Combined genome and transcriptome analyses of the ciliate *Schmidingerella arcuata* (Spirotrichea) reveal patterns of DNA elimination, scrambling, and inversion

Susan A. Smith[1]*, Xyrus X. Maurer-Alcalá[2], Ying Yan[3], Laura A. Katz[3], Luciana F. Santoferrara[1,4], George B. McManus[1]
[1]Department of Marine Sciences, University of Connecticut, Groton, USA
[2]Institute of Cell Biology, University of Bern, Bern, Switzerland
[3]Smith College, Department of Biological Sciences, Northampton, Massachusetts, USA
[4]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, USA
*Author for Correspondence: Department of Marine Sciences, University of Connecticut, Groton, USA, susan.smith@uconn.edu

Abstract

Schmidingerella arcuata is an ecologically important tintinnid ciliate that has long-served as a model species in plankton trophic ecology. We present a partial micronuclear genome and macronuclear transcriptome resource for *S. arcuata*, acquired using single-cell techniques, and we report on pilot analyses including functional annotation and genome architecture. Our analysis shows major fragmentation, elimination, and scrambling in the micronuclear genome of *S. arcuata*. This work introduces a new non-model genome resource for the study of ciliate ecology and genomic biology, and provides a detailed functional counterpart to ecological research on *S. arcuata.*

Keywords: genome architecture, single-cell 'omics, tintinnid, micronucleus, macronucleus, Ciliophora

Significance

Our understanding of genome organization in non-model ciliates is limited because (1) most species are uncultivable and (2) it requires the separate amplification of a ciliate's two nuclei (the germline micronucleus and the somatic macronucleus), which is technically difficult. By using single-cell 'omics, we were able to separately sequence both genomes of the ciliate *Schmidingerella arcuata*, which revealed patterns of extensive genome rearrangement and fragmentation, and also allowed us to analyze functional details of its somatic genome. This research contributes information on the genomic architecture and functional of an ecologically important ciliate, and expands our understanding of ciliate genomics beyond model species.

Introduction

Ciliates are an ancient and diverse clade of microbial eukaryotes that inhabit nearly every environment on Earth. They have long-served as models in the research of cell biology, and recently have become an ideal system for the study of genome fragmentation and organization (Greider and Blackburn 1985; Parfrey et al. 2011; Lynn 2008, Pederson 2010). Although genome rearrangements have been discovered throughout the eukaryotes, the phenomenon appears to be especially elaborate within ciliates, which can exhibit the mass elimination, fragmentation, and scrambling of loci (Prescott 2000; Bracht et al. 2013; Gao et al. 2015; Jahn and Klobutcher 2002; Chen et al. 2014; Maurer-Alcalá et al. 2018a; Chalker 2008; Landweber et al. 2000).

The complexity of ciliate genome architecture results from their nuclear dimorphism, a unique segregation of germline and somatic functions into separate nuclei (Lynn 2008). Chromosomal rearrangements can occur between the germline-limited micronucleus (MIC) and the transcriptionally active macronucleus (MAC). Following conjugation, the zygotic nucleus divides to form a new MIC and a new MAC. In creating the new MAC, a series of rearrangements and deletions can occur, including the extensive elimination of MIC segments called IESs (Internal eliminated Sequences) (Fass et al. 2011; Riley and Katz 2001; Gratias and Bétermier 2001; Katz et al. 2003; Zufall et al. 2005). Regions that are not eliminated (i.e. present in both MIC and MAC) are called MDSs (Macronucleus Destined Sequences), and are cut and pieced together to form MAC loci (Swart et al. 2013) (Figure 1A). In some ciliates, MDS regions are arranged out of order in the MIC, a complex pattern of genomic architecture called "scrambling". Scrambled loci can additionally be "inverted" if they are transcribed on opposing strands of the MIC scaffolds (Figure 1A).

As most ciliates are not amenable to culture, research on ciliate genomics and nuclear architecture has mostly been limited to a few model species (e.g. *Paramecium*, *Tetrahymena*) (Wang et al. 2017; Hamilton et al. 2016). In addition, traditional sequencing methods are biased toward highly-amplified MAC regions, which has made it challenging to isolate MIC regions. However, the process of multiple-displacement amplification (MDA) used in single-cell genomics is biochemically biased for long-template DNA (2-70kb), which allows for the selection of MIC chromosomes (Spits et al. 2006).

Combined with single-cell transcriptomics, we are now are able to elucidate the patterns of genome rearrangement, elimination, and scrambling, all from a single cell (Maurer-Alcalá et al. 2018a).

