Analysis of the somatic and germline genomes of the ciliate *Blepharisma stoltei*

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Minakshi Singh

Pune/Indien

Tübingen

2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	06.10.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Dr. Estienne Swart
2. Berichterstatter/-in:	Prof. Dr. Ralf-Peter Jansen

I hereby declare that the thesis I submit for my doctorate with the title:
„**The genomic analysis of somatic and germline nuclei of the ciliate *Blepharisma stoltei***“ is my own independent work, that I used only the sources and resources cited and have clearly indicated all content adopted either word-for-word or in substance. I declare that the University of Tübingen's guidelines to ensure good academic practice (Senate decision of 25.5.2000) have been observed. I solemnly swear that this information is true and that I have not concealed any relevant information. I am aware that making a false declaration is punishable by a fine or by a prison term of up to three years.

Acknowledgements

I thank Dr. Estienne Swart for giving me the opportunity to learn and work in his group. His patience and encouragement allowed me to learn a lot of new things and has made this project possible. My heartfelt thanks to Prof. Dr. Ralf Jansen and Prof. Dr. Daniel Huson for their guidance and support as part of my thesis advisory committee. I would like to thank Prof. Dr. Terue Harumoto and Prof. Dr. Mayumi Sugiura for hosting me in their lab in Nara (Japan) and kindly and patiently sharing their expertise and knowledge of *Blepharisma* with me. My thanks also to Prof. Dr. Detlef Weigel for his encouragement, mentorship and interest in my work. Thanks to Prof. Dr. Debjani Paul and Prof. Dr. Patrick Müller for their guidance during my Bachelors and Masters degrees.

Chapter 4 is dedicated to the memory of Akio Miyake and his decades of inspirational *Blepharisma* research. Thanks to Federico Buonanno for the provision of *B. stoltei* ATCC 30299 cells and culturing advice. Thanks also to Christa Lanz and the MPI for Biology's genome center, and Sebastien Colin and the MPI for Biology's Light Microscopy Facility for the 3D nuclear reconstruction.

I would like to thank Brandon Seah, Christiane Emmerich, Aditi Singh and Lilia Häußermann for all the productive discussions over cake and otherwise, for making the lab a supportive and productive environment and for the fun times outside the lab.
Thanks to Christian Feldhaus and Aurora Panzera for their expertise and patience in training and retraining me at microscopy facility. Thank you also for the discussions on how to do fun things with microscopes, dyes and strange creatures. Thanks for making sure we remember, that at the end of the day science is really, really cool.

A lot of help from a lot of people went into this work.

Thanks to …

Brandon Seah and Christiane Emmerich for going over the several drafts of my work and giving me helpful feedback.

To everyone in lab who patiently sat through my practice talks. I've learnt so much from you.

Sybille Patheiger, Jeanette Müller, George Deffner, Tina Peart and Sarah Danes for their support through matters administrative, legal and otherwise.

Christa Lanz for her kind help and patient discussions on all manner of sequencing issues and for promptly including my samples in her FEMTO runs.

Andre Noll, Harald Huss, Henry Vogt and Thomas Helle for their support with the cluster and IT infrastructure.

Sinja Mattes for her dedicated culturing and care of our precious cultures.

Luis Antoniotti for hand-crafting the cell-filtering apparatus in his workshop.

Thanks to Thanvi Srikant, Sara Wighard, Devansh Raj, Anasuya Moitra and Bridgit Waithaka for their delightful company and discussions about all and sundry at B.E.E.R. Hour and *Chai*.

A special thanks to the once and current PhD Reps. I remain inspired by the enthusiasm, creativity and dedication you brought to the team.

Carl, Raju, Vita, Ting, Ana, Devansh, Lena, Maja, Joha and Martina, thank you for your friendship, kindness and warmth. With all of you around, I never felt like I was too far from home.

Finally, I would like to thank my parents and loved ones for their unflinching support, kindness, care and encouragement.

Table of contents

Abstract *14*

Zusammenfassung *16*

Chapter 1 Introduction *18*

1.1. Nuclear dualism and genome reorganization in ciliates 18

1.2. *Blepharisma*: an early-diverging ciliate lineage 22

1.3. Bibliography 27

Chapter 2 Research Aims *30*

Chapter 3 Preliminary assembly and annotation of the draft genome of the Blepharisma stoltei somatic nucleus *31*

3.1. Introduction 31

3.2. Results 33

3.2.1. DNA extraction 33

3.2.1.1. DNA extraction from whole-cell lysate 33

3.2.1.2. DNA extraction from MAC-enriched lysate 35

3.2.2. Assembly 36

3.2.2.1. Short-read assembly 36

3.2.2.2. Long-read assembly 37

3.2.3. Annotation 39

3.2.3.1. Structural annotation with AUGUSTUS 39

3.2.3.2. Functional annotation with HMMER3, Pannzer2 and eggNOG 41

3.3. Discussion 42

3.3.1. High molecular weight DNA was extracted from MAC- and MIC-enriched fractions 42

3.3.2. Short-read assemblies were produced for two strains of B. stoltei 43

3.3.3. Structural gene prediction was performed using AUGUSTUS 43

3.3.4. Functional gene prediction was performed with HMMER3, Pannzer2 and eggNOG 44

3.4. Bibliography 46

Chapter 4 Expression of genome reorganization gene homologs during sexual reproduction in B. stoltei *49*

4.1. Introduction 50

4.2. Results 52

4.2.1. Morphological staging of conjugation time-course in *B. stoltei* 52

4.2.2. Transcriptomic analysis of the conjugation time-course in *B. stoltei* 55
4.2.2.1. Small RNA biogenesis machinery in B. stoltei 55
4.2.2.2. Development-specific histone variant upregulation 57
4.2.3. Transposase domains encoded in the *B. stoltei* somatic genome and their expression 59
4.2.3.1. PiggyBac family (DDE_Tnp_1_7) transposases in the MAC genome 60
4.2.3.2. Pogo/Tigger family (DDE_1) transposases in the MAC genome 62
4.2.3.3. Tc1/Mariner family (DDE_3) transposases in the MAC genome 62
4.2.3.4. Merlin family (DDE_Tnp_IS1595) and Mutator family (MULE) transposases in the MAC genome 63
4.3. Discussion 65
4.3.1. Small RNA biogenesis in ciliates is critical for genome reorganization 65
4.3.2. Histones in *Blepharisma* 66
4.3.3. *Blepharisma* possesses additional domesticated transposases whose roles await determination 66
4.4. Bibliography 70
Chapter 5 Genome editing excisase origins illuminated by somatic genome of Blepharisma 74
5.1. Introduction 75
5.2. Results 77
5.2.1. A compact somatic genome with a minichromosomal architecture 77
5.2.1.1. Blepharisma has short spliceosomal introns 79
5.2.1.2. Blepharisma has minichromosomal architecture in the somatic genome 83
5.2.2. PiggyBac transposases in the somatic and germline genomes of *Blepharisma* 86
5.2.2.1. A single Blepharisma PiggyBac homolog has a complete catalytic triad 87
5.2.2.2. PiggyBac transposases are subject to purifying selection 89
5.2.3. PiggyBac transposases originated early in ciliate evolution 90
5.3. Discussion 93
5.3.1. The *Blepharisma* MAC genome is organized as minichromosomes 93
5.3.2. A PiggyBac is the main IES excisase in *Blepharisma* 93
Supplementary figures 96
Supplementary tables 98
5.4. Bibliography 103
Chapter 6 MITE infestation of germline accommodated by genome editing in Blepharisma 107

6.1. Introduction 108
6.2. Results 111
6.2.1. Detection and targeted assembly of ca. forty thousand germline-limited IESs 111
6.2.2. A “hybrid” IES length distribution with periodic length peaks for short IESs 111
6.2.3. IESs are bounded by heterogeneous direct and inverted terminal repeats 111
6.2.4. Repeat elements are abundant in long, non-periodic IESs 115
6.2.5. Germline-limited repeats include few autonomous transposons but many MITEs 116
6.2.5.1. Pogo/Tigger-family transposon with abundant MITEs 117
6.2.5.2. Tc1-family transposon with microsatellites 119
6.2.6. Non-LTR retrotransposon sequences in both the somatic and germline genomes 120
6.2.7. Development-specific 24 nt small RNAs are likely scnRNAs in *Blepharisma stoltei* 122
6.2.8. Putative scnRNAs have lower coverage over periodic IESs and BogoMITE IESs 124
6.3. Discussion 126
6.3.1. Comparison to IESs in other ciliates 126
6.3.2. Are MITEs a missing link in the IBAF model? 127
6.3.3. Is “genome defense” a flawed analogy? 129
6.3.4. Why does the *Blepharisma* somatic genome have retrotransposon-derivatives? 130
6.3.5. Conclusion 132
Supplementary figures 133
Supplementary tables 141
6.4. Bibliography 148
Chapter 7 General Discussion 155
7.1. The macronuclear and micronuclear genomes of *Blepharisma* are structurally and functionally annotated 157
7.2. Expression of genes in the genome reorganization toolkit is upregulated during development of the new MAC 157
7.3. Multiple transposase families and a putative IES excisase in Blepharisma 158
7.4. *Blepharisma* IESs share several characteristics with *Paramecium* IESs 159
7.5. The last common ancestor of ciliates possessed a PiggyBac 160
7.6. The *Blepharisma* germline genome indicates the early origin of IESs 162
7.7. MITIES in *Blepharisma* represent an intermediate stage in IES generation and transposon domestication 162
7.8. Conclusion and outlook 163
7.9. Bibliography 165
Chapter 8 Materials and methods 170
8.1. Strains and localities 171

8.2. Cell cultivation, harvesting and cleanup 171
8.3. DNA isolation from whole cells and macronuclei, library preparation and sequencing 172
8.4. *Enrichment of micronuclei, isolation and sequencing of MIC genomic DNA* 172
8.5. Genome assembly 173
8.5.1. Chapter 3 173
8.5.2. Chapter 5 173
8.6. Gene prediction 175
8.7. Functional gene annotation 176
8.8. Gamone 1/ Cell-Free Fluid (CFF) isolation and conjugation activity assay 177
8.9. Conjugation time course and RNA isolation for high-throughput sequencing 177
8.10. RNA-seq read mapping 178
8.11. Gene expression analysis 178
8.12. Repeat annotation 179
8.13. Cell fixation and imaging 179
8.14. Variant calling 180
8.15. Annotation of alternative telomere addition sites 180
8.16. Genetic code prediction 181
8.17. Assessment of genome completeness 181
8.18. Gene expression analysis 182
8.19. Sequence visualization and analysis 182
8.20. Identification and correction of MIC-encoded PiggyBac homologs 182
8.21. d_N/d_S estimation 184
8.22. Phylogenetic analysis of eukaryotic PiggyBac-like elements 184
8.23. Repeat annotation 184
8.24. IES prediction from PacBio subreads 185
8.25. Identification and comparison of IES length classes 186
8.26. Probability of terminal direct repeat-bound IESs 186
8.27. Identification of terminal inverted repeats (TIRs) and palindromes in IESs 187
8.28. Comparison of intragenic:intergenic IES ratios 188
8.29. Developmental time series small RNA-seq 188
8.30. Small RNA libraries mapping and comparison 188
8.31. Gene prediction and domain annotation in IES regions 189
8.32. Repeat annotation and clustering 190
8.33. Phylogenetic analysis of Tc1/Mariner-superfamily transposases 191
8.34. Phylogenetic analysis of retrotransposon-derived sequences 192
8.35. Sequence visualization and analysis 192
8.36. Data availability 192
8.37. Bibliography 194
Appendix 202

