<td>0.02%</td>
</tr>
<tr>
<td><em>Blepharisma</em> Code</td>
<td>0.00%</td>
</tr>
<tr>
<td>UAR-glutamate Code</td>
<td>0.07%</td>
</tr>
</tbody>
</table>

We also found that 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.
suggested 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 216,022 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
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 of the introns, 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.

**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.

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 × 10^{-6} by randomization; Figure 3B). Several features suggest avoidance of off-target splicing may shape the transcriptome. First, within unspliced genic regions.

![Figure 3. Intron Sequences and Splicing in Stentor](image)

(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(1–9)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.
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]. Even 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/10.17632/37gp2djcost.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

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.).

REFERENCES