Here we use single-cell 'omics (genomics and transcriptomics) to study the MIC genome and the transcriptome (a proxy for the gene-sized chromosomes of the macronuclear genome) of *Schmidingerella arcuata,* a marine ciliate (class Spirotrichea, order Tintinnida). Long-used as an ecological model in planktonic food web studies, *S. arcuata* is ubiquitous in coastal waters, where it periodically dominates the ciliate community (Dolan and Pierce 2013; Santoferrara et al. 2018; Agatha and Strüder-Kypke 2012; Echevarria et al. 2014). *S. arcuata* is also one of the few marine ciliates amenable to culture and thus represents a ciliate that is both ecologically relevant and cultivable (Echevarria et al. 2016; Montagnes et al. 2008, 2013; Dolan 2012; Jung et al. 2016; Cobb 2017; Gruber et al. 2019). Although tintinnids have a long history of taxonomic study (Müller 1779; Haeckel 1866), there exists no published data on their MIC or genomic architecture, and only limited transcriptome data exist for *Schmidingerella* (Keeling et al. 2014), the only tintinnid genus with transcriptome data. Here, we present a genome and transcriptome resource for *S. arcuata,* acquired using single-cell techniques, and we report on pilot analyses of its genome architecture and transcriptome. This represents a new resource for the study of ciliate genomic architecture, and provides a detailed genomic counterpart to ecological research on this model microzooplankton.

Materials and Methods

Culturing

Schmidingerella arcuata was collected from the surface waters of northeastern Long Island Sound, CT (41.31˚ N, 72.06˚ W), using a 20μm-mesh plankton net. Single cells were isolated with drawn capillaries and moved to 6-well culture plates with 0.2μm-filtered sample water. Clonal cultures of *S. arcuata* were fed saturating concentrations (c. $3x10^3$ cells/ml) of the dinoflagellate *Heterocapsa triquetra* and the prymnesiophyte *Isochrysis galbana* (strain TISO). Cultures were kept at 18°C under a 12:12 h light:dark

cycle. Morphology and 18S rDNA sequences (see Santoferrara et al. 2013) confirmed the taxonomic identification of *S. arcuata* (Agatha and Strüder Kypke, 2012).

Isolation of Single Cells

Individuals were transferred from growing cultures to autoclaved, 0.2 μm-filtered seawater and starved for twelve hours to ensure the clearance and digestion of prey. The cells were then picked and rinsed a minimum of five times in autoclaved, 0.2 μm-filtered seawater using drawn capillaries under a stereo microscope. Each cell was then transferred into the appropriate buffer for transcriptome or genome sequencing and brought to volume with nuclease-free water (as specified in the kits detailed below).

Single Cell Transcriptome and Genome Amplification

The SMART-Seq2 v4 Ultra Low input RNA kit (Cat: 634889; Takara, Mountain View, CA) was used for WTA (whole transcriptome amplification) following manufacturer's protocols, with the exception that we quartered the reaction volumes. For WGA (whole genome amplification), the Repli-g single-cell kit (Cat: 150343; Qiagen, Hilden, Germany) was used following manufacturer's protocols. The products (cDNA for WTA, gDNA for WGA) were quantified with the dsDNA Qubit assay (Invitrogen, Waltham, MA) and PCR-checked with eukaryotic 18S rDNA (Medlin et al. 1988) and genus-specific ITS (Costas et al. 2007) primers. Minimum bacterial contamination was confirmed by PCR with 16S rDNA primers (Lane et al. 1991). Sequencing libraries were prepared with the Illumina Nextera XT kit (Cat: FC1311096; Illumina, San Diego, CA) then processed with Illumina HiSeq 2500 at Macrogen Sequencing (Geumcheon-gu, Seoul, South Korea).

Transcriptome and Genome Assembly

Raw reads from WTA and WGA sequencing were trimmed for quality and size (Q28 and minimum length of 200bp and 1500bp, respectively) using BBMap (V38.39; Bushnell 2014) After trimming, two single-cell WTAs were assembled together using rnaSPAdes (V3.13.1; Bankevich et al. 2012), and seven singe-cell WGAs were assembed together using both SPAdes (V3.13.1) and MEGAHIT (V1.2.9; Li et al. 2015) .The MEGAHIT genome assembly was used for the final analysis because it yielded a higher mapping continuity (i.e. the amount of transcripts mapped to the WGA assembly per kilobase).

Assemblies were processed through custom python scripts (http://github.com/maurerax/KatzLab/tree/HTS-Processing-PhyloGenPipeline) for the removal of rDNA and prokaryotic transcripts, and for the identification of orthologous gene (OG) families using USEARCH (V9.2; Edgar 2010) with OrthoMCLdatabases (V2.0.9; Fischer et al. 2011) (Maurer-Alcalá et al. 2018a). Additional steps included the prediction of open reading frames with AUGUSTUS (Hoff and Stanke 2019) using an *E.coli* model to eliminate bacterial contaminants. Stop codon usage was determined using a custom Python script, which quantified the frequency of in-frame occurences of TAG/TGA/TAA when each codon was used as a termination site. The completeness of the MIC genome assembly was analyzed using BUSCO (Benchmarking Universal Single-Copy Ortholog; Waterhouse et al. 2018) (E-value $<10^{-3}$, alveolate lineage database). OmicsBox (V5.2.5; Götz et al. 2008) was used with InterProScan (V5.42; Hunter et al. 2009) and BLASTx (NCBI non-redundant database, E-value $<10^{-4}$; V2.8.1; Altschul et al. 1990) for functional annotation and for the identification of GO terms involved in KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto 2000) pathways.