Author abbreviations 202

A.1. Author contributions for Chapters 4, 5 and 6 203

Chapter 4 203

Chapter 5 203

Chapter 6 203

A.2. Author contributions for Chapters 7 and 8 205

Chapter 7 205

Chapter 8 205

List of figures

Figure 1.1. Schematic depictions of ciliate body plans and nuclear morphology. 19

Figure 1.2. The macronuclear genome develops from the micronuclear genome through genome reorganization during sexual reproduction. 20

Figure 1.3. A *Blepharisma japonicum* cannibal giant 23

Figure 1.4. Starvation induces cells to produce gamones. 24

Figure 1.5. Schematic of nuclear processes occurring during conjugation (classified according to, and modified from (Miyake et al., 1991)). 25

Figure 3.1. *Blepharisma stoltei* ATCC cell. 33

Figure 3.2. Column-purified *B. stoltei* ATCC DNA from whole-cell lysate. 34

Figure 3.3. Phenol-chloroform-purified *B. stoltei* ATCC DNA from whole-cell lysate. 34

Figure 3.4. *B. stoltei* ATCC MAC-enriched fraction isolated using gravity-flow purification columns. 35

Figure 4.1. Developmental staging of *B. stoltei* for RNA-seq. 53

Figure 4.2. ResIII, Helicase_c and Ribonuclease_3 domains in *B. stoltei.* 55

Figure 4.3.PIWI domain in *B. stoltei.* 56

Figure 4.4. Histones and histone-domain-containing proteins in *Blepharisma*. 58

Figure 4.5. MAC genome-encoded transposases in ciliates. 59

Figure 4.6. DDE_Tnp_1_7 domain in *B.stoltei.* 60

Figure 4.7. DDE_1 domain-containing proteins in *Blepharisma*. 62

Figure 4.8. DDE_3 domain-containing proteins in *Blepharisma*. 63

Figure 4.9. DDE_Tnp_IS1595 domain-containing proteins in *Blepharisma*. 63

Figure 4.10. MULE domain-containing proteins in *Blepharisma*. 64

Figure 5.1. Analysis of assembly completeness and genetic code. 77

Figure 5.2. Intron splicing. 81

Figure 5.3. A gene-dense somatic genome with a minichromosomal architecture. 82

Figure 5.4. Properties of minichromosomes, telomeres, and alternative telomere addition sites. 84

Figure 5.5. MAC genome-encoded transposases in ciliates and properties of a putative *Blepharisma* IES excisase. 89

Figure 5.6. Phylogeny of ciliate PiggyBac homologs and eukaryotic PBLEs. 92

Figure 6.1. A "hybrid" IES length distribution with periodic length peaks for short IESs. 112

Figure 6.2. IESs are bounded by heterogeneous direct and inverted terminal repeats. 113

Figure 6.3. Repeat elements are abundant in long, non-periodic IESs. 116

Figure 6.4. Germline-limited repeats include few autonomous transposons but many MITEs. 118

Figure 6.5. Non-LTR retrotransposon sequences in both somatic and germline genomes. 121

Figure 6.6. Development-specific 24 nt small RNAs are likely scnRNAs in *B. stoltei*. 124

Figure 6.7. Model for transposon fixation as IESs in a ciliate genome with an existing domesticated excisase. 128

List of tables

Table 3.1. Assessment of SPAdes genome assemblies of *Blepharisma* species from Illumina reads. 37

Table 3.2. Assessment of genome assemblies of *B. stoltei* ATCC. 38

Table 3.3. AUGUSTUS parameters and respective gene prediction accuracy at the exon level. .40

Table 5.1. *Blepharisma* PiggyMac-like substitution rates. 90

List of abbreviations

- MIC - micronucleus
- MAC - macronucleus
- IES - internally eliminated sequence
- MDS - macronuclear-destined sequence
- PacBio - Pacific Biosciences
- CLR - continuous long read (PacBio)
- CCS - circular consensus sequence (PacBio)
- HiFi - High-fidelity read (PacBio)
- ATAS - alternative telomere addition site
- PBLE - PiggyBac-like element
- PGBD - PiggyBac element-derived
- Pgm - PiggyMac
- PgmL - PiggyMac-like
- IES - internally eliminated sequence
- LTR - long terminal repeat
- MAC - macronucleus
- MIC - micronucleus
- MITE - miniature inverted-repeat transposable element
- MITIES - miniature inverted-repeat transposable internally eliminated sequences
- TDR - terminal direct repeat
- TIR - terminal inverted repeat
- TSD - target site duplication

Abstract

Ciliates are prototypical, conventionally unicellular eukaryotes with separate germline and somatic nuclei. The somatic genome arises from the germline genome through a process of transposase-mediated DNA elimination and genome rearrangement during sexual reproduction. Current models for genome reorganization in ciliates posit that small RNAs are transported to the developing somatic nucleus during sexual reproduction, aiding transposases in identifying and excising germline-specific sequences. Accompanying these sequences, known as Internally Eliminated Sequences (IESs), and their excisases is the machinery to carry out their removal. This includes Dicer-like and Piwi/Argonaute proteins, which generate and transport small RNAs, as well as proteins that alter chromatin, and make DNA accessible for excision.

The ciliate *Blepharisma* belongs to an early diverging class of ciliates known as the Heterotrichea. Though genome reorganization has been studied in later diverging ciliates such the oligohymenophorean ciliates *Tetrahymena* and *Paramecium* and the spirotrich *Oxytricha* there are pronounced differences in how they do so. Studying this process in an early diverging ciliate like *Blepharisma* is an important contribution to the understanding of how conserved the different elements of the genome reorganization machinery among ciliates are. This thesis provides the first look, from a genomic perspective, at the various participants and putative mechanisms of genome reorganization in *Blepharisma*.

Annotated reference genomes for the somatic and germline nuclei of *Blepharisma* stoltei (strain ATCC 30299) were generated using long-read sequencing and annotation methods tailored to the atypical genome properties of *Blepharisma*. The *B. stoltei* somatic genome is compact (41 Mb), gene-dense (25710 genes) and contains short, 15-16 nucleotide spliceosomal introns.

We identified key components involved in genome reorganization in the *Blepharisma* somatic genome and compared them with those of the model ciliates *Paramecium*, *Tetrahymena* and *Oxytricha*. Four transposase families were found encoded in the somatic and germline genomes, namely the PiggyBac, Tc1/Mariner, Mutator and Merlin families. PiggyBac transposases are known to be the main transposases involved in genome reorganization in the model ciliates *Paramecium* and *Tetrahymena*, but are entirely absent in *Oxytricha*, which is thought to use a transposase from another family. In *Paramecium*, six somatically encoded PiggyBacs incapable of catalysis, plus one catalytically complete homolog called the PiggyMac,

coordinate DNA excision. This resembles the situation in *Blepharisma*, which has thirteen homologs of the PiggyBac transposase, only one of which has a complete catalytic triad and is hence likely to be the primary excisase.

The germline-limited genomic regions of *Blepharisma* were also characterized. *Blepharisma* IESs share two key features with the IESs of *Paramecium*, namely a periodic length distribution for short IESs and predominantly TA-dinucleotide delineated IES boundaries. We also identified a class of 24-nucleotide small RNAs that increasingly map to IESs as development progresses in *Blepharisma*. These trends are similar to those observed in *Paramecium* and *Tetrahymena*, hence we propose that they are also so-called "scan" RNAs (scnRNAs) that guide IES excision.

Phylogenetic analysis of the *Blepharisma* PiggyBac homologs showed that they share common ancestry with the PiggyBac homologs of *Paramecium* and *Tetrahymena*, where the latter are evolutionarily more divergent than *Blepharisma* and are located on more recently diverging branches of the ciliate phylogenetic tree. Several lines of evidence from these studies therefore indicate that a PiggyBac transposase is the most likely the main IES excisase in *Blepharisma* and that the last ciliate common ancestor also possessed this type of transposase.

Zusammenfassung

Ciliaten sind prototypische, üblicherweise einzellige Eukaryoten mit getrennten Keimbahn- und somatischen Zellkernen. Das somatische Genom entsteht aus dem Keimbahngenom durch einen Prozess der Transposase-vermittelten DNA-Eliminierung und Genom-Neuordnung während der sexuellen Fortpflanzung. Aktuelle Modelle für die Reorganisation des Genoms bei Wimpertierchen gehen davon aus, dass kleine RNAs während der sexuellen Fortpflanzung in den sich entwickelnden somatischen Kern transportiert werden und Transposasen dabei helfen, keimlinienspezifische Sequenzen zu identifizieren und auszuschneiden. Diese Sequenzen, die so genannten intern eliminierten Sequenzen (IES), und ihre Exzisasen werden von einer Maschinerie begleitet, die ihre Entfernung durchführt. Dazu gehören Dicer-ähnliche und Piwi/Argonaute-Proteine, die kleine RNAs erzeugen und transportieren, sowie Proteine, die das Chromatin verändern und die DNA für die Exzision zugänglich machen.

Blepharisma gehört zu einer früh divergiereden Klasse von Ciliaten, die als Heterotrichea bekannt sind. Obwohl die Reorganisation des Genoms bei später divergierenden Ciliaten wie den oligohymenophoren Ciliaten *Tetrahymena* und *Paramecium* und den spirotrichen *Oxytricha* untersucht wurde, gibt es deutliche Unterschiede in der Art und Weise, wie sie dies tun. Die Untersuchung dieses Prozesses in einem früh divergiereden Ciliaten wie *Blepharisma* ist ein wichtiger Beitrag zum Verständnis, wie konserviert die verschiedenen Elemente der Genom-Reorganisationsmaschinerie unter Ciliaten sind. Diese Arbeit bietet den ersten Blick aus genomischer Sicht auf die verschiedenen Teilnehmer und mutmaßlichen Mechanismen der Genomreorganisation in *Blepharisma*.

Mittels Long-Read-Sequenzierung und Annotationsmethoden, die auf die atypischen Genomeigenschaften von *Blepharisma* zugeschnitten sind, wurden annotierte Referenzgenome für die somatischen und Keimbahnkerne von *Blepharisma* stoltei (Stamm ATCC 30299) erstellt. Das somatische Genom von B. stoltei ist kompakt (41 Mb), gen-dicht (25710 Gene) und enthält kurze, 15-16 Nukleotide umfassende spliceosomale Introns.