Genome Architecture Analysis

Custom Python scripts were used to identify and organize the genome architecture of *S. arcuata* (Maurer-Alcalá et al. 2018a). Putative MIC loci for MAC MDSs were identified via mapping of the MAC transcriptome to the MIC genome sequences using BLAST. For transcripts to be considered "MIC-mapped", at least 60% of their length was required to be mapped to the MIC (length threshold as suggested in Maurer-Alcalá et al. 2018a). Loci were classified into categories of "unmapped," "nonscrambled", and "scrambled". MDS-IES borders were required to be flanked by pairs of short (2-10bp) tandem repeats called pointer regions to discriminate them from possible intron-exon boundaries (Bracht et al. 2013) (Figure 1B, C). The GC content of MDS-IES boundaries was determined by evaluating the 40bp located at both ends (i.e. the 5' and 3') of an MDS in the MIC.

Results and Discussion

Genome and Transcriptome Resources

Assemblies for the MIC and MAC of *S. arcuata* are about 49Mbp and 6Mbp in size, respectively (Table 1). Of the 11,673 transcripts, which are a proxy for the gene-sized MAC chromosomes of this species, roughly 15% (1,712) were mapped to the MIC. BUSCO analyses estimate that the MIC genome resource is about 19% complete (Complete:18.8%, Fragmented:7.6%, Missing:73.6%, n:171). This indicates the need for deeper sequencing, although BUSCO analyses have been found to underestimate highly-fragmented genomes (López-Escardó et al. 2017) and thus may not be a reliable indicator of completeness in *S. arcuata*. About 80% of MAC transcripts have significant BLASTx hits (NCBI non-redundant database, E-value $<10^{-4}$), and of those, 74% have a confident assignment of GO terms. The majority of MAC transcripts related to cellular components correspond to membrane and organelle activity, while catalytic and binding activity are the primary molecular functions, and localization and biological regulation comprise the main biological terms (Supplementary Figure S1). About 15% of MAC transcripts were placed in KEGG pathways. The three primary pathways identified are thiamine metabolism, purine metabolism, and Aminoacyl-tRNA biosynthesis (Supplementary Figure S1). Annotation details regarding the MAC transcriptome can be found in Supplementary Data, and full annotation files can be found at figshare at the link: https://doi.org/10.6084/m9.figshare.12686621.

We assessed stop codon usage and found that the codons TGA and TAA were rarely found in-frame, and their usage in *S. arcuata* matched homologs in the *Oxytricha trifallax* transcriptome (GenBank BioSample SAMN02953822). The combination of TGA/TAA as stop codons is not observed in the few published transcriptomes for this genus (Keeling et al. 2014; Heaphy 2018) or for other Spirotrichs, with TAA/TAG reported for *E. crassus* and TGA for *O. trifallax* and *S. lemnae* (Lozupone et al. 2001; Kervestin et al. 2001; Yan et al. 2019; Heaphy 2018; Swart et al. 2013). However, ciliates have frequent stop codon reassignments, and even context-dependent stop codons (Yan et al. 2019; Heaphy et al. 2016; Adachi and Calvacanti 2009).

Patterns of Genome Architecture

Schmidingerella arcuata shows extensive genome fragmentation, including the unscrambling and inversion of loci during MAC formation (Figure 1B). We considered MIC loci as scrambled if their

associated transcripts mapped to MDSs that were out of consecutive order in the MIC, and if those MDS-IES boundaries contained pointer regions (Maurer-Alcalá et al. 2018a). Scrambled loci were determined to be inverted if non-consecutive MDSs appeared on both strands of germline scaffolds (Figure 1A, B). Of the MIC-mapped transcripts, roughly 36% were found to be scrambled. The average GC content at MDS-IES boundaries for *S. arcuata* (47.5%) was slightly higher than the overall GC for all germline-supported scaffolds (46.4%); this increase also occurs in other ciliate classes, although most report a sharper rise (10-14%) in %GC around these regions (Maurer-Alcalá et al. 2018a). Pointer sequence size was variable within and among MDS groups (Figure 1B), with a range of 2-10bp. We found no evidence for alternative processing (more than one MAC sequence resulting from a single MIC region; Katz and Kovner 2010).

Variations in Genome Architecture Within and Among Ciliate Classes

In general, the micronuclear arrangement of housekeeping genes in *S. arcuata* matched that of *Oxytricha trifallax.* However, we detected a major housekeeping gene with variable micronuclear organization among six ciliates with available data (Figure 1C). In this example, the beta-tubulin gene (paralog 1) is separated into two similar-sized MDSs in the MIC of *S. arcuata*, interrupted by a single IES and connected by an 8bp TC-iterative pointer sequence (Figure 1C). In contrast, other ciliates of the class Spirotrichea (O. *trifallax, S. lemnae*) separate the paralog into three or four MDSs, with comparatively shorter IES regions. Model ciliates of the Oligohymenophorea class (*P.caudatum* and *T. thermophila*) contain this paralog as either three MDS regions of variable size, or as an uninterrupted sequence in the MIC. The Phyllopharyngean ciliate *C. unicinata* divides the MIC gene into three consecutive MDSs of variable size (75 to 600 bp each) with two 6-7bp pointer sequences (Zufall and Katz 2007; Katz and Kovner 2010; Harper and Jahn 1989; Zufall et al. 2005).

Significance of *S. arcuata* -Omics Resources

This work contributes a MIC genome and MAC transcriptome resource for the ecologically-important ciliate *S. arcuata*. Single-cell omics allowed selective amplification the MIC and MAC, which revealed genomic scrambling, elimination, and inversion in *S. arcuata*. This study provides a non-model genome

and transcriptome resource to a field represented mostly by model ciliates. The included annotation details are a valuable resource for future ecological research on *S. arcuata* and closely-related ciliates, which are currently underrepresented in detailed genome-scale analyses. Additionally, research on non-model ciliates are beginning to reveal the significance of genome architecture in molecular evolution (Yan et al. 2019; Maurer-Alcalá et al. 2018b; Maurer-Alcalá and Nowacki 2019). Recent models indicate that slight differences in IESs and specific architectural patterns (e.g. alternative processing and scrambling) among intraspecific ciliate populations can cause rapid incompatibility, potentially leading to incipient speciation (Gao et al. 2015; Yan et al. 2019; Katz and Kovner 2010; Goldman and Landweber 2012). These slight errors in the rearrangement of loci could theoretically accumulate more frequently than (non-neutral) point mutations, which may help to explain the large disparity between the molecular and morphological diversity in ciliates (Gao et al. 2015).

Data deposition: Data has been deposited at NCBI under the Bioproject ID:PRJNA626068; under the accession numbers JABUIQ000000000 (MIC genome), GISN00000000 (MAC transcriptome), and SRR11933493 (raw reads); transcriptome annotation is available on figshare at the link: https://doi.org/10.6084/m9.figshare.12686621.

Acknowledgements:

This research was supported by the U.S. National Science Foundation (awards to G.B.M. and L.F.S.: OCE1924527; L.A.K.: OCE-1924570 and DEB-1541511), Smith College, and the University of Connecticut's Center for Genome Innovation and Computational Biology Core.