Wir haben Schlüsselkomponenten identifiziert, die an der Reorganisation des Genoms im somatischen Genom von *Blepharisma* beteiligt sind, und sie mit denen der Modell-Ciliaten *Paramecium*, *Tetrahymena* und *Oxytricha* verglichen. Es wurden vier Transposase-Familien gefunden, die in den somatischen und Keimbahn-Genomen kodiert sind, nämlich die PiggyBac-,

Tc1/Mariner-, Mutator- und Merlin-Familien. Es ist bekannt, dass PiggyBac-Transposasen die wichtigsten Transposasen sind, die in den Modell-Ciliaten *Paramecium* und *Tetrahymena* an der Reorganisation des Genoms beteiligt sind, während sie in *Oxytricha,* wo vermutlich eine Transposase aus einer anderen Familie verwendet wird, gänzlich fehlen. In *Paramecium* koordinieren sechs somatisch kodierte PiggyBacs, die nicht zur Katalyse fähig sind, sowie ein katalytisch vollständiges Homolog, namens PiggyMac, die DNA-Exzision. Dies ähnelt der Situation in *Blepharisma*, wo es dreizehn Homologe der PiggyBac-Transposase gibt, von denen nur eine eine vollständige katalytische Triade besitzt und daher wahrscheinlich die primäre Exzisase ist. Die keimbahnbegrenzten genomischen Regionen von *Blepharisma* wurden ebenfalls charakterisiert. Die IES von *Blepharisma* haben zwei wesentliche Merkmale mit den IES von *Paramecium* gemeinsam, nämlich eine periodische Längenverteilung für kurze IES und überwiegend durch TA-Dinukleotide abgegrenzte IES-Grenzen. Wir haben auch eine Klasse von kleinen RNAs („small RNAs“) mit 24 Nukleotiden identifiziert, die mit fortschreitender Entwicklung in *Blepharisma* zunehmend den IESs zugeordnet werden. Diese Tendenzen ähneln denen, die in *Paramecium* und *Tetrahymena* beobachtet wurden, weshalb wir vorschlagen, dass es sich auch hier um so genannte "Scan"-RNAs (scnRNAs) handelt, die die IES-Exzision steuern.

Die phylogenetische Analyse der PiggyBac-Homologe von *Blepharisma* hat gezeigt, dass sie einen gemeinsamen Ursprung mit den PiggyBac-Homologen von *Paramecium* und *Tetrahymena* haben, wobei letztere evolutionär stärker divergieren als *Blepharisma* und auf jüngeren Zweigen des phylogenetischen Stammbaums der Ciliaten zu finden sind. Mehrere Indizien aus diesen Studien deuten daher darauf hin, dass eine PiggyBac-Transposase höchstwahrscheinlich die wichtigste IES-Exzisase in *Blepharisma* ist und dass der letzte gemeinsame Vorfahre der Ciliaten ebenfalls diesen Transposasetyp besaß.

Chapter 1
Introduction

Plants, animals and fungi only constitute two of the eleven major eukaryotic lineages (Keeling and Burki 2019). The majority of eukaryotes are microscopic, free-living and unicellular. Ciliates are one such group of eukaryotes, with separate germline and somatic nuclei in a single cell. The somatic nuclei develop from germline nuclei during sexual reproduction, during which large sections of the germline genome are eliminated from the developing somatic nucleus (Prescott 1994). The process of DNA elimination and reorganization which occurs in ciliates is quite different form the mechanisms of DNA elimination which have been studied in metazoans such as *Ascaris* worms and later in zebra finches, lampreys and rotifers (Wang and Davis 2014). In metazoans, DNA is usually eliminated in the form of chromosome breakage, diminution or rearrangement, and regulates gene dosage or mediates the sexual or immunological identity of the cell. In ciliates, in contrast, DNA elimination is associated with the segregation of the soma from the germline in two separate nuclei.

1.1. Nuclear dualism and genome reorganization in ciliates

Ciliates are characterized by hair-like appendages called cilia on the cell surface, which they use for locomotion and food acquisition. Ciliates have remarkably complex cell organization with dedicated feeding structures, usually lined with cilia or a ciliary membrane used to grab or hydrodynamically funnel food into the cell (Figure 1.1). The dual nuclei exhibit various morphologies (Figure 1.1). The somatic nucleus, known as a macronucleus (MAC), is often large and contains most of the DNA of the cell. The germline nuclei, known as micronuclei (MIC) are smaller and are transcriptionally inactive during the vegetative lifecycle of the cell., located near the MACs.

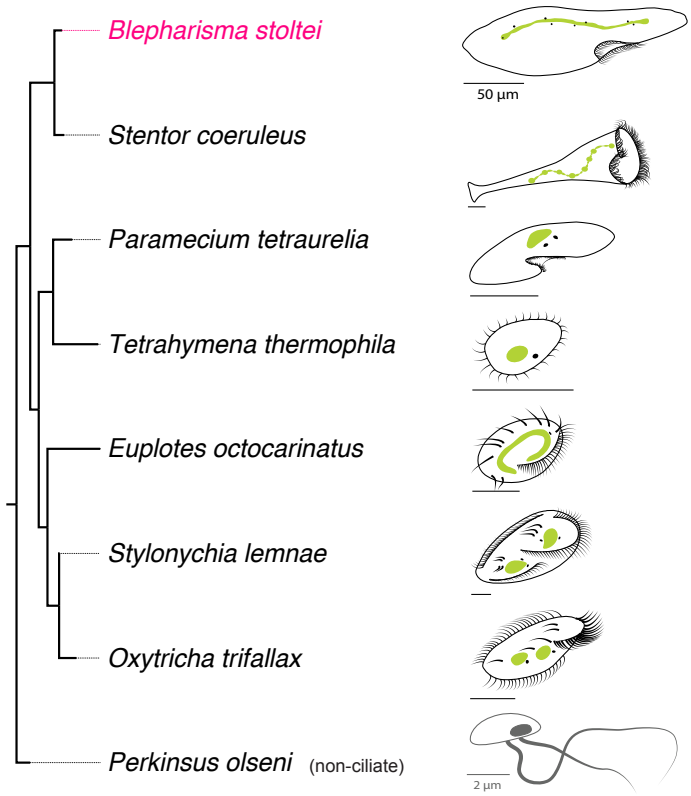


Figure 1.1. Schematic depictions of ciliate body plans and nuclear morphology. MACs are green and MICs are small black dots in close proximity to MACs. Scale bars – 50 µm, except for ***Perkinsus olseni* 2 µm**. A phylogenetic tree constructed with small subunit rRNA sequences is shown next to the ciliates, using the non-ciliate dinoflagellate *Perkinsus marinus* as an outgroup**.**

Only the pre-eminent ciliary structures of the cells are depicted in Figure 1.1, such as oral grooves and cellular ridges. The cell surfaces of the ciliates *Blepharisma stoltei, Stentor coeruleus, Paramecium tetraurelia* and *Tetrahymena thermophila* are actually densely packed with rows of cilia (not shown). The ciliates *Euplotes octocarinatus, Stylonychia lemnae* and *Oxyrticha trifallax* have rows of cilia interspersed with pellicular ridges with large, spiny ciliary structures (Figure 1.1).

Ciliates divide clonally, through vegetative reproduction, where the mother cell divides into two daughter cells. The MAC of the mother cell divides amitotically and pinches into two sections roughly equal in size, corresponding to each of the daughter cells (Orias 1991). Amitosis of the MAC is coordinated by nucleoplasmic microtubules (Tucker et al. 1980), though it lacks some of the more conspicuous features of eukaryotic mitosis such as chromatin condensation, spindle formation and histone H3 phosphorylation (Flickinger 1965). Simultaneously, the MICs in the mother cell are distributed among the daughter cells, where they replicate mitotically and

restore their original number (Prescott 1994). The DNA content of the amitotically divided MAC also undergoes amplification in the daughter cells and is restored to its former volume (Woodard, Kaneshiro, and Gorovsky 1972). Clonal propagation is how most lab strains of ciliates are maintained in the laboratory.

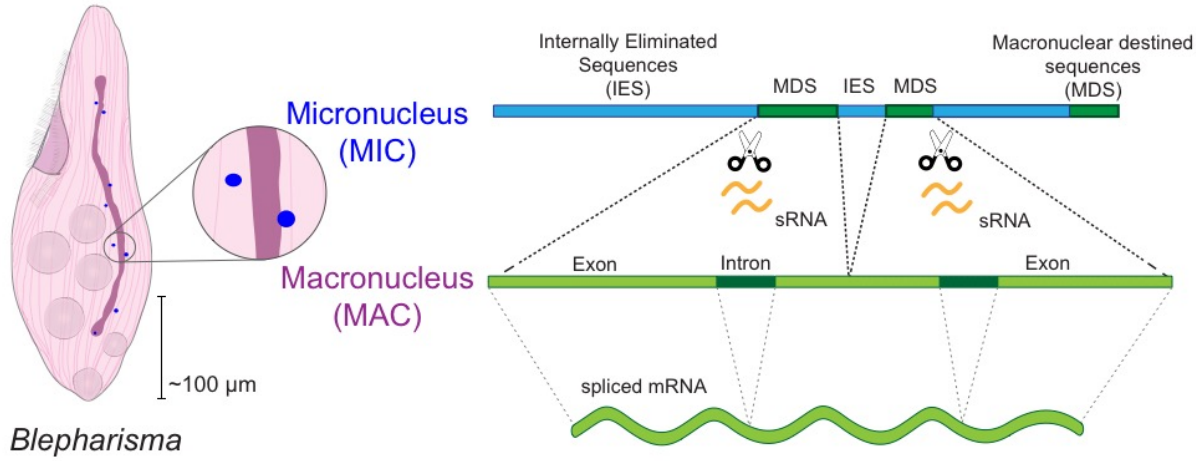


Figure 1.2. The macronuclear genome develops from the micronuclear genome through genome reorganization during sexual reproduction. The scissors represent transposases involved in the elimination of MIC-limited genomic regions (IESs).

In most ciliates, the germline nucleus is physically smaller in size than the somatic nucleus. Some ciliates have multiple MICs and an equal or fewer number of MACs. The MIC is diploid and is transcriptionally silent during the vegetative life cycle of the cell. The MAC, in contrast, contains a segmented, amplified version of the MIC genome (Prescott 1994). These amplified segments correspond to specific regions of the MIC, known as macronuclear destined sequences (MDS). The germline genome contained in the MIC consists of the all the genetic information present in the MAC plus MIC-limited sequences. During sexual reproduction however, the somatic nucleus emerges from the germline nucleus, by elimination of regions of the germline genome through a process of DNA elimination, which is mediated by domesticated transposases (Prescott 1994)(Figure 1.2).