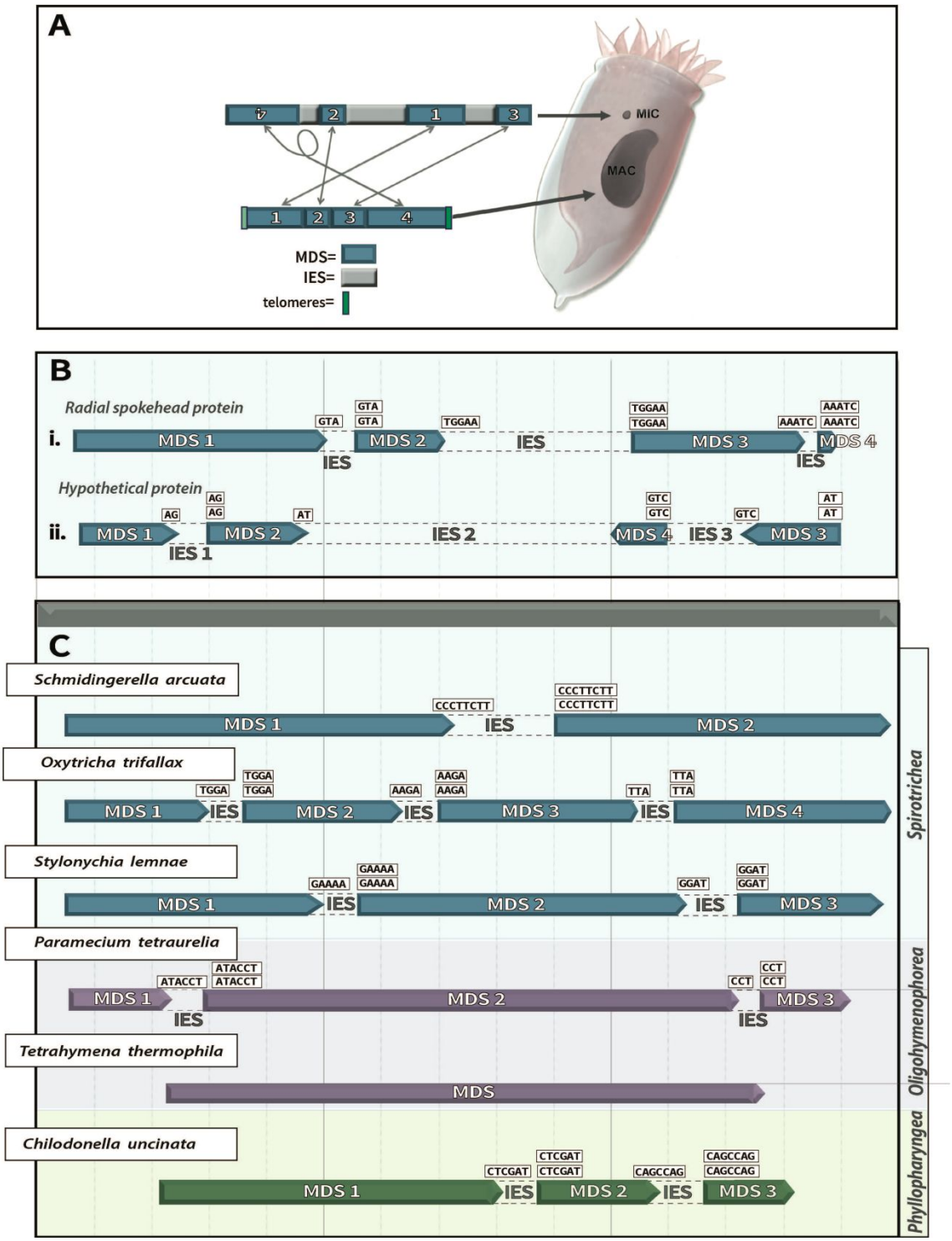

Figure 1: Genome architecture in S. *arcuata.* A: Patterns of nuclear architecture in the germline micronucleus (MIC) and somatic macronucleus (MAC). The MIC contains the required

macronuclear-destined sequences (MDSs) for the generation of functional genes during development of the MAC. MDSs can be interrupted by internally-eliminated sequences (IESs). Development of the MAC requires the precise excision of IESs and the correct rearrangement of MDS regions. MDS loci may be scrambled (e.g. MDS 1-4), inverted (e.g. MDS 4), or a combination of both. The organization of these MDSs is guided by pointer sequences (2-10bp) that occur at MDS/IES boundaries. Green blocks capping ends indicate telomeres, which are added *de novo* to the ends of MAC chromosomes. **B:** Exemplar micronuclear patterns of loci elimination, scrambling and inversion identified in *S. arcuata*. **i.** Consecutive MDS regions of varying size separated by IESs and guided by 3-5bp pointers. **ii.** MDSs separated by IESs of variable lengths; a mix of scrambled, non-scrambled, and partially-inverted loci, with short pointers (2-3bp). Pointer sequences are shown in white blocks; those appearing twice indicate their secondary location in the MIC (pointers occur twice in MIC and once in MAC). MDSs are numbered according to their somatic order in the MAC. Arrows at the end of each MDS indicate MDS directionality in the MIC. **C.** Micronuclear architecture of a beta-tubulin gene in various ciliates. *S. arcuata* separates the gene region into two MDSs, interrupted by a single IES and guided by an 8bp pointer region. Different colors of MDS correspond to different classes, indicated at right. Accession numbers or gene identifiers: *O. trifallax* (PRJNA194431; OxyDB: Contig11167.0.g9); *S. lemnae* (X06874.1); *T. thermophila* (L01416.1); *P.caudatum* (AB070222.1); *C. uncinata* (MH388464).

Table 1: Summary data on micronuclear (MIC) and macronuclear (MAC) characteristics of *Schmidingerella arcuata.* MDS = macronuclear destined sequence; IES = internally eliminated sequence.

Feature	
Size of MIC assembly (Mbp)	**48.6**
Size of MAC assembly (Mbp)	**6.3**
Number of MAC transcripts	**11,673**
Number of MIC-mapped transcripts	**1,718**
Percentage of MIC covered by MAC	**14.6**
Number of scrambled transcripts	**616**
Percentage of MIC genome that contains scrambled transcripts	**35.8**
Average %GC content for all MIC-supported scaffolds	**46.4**
Average %GC content at MDS-IES boundaries	**47.5**
Average pointer length (bp)	**3.7**
Average %GC content of pointers	**40.8**
Average length of scrambled MDSs (bp)	**361.7**
Stop-codon usage	**TGA/TAA**

Literature Cited

Adachi M, Cavalcanti AR. 2009. Tandem stop codons in ciliates that reassign stop codons. J Mol Evo. 68(4):424-431.

Agatha S, Strüder-Kypke MC. 2012. Reconciling cladistic and genetic analyses in choreotrichid ciliates (Ciliophora, Spirotricha, Oligotrichea). J Eukaryot Microbiol 59:325-350.

Altschul SF, et al. 1990. Basic local alignment search tool. J Mol Bio. 215(3): 403-410.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 19(5): 455-477.

Bracht JR, et al. 2013. Genomes on the edge: programmed genome instability in ciliates. Cell. 152:406-416.

Bushnell B. 2014. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab. (LBNL). OSTI.

Chalker DL. 2008. Dynamic nuclear reorganization during genome remodeling of *Tetrahymena*. BBA-Mol Cell Res. 1783(11):2130-2136.

Chen X, et al. 2014. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell. 158:1187–1198.

Cobb S. 2017. Investigation of contact-based cues mediating food uptake in the marine tintinnid ciliate *Favella sp.* Masters Sci. California State University. Chico, CA.

Costas BA, McManus G, Doherty M, Katz LA. 2007. Use of species-specific primers and PCR to measure the distributions of planktonic ciliates in coastal waters. Limnol Oceanogr-Meth, 5(6):163-173.

Dolan JR, Pierce RW. 2013. Diversity and distributions of tintinnids. In: Dolan JR, Agatha S, Coats DW, Montagnes DJS, Stoecker DK. *Biology and Ecology of Tintinnid Ciliates: Models for Marine Plankton*, Oxford: Wiley-Blackwell. p. 214-243.

Dolan JR. 2012. Tintinnid Ciliates: an Introduction and Overview. In: Dolan JR, Agatha S, Coats DW, Montagnes DJS, Stoecker DK. *Biology and Ecology of Tintinnid Ciliates: Models for Marine Plankton*, Oxford: Wiley-Blackwell. p 1-16.

Dupuis P. 1992. The beta-tubulin genes of *Paramecium* are interrupted by two 27 bp introns. EMBO J. 11(10): 3713-3719.

Echevarria ML, Wolfe G, Strom S, Taylor A. 2014. Connecting alveolate cell biology with trophic ecology in the marine plankton using the ciliate *Favella* as a model. FEMS Microb Ecol. 90:18-38.

Echevarria ML, Wolfe GV, Taylor AR. 2016. Feast or flee: bioelectrical regulation of feeding and predator evasion behaviors in the planktonic alveolate *Favella sp.* (Spirotrichia). J Exp Biol. 219(3): 445-456.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 26(19): 2460-2461.

Hoff KJ, Stanke M. 2019. Predicting genes in single genomes with AUGUSTUS. Current Protocols in Bioinformatics. 65(1): 57.

Fass JN, et al. 2011. Genome-scale analysis of programmed DNA elimination sites in *Tetrahymena thermophila*. G3-Genes Genom Genet. 1(6):515-22.

Fischer S, et al. 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Current protocols in bioinformatics. 35(1):6-12.

Gao F, Roy SW, Katz LA. 2015. Analyses of alternatively processed genes in ciliates provide insights into the origins of scrambled genomes and may provide a mechanism for speciation. MBio. 2015 Feb 27. 6(1).

Goldman AD, Landweber LF. 2012. *Oxytricha* as a modern analog of ancient genome evolution. Trends Genet. 28(8):382-388.

Götz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids res, 36(10), 3420-3435.

Gratias A, Bétermier M. 2001. Developmentally programmed excision of internal DNA sequences in *Paramecium aurelia*. Biochimie 83.11:1009-1022.
Greider, CW, Blackburn EH. 1985. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. Cell. 43.2:405-413.
Gruber MS, Mühlthaler A, Agatha S. 2019. Ultrastructural studies on a model tintinnid–*Schmidingerella meunieri* (Kofoid and Campbell, 1929) Agatha and Strüder-Kypke, 2012 (Ciliophora). I. Somatic kinetids with unique ultrastructure. Acta protozool. 57(3):195.
Haeckel E. 1866. VII. Character des Protistenreiches. (VII. Character of the kingdom of Protists.) In: Haeckel E, Reimer G. *Prinzipien der generellen Morphologie der Organismen: wörtlicher Abdruck eines Teiles der 1866 erschienenen generellen Morphologie (Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformierte Deszendenz-Theorie),* Berlin. p. 215
Hamilton EP, et al. 2016. Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. Elife 5:e19090.
Harper DS, Jahn CL. 1989. Actin, tubulin and H4 histone genes in three species of hypotrichous ciliated protozoa. Gene.75:93-107.
Heaphy SM, et al. 2016. Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*. Mol Biol Evol. 33(11): 2885-2889.
Heaphy SM. 2018. Recoding and reassignment in protists. Doct Diss. University College Cork, Cork
Heywood J, Sieracki M, Bellows W, Poulton NJ, Stepanauskas R. 2011. Capturing diversity of marine heterotrophic protists: one cell at a time. ISME J 5(4):674–684.
Hogan DJ, Hewitt EA, Orr KE, Prescott DM, Müller KM. 2001. Evolution of IESs and scrambling in the actin I gene in hypotrichous ciliates. Proc Natl Acad Sci. 98:15101–15106.
Hunter S, et al. 2009. InterPro: the integrative protein signature database. Nucleic acids research..1(39):211-215.
Jahn CL, Klobutcher LA. 2002. Genome remodeling in ciliated protozoa. *Annu Rev Microbiol.* 56.1: 489-520.
Jung JH, Choi JM, Coats DW, Kim YO. 2016. *Eududoscquella costata* n. sp. (Dinoflagellata, Syndinea), an intracellular parasite of the ciliate *Schmidingerella arcuata*: morphology, molecular phylogeny, life cycle, prevalence, and infection intensity. J Eukaryot Microbiol. 63(1): 3-15.
Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28:27-30.
Katz LA, Lasek-Nesselquist E, Snoeyenbos-West OL. 2003. Structure of the micronuclear alpha-tubulin gene in the phyllopharyngean ciliate *Chilodonella uncinata*: implications for the evolution of chromosomal processing. Gene 315:15-9.
Katz LA, Kovner AM. 2010. Alternative processing of scrambled genes generates protein diversity in the ciliate *Chilodonella uncinata*. J Exp Zool Part B. 314.6: 480-488.
Keeling PJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Bio. 12(6).
Kervestin S, Frolova L, Kisselev L, Jean-Jean, O. 2001. Stop codon recognition in ciliates: *Euplotes* release factor does not respond to reassigned UGA codon. EMBO Rep. 2(8):680-684.
Landweber LF, Kuo TC, Curtis EA. 2000. Evolution and assembly of an extremely scrambled gene. P Natl Acad Sci. 97(7):3298-3303.
Lane DJ. 1991.16S/23S rRNA sequencing. In:Stackebrandt E, Goodfellow M. *Nucleic acid techniques in bacterial systematics.* New York: John Wiley & Sons, p.115-175.
Libusová L, Dráber P. 2006. Multiple tubulin forms in ciliated protozoan *Tetrahymena* and *Paramecium* species. Protoplasma. 227(2-4):65-76.
Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 15;31(10):1674-6.