The ciliate-specific process of genome reorganization which allows the emergence of the MAC genome from a MIC genome occurs in addition to the other canonical processes of meiosis, gametic fusion and recombination, which are characteristic of eukaryotic sexual reproduction. The regions of the germline genome which are eliminated during the development of the new MAC are known as Internally Eliminated Sequences (IESs) and the MDSs constitute the remaining regions of the germline genome present in the MAC (Figure 1.2). The process of

IES elimination occurs in an intermediate nuclear body known as a "macronuclear anlagen" (MA) or the developing MAC, which appears during the later stages of sexual reproduction. The MA matures and serves as the new MAC of the cell, while the old MAC of the cell is destroyed during the time between the beginning of sexual reproduction and the eventual development of the new MAC (Miyake, Rivola, and Harumoto 1991). The excision of IESs is usually accompanied by the organization of the MDSs into chromosome-like structures, with telomeres at their ends. These chromosome-like regions can vary in their length, from containing multiple genes, as is the case for the ciliates *Tetrahymena* (Sheng et al. 2020; Eisen et al. 2006) and *Paramecium* (Aury et al. 2006; Duret et al. 2008) or be limited to gene-sized segments like those of *Oxytricha* (Swart et al. 2013).

MAC chromosomes are amplified to different extents in different species, resulting in the much larger DNA content of the MAC in comparison to the MIC. In *Tetrahymena thermophila*, the genomic content of the MAC is ~46 times that of the MIC (Woodard, Kaneshiro, and Gorovsky 1972), but there is differential amplification of the MAC segments and not all MAC segments are amplified to the same extent. Similarly, in *Paramecium tetraurelia*, the MAC contains ~800 times the differentially amplified genic content of the diploid MIC (Aury et al. 2006). In *Oxytricha trifallax,* the MAC content is differentially amplified ~2000 times (Prescott 1994). Additionally, in some ciliates such as *Oxytricha* and *Chilodonella unicata*, a rearrangement of MDS regions occurs during MAC development, known as genome unscrambling, where the order and orientation of MDSs in the MIC genome is not preserved in the new MAC genome (Chen et al. 2014; Katz and Kovner 2010). The process of genomic reorganization thus alters the genetic content, its total amount, and in certain cases the order of MAC genome regions in the new MAC.

In the model ciliates, *Paramecium* and *Tetrahymena,* the excision of IESs is carried out by domesticated transposases encoded in the MAC genome. These are domesticated transposases of the PiggyBac family (Cheng et al. 2010; Baudry et al. 2009). In *Paramecium* the main IES excisase called PiggyMac, was proposed to act in a heteromeric complex with six other PiggyMac-like proteins (PgmLs) (Bischerour et al. 2018) and excises IES precisely at IES boundaries, typically identifiable by the TA-dinucleotide. The main *Tetrahymena* PiggyBac, called Tpb2 acts to excise IESs (Cheng et al. 2010), but performs IES excision in an imprecise manner, though an extremely weak sequence bias for the terminal direct repeat TTAA is detected at IES boundaries

(Hamilton et al. 2016). It acts in concert with two other PiggyBac-derived genes, Tpb1 and Tpb6, which exclusively excise a subset of intragenic IESs in a precise manner (Feng et al. 2017; Cheng et al. 2016). In *Oxytricha*, transposases encode by a family of transposons known as Telomere Bearing Elements (TBEs) are thought to be required for the excision of IESs from the germline genome during genome reorganization (Williams, Doak, and Herrick 1993).

IESs themselves are usually repeat rich, consisting of transposable elements (TE) in various stages of activity and decay. The TBEs in *Oxytricha,* and similar elements from *Euplotes*, called Tec, are Tc1/Mariner transposons, a family of Type II ("cut -and-paste") transposons (Herrick et al. 1985; Jahn et al. 1993)A class of IESs in *Paramecium* are the same kind of transposons (Arnaiz et al. 2012). The resemblance of IESs in different ciliates to Tc1/Mariner transposons and their excision by domesticated transposases lead to the hypothesis that IES originated from cut-and-paste DNA transposons (Klobutcher and Herrick 1997).

There is enormous diversity of form and nuclear morphology, even among the ciliates (Figure 1). The magnitude of this variety is illustrated by the fact that the evolutionary distance between *Tetrahymena* and *Euplotes* is comparable to that between corn and the rat (Prescott 1994). This makes studies of any particular ciliate sub-group, relatively specific to that sub-group and it is through comparative studies that we have the opportunity to observe generalized principles of ciliate biology. *Paramecium, Tetrahymena* and *Oxytricha* are present on less divergent branches of the ciliate tree, in comparison to *Blepharisma* and *Stentor,* both members of the Heterotrichea class of ciliates constitute one of the earliest diverging lineages of ciliates (Figure 1.1).

1.2. *Blepharisma*: an early-diverging ciliate lineage

Blepharisma is a heterotrichous ciliate, characterized by its pink photosensitive pigment 'blepharismin'. It is relatively large for a unicellular organism, with specimens of the largest *Blepharisma* species, *B. japonicum*, growing up to 500 μm on their longitudinal axis.

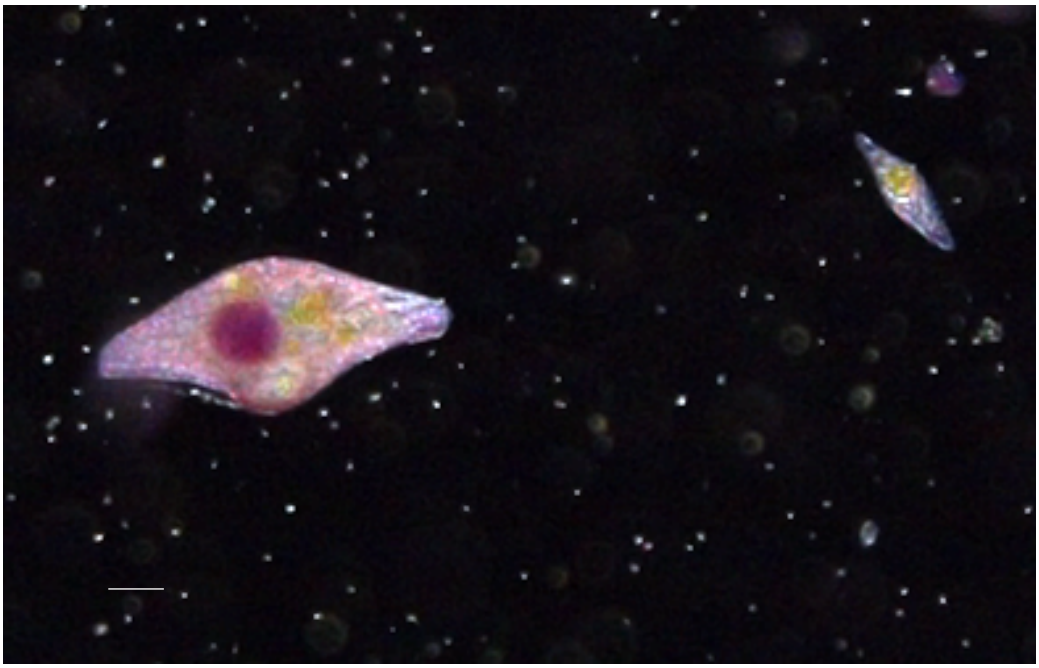


Figure 1.3. A *Blepharisma japonicum* cannibal giant (left) next to a normal-sized *B. japonicum* cells (top right). The small, reflective granules in the background are *Chlorogonium* algae, which are the food source for *Blepharisma*. Scale bar 100 μm. Transmitted light-darkfield.

While dividing vegetatively, *Blepharisma* cells consume bacteria, algae and even smaller ciliates for nourishment (Giese 1973). Under conditions of high density and starvation, they can also resort to consuming smaller conspecifics giving rise to a cannibal phenotype which is much larger than normal cells (Giese 1938) (Figure 1.3). These cannibal giants are very conspicuous in a cell culture and can be easily identified by the intensely colored, magenta vacuoles, which are the digested remains of their conspecifics (Figure 1.3).

Conditions of starvation can also induce pair-formation and conjugation in *Blepharisma,* which marks the onset of sexual reproduction (Miyake and Beyer 1973). Conjugation in *Blepharisma* is mediated by pheromone-like substances called gamones. It is one of only two ciliate genera, along with *Euplotes* where conjugation has been shown to be mediated through soluble factors like gamones (Katashima, 1959; Kimball, 1942; Luporini et al., 1983; Vallesi et al., 1995). *Blepharisma* has two mating types, distinguished by their gamone production. Mating type I cells release gamone 1, a ~30 kDa glycoprotein (Miyake and Beyer, 1974; Sugiura and Harumoto, 2001); mating type II cells release gamone 2, formally calcium-3-(2'- formylamino-5'-hydroxybenzoyl) lactate, a small-molecule effector (Kubota et al., 1973).

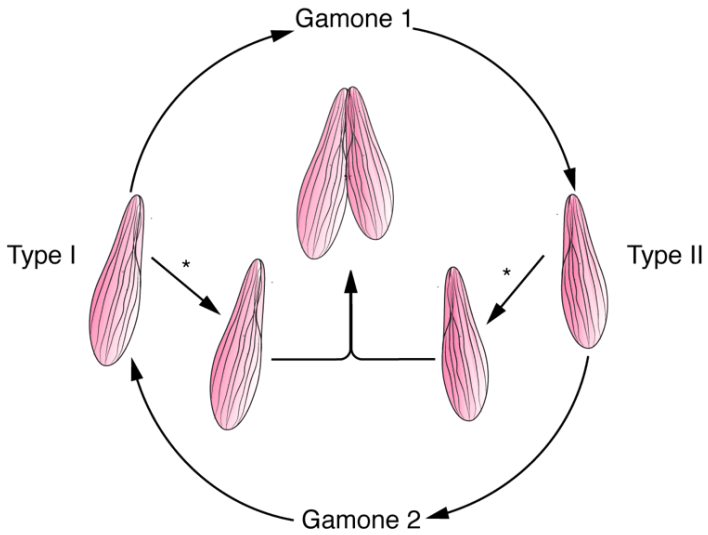


Figure 1.4. Starvation induces cells to produce gamones. Mating type I cells produce of Gamone 1, which in turn induces mating type II cells to produce of Gamone 2. Gamone 2 induces and upregulation in the production of Gamone 1, which then also causes an upregulation in the production of Gamone 2. Cells exposed to the gamone produced by the opposite mating type are primed for pair formation (*) and can form conjugating pairs. Figure adapted from Miyake & Beyer, 1973.

Complementary mating types have been isolated for several species of *Blepharisma* such as *B. stoltei, B. japonicum, B. americanum* and *B. undulans*. The *Blepharisma stoltei* strains used in the present study, were originally isolated in Germany (strain ATCC 30299) and Japan (strain HT-IV), with the former continuously cultured for over fifty years (Repak 1968), and the latter for over a decade (personal communication, Terue Harumoto). The *B. stoltei* ATCC 30299 strain is of mating type I and the *B. stoltei* HT-IV strain is of mating type II.

Blepharisma cells commit to conjugation when complementary mating types recognize and respond to each other's gamones, the production of which is induced and upregulated in a reciprocative feedback loop (Figure 1.4). Cell pairs involving cells from complementary mating types are called heterotypic pairs. Meiosis and recombination of gametic nuclei occurs in these heterotypic pairs leading to the formation of new MACs.