López-Escardó D, et al. 2017. Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate *Monosiga brevicollis*. Sci Rep. 7:11025
Lozupone CA, Knight RD, Landweber LF. 2001. The molecular basis of nuclear genetic code change in ciliates. Curr Biol. 11(2):65-74.
Lynn DH. 2008. The Ciliated Protozoa. Springer. Orig. Published 1979.New York: Pergamon Press p. 65-199.
Maurer-Alcalá XX, Knight R, Katz LA. 2018a. Exploration of the Germline Genome of the Ciliate *Chilodonella uncinata* through Single-Cell Omics (Transcriptomics and Genomics). mBio. 9:1
Maurer-Alcalá XX, et al. 2018b. Twisted tales: insights into genome diversity of ciliates using single-cell 'omics.' Genome biol evol. 10.8:1927-1938.
Maurer-Alcalá XX, Nowacki M. 2019. Evolutionary origins and impacts of genome architecture in ciliates. *Ann NY Acad Sci.*1447.1: 110.
Medlin L, Elwood HJ, Stickel S, Sogin ML. 1988. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene, 71(2):491-499.
Miceli C, et al. 1994. Identification of the tubulin gene family and sequence determination of one β-tubulin gene in a cold-poikilotherm protozoan, the Antarctic ciliate *Euplotes focardii*. J Eukaryot Microbiol. 41(4): 420-427.
Montagnes DJS. 2013. Ecophysiology and behaviour of Tintinnids. In: Dolan JR, Agatha S, Coats DW, Montagnes DJS, Stoecker DK. *Biology and Ecology of Tintinnid Ciliates: Models for Marine Plankton*, Oxford: Wiley-Blackwell. p 86-122.
Montagnes DJS, et al. 2008. Selective feeding behaviour of key free-living protists: avenues for continued study. *Aquat. Microb. Ecol.* 53(1): 83–98.
Müller OF. 1779. Zoologia Danica seu animalium Daniae et Norvegiae rariorum ac minus notrorum descriptiones et historia 1. Weygandinis, Havniae et Lipsiae.
Nurk S, et al. 2013. Assembling genomes and mini-metagenomes from highly chimeric reads. International Conference on Research in Computational Molecular Biology. 1:158-170.
Parfrey LW, Lahr DJ, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci -USA 108(1): 13624–13629.
Pederson T. 2010. An Olympian Protozoan. *Nucleus* 1(1): 2–3.
Prescott DM. 2000. Genome gymnastics: unique modes of DNA evolution and processing in ciliates. Nat Rev Genet 1(3): 191-198.
Riley JL, Katz LA. 2001. Widespread distribution of extensive genome fragmentation in ciliates. Mol Biol Evol.18(7):1372–1377.
Santoferrara LF, McManus GB, Alder VA. 2013. Utility of genetic markers and morphology for species discrimination within the order Tintinnida (Ciliophora, Spirotrichea). Protist. (1):24-36.
Santoferrara LF, Rubin E, McManus GB. 2018. Global and local DNA (meta)barcoding reveal new biogeography patterns in tintinnid ciliates. J Plankton Res. 40(1):209–221.
Spits C, et al. 2006. Whole-genome multiple displacement amplification from single cells. *Nat protoc*. *1*(4): 1965.
Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 24(5): 637-644.
Swart EC, et al. 2013. The *Oxytricha trifallax* Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes. *PLoS Biol* 11(1).
Wang Y, et al. 2017. A comparative study of genome organization and epigenetic mechanisms in model ciliates, with an emphasis on *Tetrahymena, Paramecium and Oxytrichda.* Eur J Protistol. 61:376-387.
Waterhouse RM, et al. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular biology and evolution. 35(3):543-548.
Yan Y, Maurer-Alcalá XX, Knight R, Pond SLK, Katz LA. 2019. Single-Cell Transcriptomics Reveal a Correlation between Genome Architecture and Gene Family Evolution in Ciliates. mBio. 10(6).

Yi Z, Huang L, Yang R, Lin X, Song W. 2016. Actin evolution in ciliates (Protist, Alveolata) is characterized by high diversity and three duplication events. Mol Phylogenet Evol. 96:45-54.
Zufall RA, Katz LA. 2007. Micronuclear and macronuclear forms of betatubulin genes in the ciliate *Chilodonella uncinata* reveal insights into genome processing and protein evolution. J Eukaryot Microbiol 54:275–282.
Zufall RA, Robinson T, Katz LA. 2005. Evolution of developmentally regulated genome rearrangements in eukaryotes. J Exp Zool Part B. 304(5): 448-455.