Once *Blepharisma* cells form heterotypic pairs, conjugation progresses and is evident though a through the intricate sequence of nuclear morphological changes seen in the cells. During this process, some of the MICs in each of the cells undergo meiosis (meiotic MICs) and the rest do not (somatic MICs) (Figure 1.5). A meiotic MICs eventually gives rise to two haploid gametic nuclei. A gametic MIC (the migratory nucleus) from each conjugating cell is exchanged with that of its partner. Subsequent fusion of the migratory and stationary haploid nuclei generates a zygotic nucleus (synkaryon), and after successive mitotic divisions gives rise to the

new MICs, some of which develop into the new MACs (primary macronuclear anlagen). Primary macronuclear anlagen (MA) continue to mature, eventually growing in size and DNA content (Miyake et al., 1991). These nuclear processes are shown in a schematic in Figure 1.5 and are described in further detail in Chapter 4.

It is in the macronuclear anlagen, that the IESs are removed from the developing MAC. As the anlagen develop and grow larger in size, the conjugating cells dissociate. These "ex-conjugants" now possess an entirely new set of MICs and a new developing MAC, with a different genotype from the old MAC. The old MAC of both the cells is degraded during conjugation. The exconjugants, usually containing two primary MAs undergo a series of up to three cytokinetic divisions, giving rise to two cells with one MA each, and subsequently to more cells with a new and fully formed MAC.

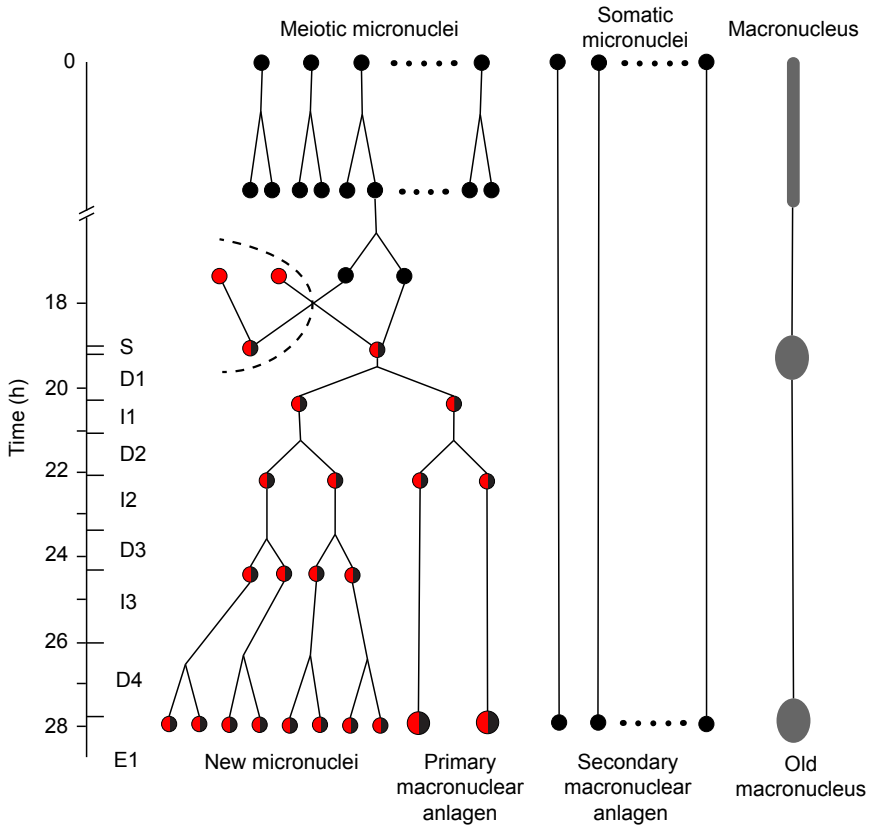


Figure 1.5. Schematic of nuclear processes occurring during conjugation (classified according to, and modified from (Miyake et al., 1991)). Nuclear events occurring before and up to, but not including fusion of the gametic nuclei (syngamy) are classified into sixteen pre-gamic stages where the MICs undergo meiosis and the haploid products of meiotic MICs are exchanged between the conjugating cells, followed by karyogamy. After karyogamy, cells are classified into 10 stages S (synkaryon), D1 (1st mitosis), I1 (1st interphase), D2 (2nd mitosis), I2 (2nd interphase), D3 (3rd mitosis), I3 (3rd interphase), D4 (4th mitosis), E1 (1st embryonic stage), E2 (2nd embryonic stage, not shown in diagram). After E2, the exconjugants divide further and are classified into 6 stages of cell division (CD1-6) not shown here.

MAC development in *Blepharisma* can also occur through a secondary pathway, which is predominantly seen in strains with high selfing frequency (conjugation among cells within a clonal population). In this pathway, MICs which have not undergone meiosis can give rise to secondary anlagen, which can develop into mature macronuclei (Miyake, Rivola, and Harumoto 1991; Suzuki 1957). This form of quasi-parthenogenetic reproduction, known as "apomixis", allows germline nuclei to give rise to the somatic nucleus, foregoing meiosis, the entire haploid phase, karyogamy and the divisions of the synkaryon (Figure 1.5). A secondary pathway of MAC development been observed in only one other ciliate, *Paramecium putrinum* (Jankowski 1962).

1.3. Bibliography

Arnaiz, Olivier, Nathalie Mathy, Céline Baudry, Sophie Malinsky, Jean-Marc Aury, Cyril Denby Wilkes, Olivier Garnier, et al. 2012. "The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences." Edited by Harmit S. Malik. *PLoS Genetics* 8 (10): e1002984. https://doi.org/10.1371/journal.pgen.1002984.

Aury, Jean-Marc, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M. Porcel, Béatrice Ségurens, et al. 2006. "Global Trends of Whole-Genome Duplications Revealed by the Ciliate *Paramecium* Tetraurelia." *Nature* 444 (7116): 171–78. https://doi.org/10.1038/nature05230.

Baudry, Céline, Sophie Malinsky, Matthieu Restituito, Aurélie Kapusta, Sarah Rosa, Eric Meyer, and Mireille Bétermier. 2009. "PiggyMac, a Domesticated PiggyBac Transposase Involved in Programmed Genome Rearrangements in the Ciliate *Paramecium* Tetraurelia." *Genes & Development* 23 (21): 2478–83. https://doi.org/10.1101/gad.547309.

Bischerour, Julien, Simran Bhullar, Cyril Denby Wilkes, Vinciane Régnier, Nathalie Mathy, Emeline Dubois, Aditi Singh, et al. 2018. "Six Domesticated PiggyBac Transposases Together Carry out Programmed DNA Elimination in *Paramecium*." *ELife* 7 (September): 1–24. https://doi.org/10.7554/eLife.37927.

Chen, Xiao, John R. R. Bracht, Aaron David David Goldman, Egor Dolzhenko, Derek M. M. Clay, Estienne C. C. Swart, David H. H. Perlman, et al. 2014. "The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development." *Cell* 158 (5): 1187–98. https://www.sciencedirect.com/science/article/pii/S0092867414009842.

Cheng, Chao-Yin, Alexander Vogt, Kazufumi Mochizuki, and Meng-Chao Yao. 2010. "A Domesticated PiggyBac Transposase Plays Key Roles in Heterochromatin Dynamics and DNA Cleavage during Programmed DNA Deletion in *Tetrahymena* Thermophila." Edited by Kerry S. Bloom. *Molecular Biology of the Cell* 21 (10): 1753–62. https://doi.org/10.1091/mbc.e09-12-1079.

Cheng, Chao-Yin, Janet M Young, Chih-Yi Gabriela Lin, Ju-Lan Chao, Harmit S Malik, and Meng-Chao Yao. 2016. "The PiggyBac Transposon-Derived Genes TPB1 and TPB6 Mediate Essential Transposon-like Excision during the Developmental Rearrangement of Key Genes in *Tetrahymena* Thermophila." https://doi.org/10.1101/gad.290460.

Duret, Laurent, Jean Cohen, Claire Jubin, Philippe Dessen, Jean François Goût, Sylvain Mousset, Jean Marc Aury, et al. 2008. "Analysis of Sequence Variability in the Macronuclear DNA of *Paramecium* Tetraurelia: A Somatic View of the Germline." *Genome Research* 18 (4): 585–96. https://doi.org/10.1101/gr.074534.107.

Eisen, Jonathan A., Robert S. Coyne, Martin Wu, Dongying Wu, Mathangi Thiagarajan, Jennifer R. Wortman, Jonathan H. Badger, et al. 2006. "Macronuclear Genome Sequence of the Ciliate *Tetrahymena* Thermophila, a Model Eukaryote." *PLoS Biology* 4 (9): 1620–42. https://doi.org/10.1371/journal.pbio.0040286.

Feng, Lifang, Guangying Wang, Eileen P. Hamilton, Jie Xiong, Guanxiong Yan, Kai Chen, Xiao Chen, et al. 2017. "A Germline-Limited PiggyBac Transposase Gene Is Required for

Precise Excision in *Tetrahymena* Genome Rearrangement." *Nucleic Acids Research* 45 (16): 9481–9502. https://doi.org/10.1093/nar/gkx652.

Flickinger, C. J. 1965. "The Fine Structure of the Nuclei of *Tetrahymena* Pyriformis throughout the Cell Cycle." *The Journal of Cell Biology* 27 (3): 519–29. https://doi.org/10.1083/JCB.27.3.519.

Giese, Arthur C. 1938. "Cannibalism and Gigantism in *Blepharisma*." *Transactions of the American Microscopical Society* 57 (3): 245. https://doi.org/10.2307/3222693.

———. 1973. *Blepharisma: The Biology of a Light-Sensitive Protozoan*. Stanford University Press. https://books.google.de/books?id=5S6sAAAAIAAJ.

Hamilton, Eileen P, Aurélie Kapusta, Piroska E Huvos, Shelby L Bidwell, Nikhat Zafar, Haibao Tang, Michalis Hadjithomas, et al. 2016. "Structure of the Germline Genome of *Tetrahymena* Thermophila and Relationship to the Massively Rearranged Somatic Genome." *ELife* 5 (November). https://doi.org/10.7554/elife.19090.

Herrick, Glenn, Samuel Cartinhour, Dean Dawson, Deborah Ang, Rebecca Sheets, Alice Lee, and Kevin Williams. 1985. "Mobile Elements Bounded by C4A4 Telomeric Repeats in *Oxytricha* Fallax." *Cell* 43 (3): 759–68. https://doi.org/10.1016/0092-8674(85)90249-1.

Jahn, Carolyn L., Stella Z. Doktor, John S. Frels, John W. Jaraczewski, and Mark F. Krikau. 1993. "Structures of the Euplotes CrassusTec1 and Tec2 Elements: Identification of Putative Transposase Coding Regions." *Gene* 133 (1): 71–78. https://doi.org/10.1016/0378-1119(93)90226-S.

Jankowski, A. W. 1962. "Conjugation Processes in *Paramecium* Putrinum. Clap. et Lachm. II. Apomictic Reorganization Cycles and the System of Mixotypes.:" *Tsitologiya* 4: 434–44.

Katz, Laura A., and Alexandra M. Kovner. 2010. "Alternative Processing of Scrambled Genes Generates Protein Diversity in the Ciliate Chilodonella Uncinata." *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 314B (6): 480–88. https://doi.org/10.1002/jez.b.21354.

Keeling, Patrick J., and Fabien Burki. 2019. "Progress towards the Tree of Eukaryotes." *Current Biology* 29 (16): R808–17. https://doi.org/10.1016/J.CUB.2019.07.031.

Klobutcher, Lawrence A., and Glenn Herrick. 1997. "Developmental Genome Reorganization in Ciliated Protozoa: The Transposon Link." *Progress in Nucleic Acid Research and Molecular Biology* 56 (April): 1–62. https://doi.org/10.1016/s0079-6603(08)61001-6.

Miyake, Akio, and J. Beyer. 1973. "Cell Interaction by Means of Soluble Factors (Gamones) in Conjugation of *Blepharisma* Intermedium." *Experimental Cell Research* 76 (1): 15–24. https://doi.org/10.1016/0014-4827(73)90413-8.

Miyake, Akio, Valeria Rivola, and Terue Harumoto. 1991. "Double Paths of Macronucleus Differentiation at Conjugation in *Blepharisma* Japonicum." *European Journal of Protistology* 27 (2): 178–200. https://doi.org/10.1016/S0932-4739(11)80340-8.

Orias, Eduardo. 1991. "Evolution of Amitosis of the Ciliate Macronucleus: Gain of the Capacity to Divide." *The Journal of Protozoology* 38 (3): 217–21. https://doi.org/10.1111/J.1550-

7408.1991.TB04431.X.

Prescott, David M. 1994. "The DNA of Ciliated Protozoa." *Microbiological Reviews* 58 (2): 233–67. http://www.ncbi.nlm.nih.gov/pubmed/8078435.

Repak, Arthur J. 1968. "Encystment and Excystment of the Heterotrichous Ciliate *Blepharisma* Stoltei Isquith." *Journal of Protozoology* 5: 407–12.

Sheng, Yalan, Lili Duan, Ting Cheng, Yu Qiao, Naomi A. Stover, and Shan Gao. 2020. "The Completed Macronuclear Genome of a Model Ciliate *Tetrahymena* Thermophila and Its Application in Genome Scrambling and Copy Number Analyses." *Science China Life Sciences* 63 (10): 1534–42. https://doi.org/10.1007/s11427-020-1689-4.

Suzuki, S. 1957. "Parthenogenetic Conjugation in *Blepharisma* Undulans Japonicus Suzuki." *Bulletin of Yamagata University of Natural Sciences* 4: 69–84.

Swart, Estienne C., John R. Bracht, Vincent Magrini, Patrick Minx, Xiao Chen, Yi Zhou, Jaspreet S. Khurana, et al. 2013. "The *Oxytricha* Trifallax Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes." Edited by Jonathan A. Eisen. *PLoS Biology* 11 (1): e1001473. https://doi.org/10.1371/journal.pbio.1001473.

Tucker, J. B., J. Beisson, D. L.J. Roche, and J. Cohen. 1980. "Microtubules and Control of Macronuclear 'amitosis' in *Paramecium*." *Journal of Cell Science* 44: 135–51. https://doi.org/10.1242/JCS.44.1.135.

Wang, Jianbin, and Richard Davis. 2014. "Programmed DNA Elimination in Multicellular Organisms." *Current Opinion in Genetics & Development* 27 (August): 26–34. https://doi.org/10.1016/J.GDE.2014.03.012.

Williams, K., T.G. Doak, and G. Herrick. 1993. "Developmental Precise Excision of *Oxytricha* Trifallax Telomere-Bearing Elements and Formation of Circles Closed by a Copy of the Flanking Target Duplication." *The EMBO Journal* 12 (12): 4593–4601. https://doi.org/10.1002/j.1460-2075.1993.tb06148.x.

Woodard, John, Edna Kaneshiro, and Martin A. Gorovsky. 1972. "Cytochemical Studies on the Problem of Macronuclear Subnuclei in *Tetrahymena*." *Genetics* 70 (2): 251–60. https://doi.org/10.1093/genetics/70.2.251.

Chapter 2
Research Aims

The basal position of *Blepharisma* on the ciliate tree make it an ideal model ciliate to study the emergence of genome reorganization pathways and determine the degree of conservation observed in these mechanisms from *Blepharisma* to the model ciliates *Paramecium*, *Tetrahymena* and *Oxytricha*.

This thesis aimed to achieve the following objectives:

1. Generation of annotated draft genomes of the macronucleus and micronucleus of *Blepharisma stoltei* (strain ATCC 30299)
2. Transcriptome sequencing of cells at multiple timepoints across conjugation and development of the new MAC
3. Investigation of the genome rearrangement properties and potential responsible molecules in *Blepharisma*

This thesis presents the work performed to achieve the objectives listed above. Chapter 3 introduces the methods and considerations underlying sequencing, assembly, structural annotation and functional annotation of the macronuclear genome. Chapter 4 addresses conjugation and sexual reproduction in *B. stoltei* and the utilization of the developmental time series to generate RNA-seq data. The RNA-seq data together with the annotated genome of *B. stoltei* ATCC 30299 provides a glimpse of the gene expression patterns during development of the new MAC for various genes known to be involved in the genome reorganization machinery in ciliates. Chapter 5 focuses on a specific class of transposase found in the *B. stoltei* MAC and MIC genomes, encoding the PiggyBac transposase domain. One of these homologs presents a strong case for being the main IES excisase in *Blepharisma*. Chapter 6 presents the collaborative work performed on the MIC genome of *B. stoltei* ATCC30299 and discusses the MIC-limited transposons and the non-autonomous miniature inverted-repeat transposable element (MITEs) which have emerged from them. Chapter 7 presents a general discussion of Chapters 1 - 6 and provides a brief comment on future work with *Blepharisma* as a model ciliate. Chapter 8 lists all the materials and methods used in the work described in Chapters 3-6.

Chapter 3

Preliminary assembly and annotation of the draft genome of the *Blepharisma stoltei* somatic nucleus

3.1. Introduction

The genome architecture of the two ciliate nuclei is of particular interest, as the somatic genome is actively shaped during sexual reproduction. This occurs through a process of genome reorganization, where large portions of the germline genome are eliminated from the developing somatic nucleus. *Blepharisma* is a ciliate belonging to the class of Heterotrichs, which diverged from the rest of the ciliate classes at an early stage. Studying genome reorganization in *Blepharisma* thus has the potential to illuminate the ancestral state of this process. However, the current lack of a reference genome for the somatic nucleus of *Blepharisma* has limited investigations into its genome architecture and molecular mechanisms. We aim to address this gap by generating a reference genome for the *Blepharisma* somatic nucleus.

Sequencing technology for sequencing entire genomes has developed fast and has come far since the initiation of the Human Genome Project in 1990. One of the key developments from this project was "shotgun sequencing", a technique which Next Generation Sequencing (NGS) relies on today. This involves breaking strands of DNA into smaller fragments at random points, sequencing these short fragments and then computationally assembling the short sequences to recreate the original sequences (Venter et al. 1998). At the most basic level, a genome assembly involves organizing the sequenced DNA fragments in the form of contiguous sequences, known as “contigs”, and further into scaffolds and ultimately chromosomes (Hunt et al. 2014). The complexity of the assembly increases if there are many regions that are similar (DNA repeats, alleles), if the reads are too short or contain too many errors (inaccurate reads) or if there are not enough overlapping reads representing a DNA fragment (low coverage).

Once a genome assembly is produced, it can be used to determine the genes encoded within it. Demarcating regions of the genome that code for proteins is known as structural annotation or gene prediction, while assigning indications of biological function to these predicted genes is known as functional annotation. The genetic code for a genome plays an integral role in structural gene annotation, as it allows the determination of the longest open reading frames. Non-standard nuclear genetic codes are most often observed in eukaryotes

among the ciliates. *Blepharisma* has a non-standard genetic code, where UGA is translated as tryptophan, instead of serving as a stop codon (Liang and Heckmann 1993). This is unlike the ciliate genetic code utilized by most other ciliates like *Paramecium, Tetrahymena and Oxytricha,* which only have one stop codon, UGA, and translate the canonical stop codons UAA and UAG as Glutamine (Lozupone, Knight, and Landweber 2001). The non-standard genetic code of *Blepharisma* is another factor to be accounted for during annotation.

The quality of the sequencing data, i.e., the raw reads, depends on the quality of DNA used for sequencing (Chen et al., 2017). Short-read sequencing produces reads of 100-150 base pairs (bp), which are highly accurate (Chen et al., 2017). In contrast, long-read sequencing can produce reads longer than 10 kilobase pairs (kbp) but were previously prone to sequencing errors (Rhoads and Au 2015). With one popular long-read sequencing technology from Pacific Biosciences, this shortcoming has been addressed by recent advances in the High Fidelity (HiFi) sequencing format, which generates long reads built from multiple circular consensus reads, which have an accuracy comparable to that of short-reads (Wenger et al. 2019). To benefit from the extended read lengths now possible with long-read sequencing, genomic DNA must have a high molecular weight, i.e. correspond to long sequences. If the DNA is excessively fragmented or degraded before sequencing, the quality of the sequenced reads will also be compromised. Moreover, if the read quality is poor, then any assembly generated using those reads will reflect this lack of accuracy. DNA extraction is therefore a critical step for quality control.

In this section, I present the exploration of suitable experimental protocols for DNA extraction from the somatic nucleus, and several iterations and variations of DNA sequencing and genome assembly performed before arriving at the final reference assembly of the *Blepharisma stoltei* somatic genome. The *B. stoltei* lab strain ATCC 30299, first isolated from Federsee (Germany) (Repak 1968), was used for all experiments and will henceforth be referred to as simply "*B. stoltei* ATCC". I also outline the iterative assessment and refinement of gene prediction models to address challenges in annotating an atypical eukaryotic genome. Finally, I describe the functional annotation of the reference genome. The genome assembly, augmented with structural and functional annotations, constitutes an essential resource for further analysis of *Blepharisma*'s biological pathways.

3.2. Results

3.2.1. DNA extraction

3.2.1.1. DNA extraction from whole-cell lysate

The macronucleus (MAC) of *B. stoltei* contains most of the DNA present in a single cell. In *B. stoltei*, the MAC is ribbon-like in shape, measuring 150-200 µm on its longer axis and 10-20 µm in diameter. It is several times larger than the small germline nuclei, called micronuclei (MIC), which are spherical and measure only 1-2 µm in diameter (Figure 3.1).

Figure 3.1. *Blepharisma stoltei* ATCC cell. Magenta regions indicate cytoplasm and ciliary structures on the cell surface. Cyan regions indicate nuclear material stained with DAPI. Small, round cyan bodies are micronuclei; large ribbon-like cyan structure is macronucleus. Scale bar: 50 µm.

The amount of DNA in the MAC considerably outweighs that in the MIC (Figure 3.1). Therefore, in a lysate prepared by digesting entire cells, the majority of DNA will be from the MAC, with very little originating from the MIC. MAC DNA extracted from the whole-cell lysates was purified using two different methods: a spin column or spool out of a solution.