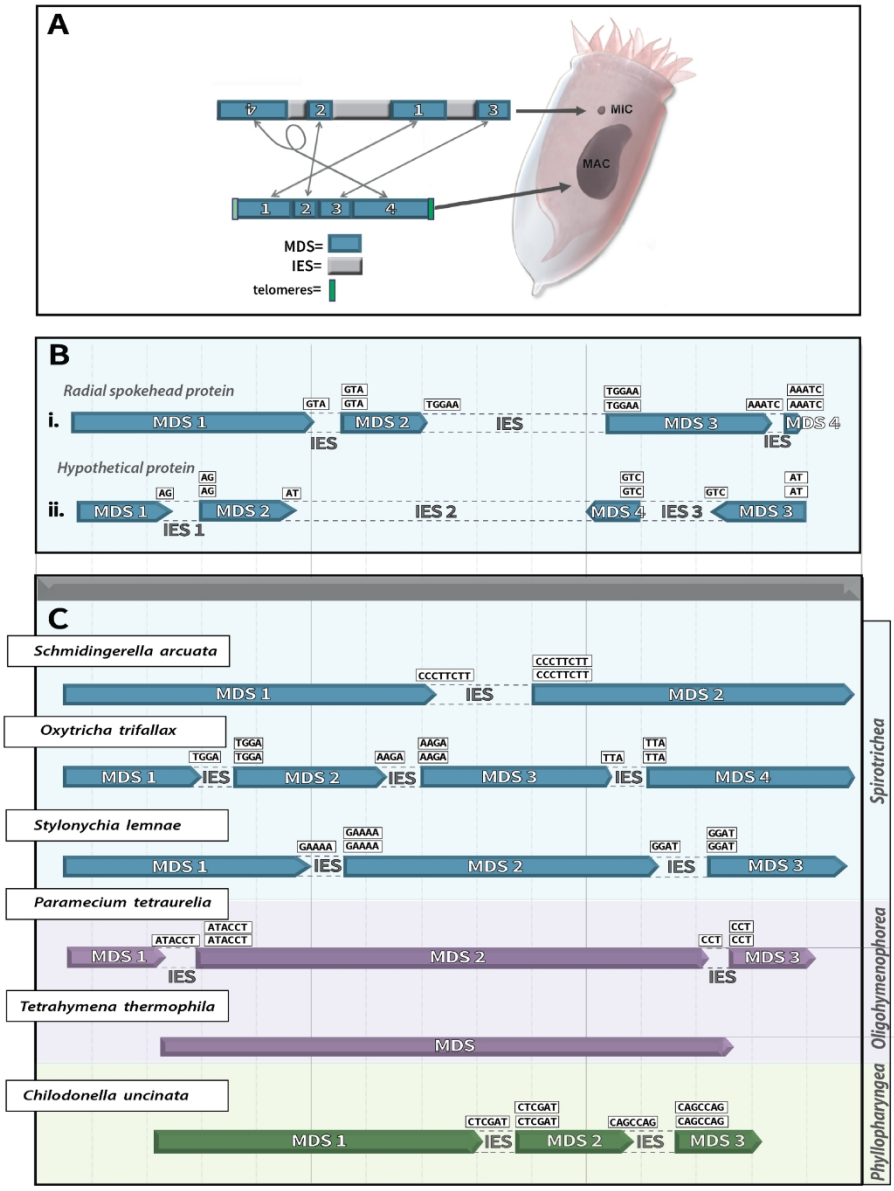

Figure 1: Genome architecture in *S. arcuata*. A: Patterns of nuclear architecture in the germline micronucleus (MIC) and somatic macronucleus (MAC). The MIC contains the required macronuclear-destined sequences (MDSs) for the generation of functional genes during development of the MAC. MDSs can be interrupted by internally-eliminated sequences (IESs). Development of the MAC requires the precise excision of IESs and the correct rearrangement of MDS regions. MDS loci may be scrambled (e.g. MDS 1-4), inverted (e.g. MDS 4), or a combination of both. The organization of these MDSs is guided by pointer sequences (2-10bp) that occur at MDS/IES boundaries. Green blocks capping ends indicate telomeres, which are added de novo to the ends of MAC chromosomes. B: Exemplar micronuclear patterns of loci elimination, scrambling and inversion identified in *S. arcuata*. i. Consecutive MDS regions of varying size separated by IESs and guided by 3-5bp pointers. ii. MDSs separated by IESs of variable lengths; a mix of scrambled, non-scrambled, and partially-inverted loci, with short pointers (2-3bp). Pointer sequences are shown in white blocks; those appearing twice indicate their secondary location in the MIC (pointers occur twice in MIC and once in MAC). MDSs are numbered according to their somatic order in the MAC. Arrows at the end of each MDS indicate MDS directionality in the MIC. C. Micronuclear architecture of a beta-tubulin gene in various

ciliates. *S. arcuata* separates the gene region into two MDSs, interrupted by a single IES and guided by an 8bp pointer region. Different colors of MDS correspond to different classes, indicated at right. Accession numbers or gene identifiers: *O. trifallax* (PRJNA194431; OxyDB: Contig11167.0.g9); *S. lemnae* (X06874.1); *T. thermophila* (L01416.1); *P.caudatum* (AB070222.1); *C. uncinata* (MH388464).

300x357mm (300 x 300 DPI)