The size distribution of DNA obtained using these methods can be seen in a profile generated using pulsed-field capillary electrophoresis (Femto Pulse, Agilent) as shown for the two different methods: isolation using a spin column (Figure 3.2) and isolation by spooling precipitated DNA out of the solution (Figure 3.3) (Chapter 8, section 8.2).

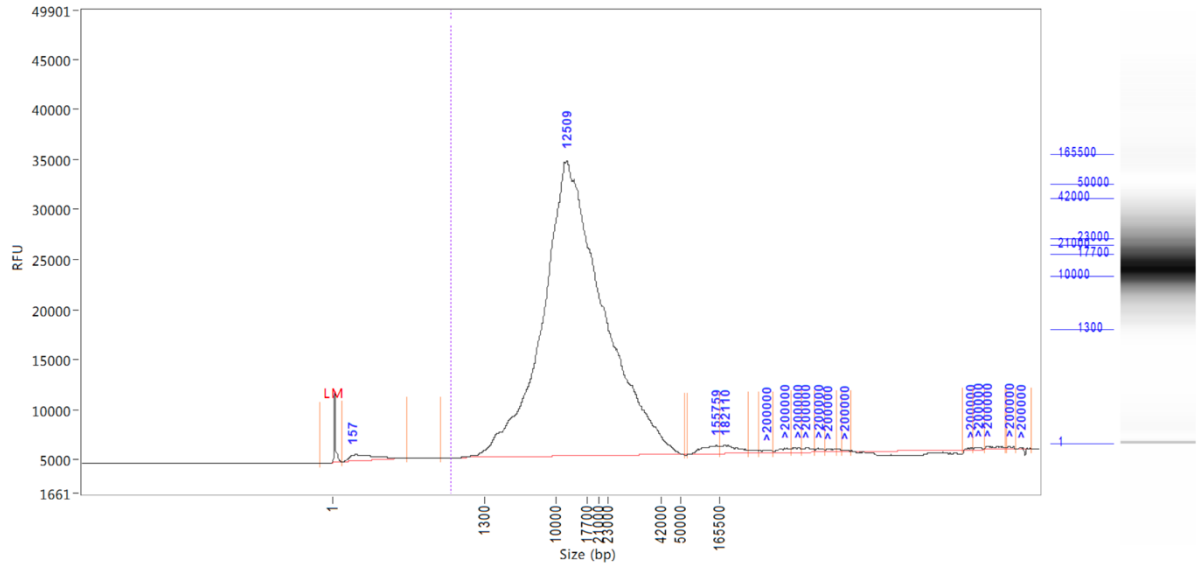


Figure 3.2. Column-purified *B. stoltei* ATCC DNA from whole-cell lysate. X-axis indicates size of DNA fragments. Y-axis indicates relative fluorescence units (RFU) as a measure of frequency of fragments of a particular size.

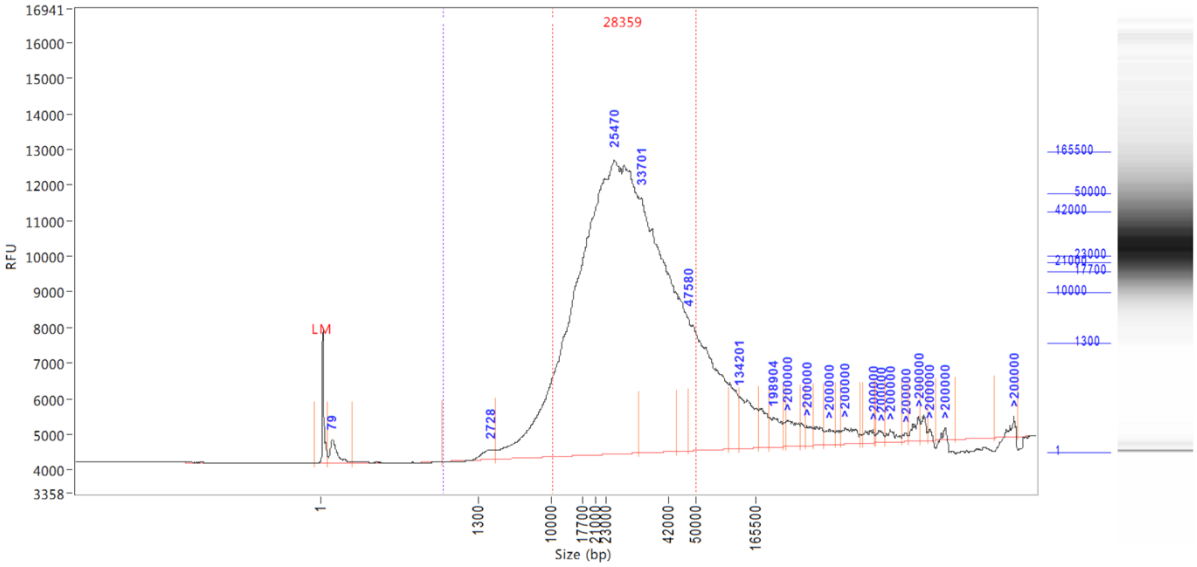
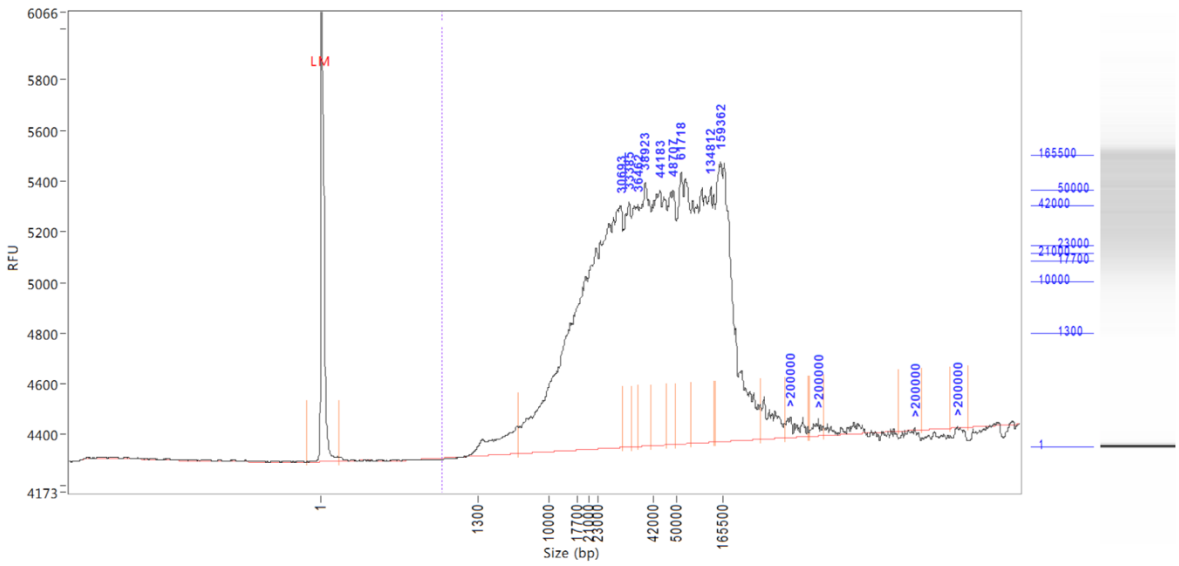


Figure 3.3. Phenol-chloroform-purified *B. stoltei* ATCC DNA from whole-cell lysate. X-axis indicates size of DNA fragments. Y-axis indicates relative fluorescence units (RFU) as a measure of frequency of fragments of a particular size.

The modal fragment length of DNA extracted using both these methods is larger than 10 kb, demonstrating that high-molecular weight DNA suitable for both short-read and long-read sequencing can be isolated from *B. stoltei*.

3.2.1.2. *DNA extraction from MAC-enriched lysate*

In addition to whole-cell lysate, DNA was extracted from purified macronuclei. The size difference between the MAC and MIC allows the two organelles to be separated by density gradient centrifugation. This allows nuclei with intact nuclear membranes to be distributed at different heights of a solution with an increasing density gradient from top to bottom. Upon centrifugation, the heavier and larger MACs settle to the bottom of the gradient as a pellet and the lighter and smaller MICs, together with fragments of broken MACs, remain in the top layer of the solution. Fractions enriched in MACs and MICs can thus be obtained by harvesting the nuclei from the bottom and top of the centrifuged gradient, respectively (Chapter 8, section 8.3). DNA from the MAC-enriched fraction was purified using gravity flow-based purification columns. Figure 3.4 shows the fragment size profile, where the modal fragment size from this preparation is well above 21 kb, thus making this sample well suited for long-read sequencing. This sample was sequenced using PacBio HiFi reads, which are highly accurate and used to



construct the MAC genome assembly.

Figure 3.4. *B. stoltei* ATCC MAC-enriched fraction isolated using gravity-flow purification columns. X-axis indicates size of DNA fragments. Y-axis indicates relative fluorescence units (RFU) as a measure of frequency of fragments of a particular size.

3.2.2. Assembly

3.2.2.1. Short-read assembly

For a preliminary assembly, DNA extracted from a whole-cell lysate was sequenced as 150 bp paired-end Illumina reads. These reads were either trimmed to exclude regions with a PHRED quality score lower than 28 (referred in Table 3.1 as "q28") or left untrimmed (referred in Table 3.1 as "untrimmed") and assembled using the SPAdes genome assembler (Bankevich et al. 2012). This yielded a 44.7 Mb genome with 2840 contigs (Table 3.1) and the length of the longest contig of this assembly was 0.81 Mb. This genome was assessed based on several metrics of genome quality, principal among which were the number of contigs, the N50 and L50. N50 is the contig length such that using longer or equal length contigs accounts for at least half (50%) of the total length of the assembly (Earl et al. 2011). This important metric indicates how well the reads have assembled. The N50 for this assembly was 0.32 Mb, demonstrating that at least half of the contigs in this assembly were longer than 0.3 Mb. Another metric, the L50, indicates the fragmentation of the assembly. It constitutes the minimum number of contigs that produce half (50%) of the bases of the assembly, i.e. the number of contigs of length at least N50 (Earl et al. 2011). The L50 of this assembly was 47, showing that at least half of the assembly could be re-constructed with just 47 contigs, all of which are at least 0.3Mb long (Table 3.1).

Short-read sequencing was also performed for DNA extracted from the whole-cell lysate of other *Blepharisma* species, *B. japonicum* and *B. undulans* and another strain of *B. stoltei* (HT-IV). SPAdes genomes were also prepared for these *Blepharisma* species and strains (Table 3.1).

Table 3.1. Assessment of SPAdes genome assemblies of *Blepharisma* species from Illumina reads.

Species	*B. stoltei* ATCC 30299 (q28)	*B. stoltei* HT-IV (q28)	*B. japonicum* (q28)	*B. undulans* (un-trimmed)
Number of contigs	2840	9229	1559	30724
Total length (Mb)	44.7	52.9	47.3	82.5
Largest contig (Mb)	0.81	0.54	0.97	0.15
N50 (Mb)	0.32	0.02	0.33	0.005
L50	47	493	47	3403

3.2.2.2. Long-read assembly

The DNA from the whole-cell lysate of *B. stoltei* was also sequenced with PacBio long-reads. These long reads served as the basis for genome assembly, using the long-read genome assembler Ra (Vaser and Šikić 2019). The genome assembled by Ra was 44.8 Mb in length. It was more contiguous than the SPAdes assembly, as it consisted of only 137 contigs. The length of the longest contig of the Ra assembly was 1.19 Mb, the N50 was 0.63 Mb and the L50 was 28 (Table 3.2), all indicating that the long-read assembly was better assembled and more contiguous than the short-read SPAdes assembly.

The long-read assembler Raven (Vaser and Šikić 2021), a successor of Ra, was also used to assemble the *B. stoltei* long-reads (Chapter 8, section 8.4.1). This assembly was 44.4 Mb in length and consisted of 118 contigs. Its N50 and L50 were 0.79 Mb and 22, respectively. Another long-read assembler, Flye (Kolmogorov et al. 2019), performed even better, generating a 41.9 Mb genome with 91 contigs and an N50 and L50 of 0.75 and 24. Based on optimized parameters observed from the previous assemblies, the Flye assembly was produced by Dr. Estienne Swart using HiFi reads, long-reads of very high accuracy obtained from the MAC-enriched DNA (Chapter 6). While shorter than other assemblies, it was more contiguous and was therefore used as the basis for the reference genome. This Flye assembly was further manually curated to produce the final assembly of the *B. stoltei* somatic genome (Chapter 8, section 8.4.2).

Table 3.2. Assessment of genome assemblies of *B. stoltei* ATCC. The reference genome is a derivative of the Flye assembly (Chapter 5).

Statistics	SPAdes short-read assembly	Ra long-read assembly	Raven long-read assembly	Flye long-read assembly	Reference genome
Number of contigs	2840	137	118	91	64
Total length (Mb)	44.7	44.8	44.4	41.9	41.4
Largest contig (Mb)	0.81	1.19	1.7	1.2	1.51
N50 (Mb)	0.32	0.63	0.79	0.75	0.78
L50	47	28	22	24	23

3.2.3. Annotation

3.2.3.1. Structural annotation with AUGUSTUS

Demarcating regions of the genome that code for proteins, i.e., genes and non-coding regions such as introns, is known as structural annotation. The eukaryotic gene prediction software AUGUSTUS (Hoff and Stanke 2019), which remains best in class for eukaryotic gene prediction (Scalzitti et al. 2020), was used for structural annotation of the *B. stoltei* genome. AUGUSTUS provides pre-trained gene prediction models for several eukaryotes, among them the ciliate *Tetrahymena*. *Tetrahymena* and *Blepharisma* differ greatly in genome architecture, have a large evolutionary distance between them and use different genetic codes. The *Tetrahymena* gene model was therefore not expected to be very effective at gene prediction in *Blepharisma* and was used to evaluate the effectiveness of an un-trained model of gene prediction for *Blepharisma*.

To assess the performance of the gene prediction using various models, a set of 284 coding regions (CDSs) and 103 introns were manually curated in the *B. stoltei* genome, with the help of transcriptomic data mapped to the genome assembly. This set was split into two equal subsets, one to serve as a training dataset and the other as a testing dataset. It was already evident from the transcriptome mapping to the assembly, that the *Blepharisma* somatic genome had some unusual properties. The 103 manually annotated introns were all 15-16 bp in length, much like the ciliate *Stentor coeruleus* (Slabodnick et al. 2017), a fellow heterotrich and close relative of *Blepharisma.* Moreover, the genic features were packed densely, with short or negligible intergenic regions. An initial round of gene prediction using the *Tetrahymena* gene model to predict features in the testing dataset of the *Blepharisma* MAC genome showed an underwhelming exon-level sensitivity of 4.6% and specificity of 8.3% (Table 3.3). Sensitivity and specificity are measures of the rate of identification of true positives and true negatives, respectively, in this case indicating the percentage of exonic and non-exonic regions AUGUSTUS was able to correctly identify.

$$Sensitivity = \frac{True\ positives}{True\ positives + \ False\ neagtives}$$

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives}$$

Gene predictions made by AUGUSTUS can be improved by supplementing the gene prediction model with the location of exons and introns inferred from RNA-seq data. These "hints" from transcriptomic data allowed the gene prediction accuracy using the *Tetrahymena* gene model to climb to 8.0% and 9.5% for sensitivity and specificity, respectively.

Due to the demonstrable insufficiency of the *Tetrahymena* gene model, an *ab initio* gene prediction model was required for *Blepharisma*. AUGUSTUS can be re-trained under supervision to develop gene models for other species *ab initio*. A gene prediction model has been developed for *Stentor* (Slabodnick et al. 2017), which was graciously shared by the authors (Pranidhi Sood, personal communication). This model is based on a modified version of AUGUSTUS, which made allowances for unusual, short introns of *Stentor*, similar to those of *Blepharisma*. Benchmarking the performance of the *Stentor* model against the *Blepharisma* test dataset showed that exon-level sensitivity was 72.4% and specificity was 75.9%, a dramatic improvement on the previous model. Supplied with "hints" from RNA-seq, the *Stentor* model predicted *Blepharisma* exons with a sensitivity and specificity of 86.2% and 82.4%, respectively.

Table 3.3. AUGUSTUS parameters and respective gene prediction accuracy at the exon level. The parameters represent only the internal parameters for gene prediction models used by AUGUSTUS, while being trained and tested on *Blepharisma* data. All Sensitivity and specificity values shown here have been generated using *Blepharisma* genome data.

Parameters	Exon-level	
	Sensitivity (%)	Specificity (%)
Tetrahymena	4.6	8.3
Stentor	72.4	75.9
Blepharisma Set 1	65.5	68.7
Blepharisma Set 2	48.8	60.1
With RNA-seq hints		
Tetrahymena	8.0	9.5
Stentor	86.2	82.4
Blepharisma Set 1	79.3	75.0
Blepharisma Set 2	84.6	85.4

While gene predictions with the *Stentor* model were considerably more accurate than those produced with the *Tetrahymena* model, it remained to be seen whether these could be improved by training the model specifically for *Blepharisma*. The training set of manually curated *Blepharisma* features was used to train a version of AUGUSTUS, modified in the same way as for

Stentor. The first gene prediction model trained with *Blepharisma* features had an accuracy of 65.5% sensitivity and 68.7% sensitivity. This could be boosted with "hints" to 79.3% sensitivity and 75% specificity. To improve this further, more CDSs representing longer exons were added to both the training and testing datasets, since these longer exons were being annotated incorrectly as intergenic regions. This augmented dataset of features was used to train and test gene prediction.

The second model yielded a sensitivity of 48.8% and 60.1% and gene prediction with hints allowed a sensitivity and specificity of 84.6% and 85.4%, respectively. This gene model, though an improvement over previous versions, still had drawbacks. While it predicted almost all 15-bp introns correctly, it also predicted additional, spurious 15-bp introns in coding regions and intergenic regions, where there were no spliced- reads to indicate the presence of real 15-bp introns. It missed 16-bp introns with a higher frequency than it missed 15-bp introns and also predicted 17-20 bp and 300-450-bp introns, which were almost always incorrect predictions (as verified by inspecting the spliced-read alignments). To circumvent these limitations, a wrapper script for AUGUSTUS was created (https://github.com/Swart-lab/Intronarrator) to infer introns directly from RNA-seq data and to predict genes using the "intronless" mode in AUGUSTUS (Chapter 8, section 8.5.). This preserved the accuracy of intron annotation, while benefiting from the high sensitivity of exon prediction by AUGUSTUS. The final set of gene predictions for the *B. stoltei* MAC genome was generated using Intronarrator.

3.2.3.2. Functional annotation with HMMER3, Pannzer2 and eggNOG

The structurally annotated somatic genome of *B. stoltei* was used as a basis for functional annotation. Functional annotation aims to assign a biological significance to the genes identified through structural annotation. This involves translating the coding regions of the genes and allowing specialized software to compare these hypothetical proteins with a database of annotated homologs.

For the *B. stoltei* somatic genome, functional annotation was conducted with HMMER3 (Eddy 2011), using hmmscan, which searches protein sequences against HMM-profiles in the PFAM-A protein database. Of the 25710 genes annotated by Intronarrator in the MAC genome, 57.6% (14817) genes were assigned PFAM domain annotations by hmmscan.

Pannzer2 (Törönen, Medlar, and Holm 2018) and eggNOG (Huerta-Cepas et al. 2019) were used to provide further functional annotation. Pannzer2 (Protein ANNotation with Z-scoRE) performs homology searches to produce gene ontology (GO) annotations, together with a free-text description. eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) is database of orthology relationships and can be used to identify orthologs in a query genome and infer their function. Pannzer2 was able to assign gene names to 2.8% (737) genes in the MAC genome and provided a description for 23.9% (6154) genes. EggNOG performed marginally better, by assigning gene names to 18.6% (4786) genes and providing descriptions for 49% (12598) annotated genes.

3.3. Discussion

The procedures described here generated the predecessors of the final somatic genome assembly and mode of structural annotation (Chapter 5), paving the way to address the principal research questions described in Chapter 2. The somatic genomes of model ciliates *Paramecium tetraurelia* (Aury et al. 2006), *Tetrahymena thermophila* (Eisen et al. 2006), *Oxytricha trifallax* (Swart et al. 2013), *Stylonychia lemnae* (Aeschlimann et al. 2014), *Stentor coeruleus* (Slabodnick et al. 2017), and *Euplotes octocarinatus* (Wang et al. 2018) have been sequenced in the last few years, with the increasing availability and decreasing cost of high-throughput sequencing. The addition of *Blepharisma* to this repertoire necessitated the optimization of several basic procedures to the specific demands of *Blepharisma* culture and biology.

3.3.1. High molecular weight DNA was extracted from MAC- and MIC-enriched fractions

We were able to extract high molecular weight DNA of *B. stoltei* from the whole-cell lysate, the MAC-enriched fraction and the MIC-enriched fraction of the cell lysate. Features particular to *Blepharisma*, such as the pigmented pellicular layer, not accounted for in standard protocols for DNA extraction from eukaryotic cells did not interfere with the process. The substantial difference between the physical dimensions of the MAC and the MIC is distinctive for *Blepharisma*, though there is a large amount of diversity in nuclear shapes and sizes observed among the ciliates. The germline genomes of *Paramecium* was sequenced (Arnaiz et al. 2012). using genetic material from the developing macronucleus of cells with a silenced IES-excisase (Pgm), using short-read sequencing (Arnaiz et al. 2012). This genetic material contained a mixture of un-modified germline DNA (60-65%) and IES-eliminated DNA (35-40%), due to the presence of old MACs in the cytoplasm (Arnaiz et al. 2012). The enrichment of MICs

directly from cell lysate avoids this particular type of contamination in the case of *Blepharisma*, though this is solely a theoretical consideration, given that gene silencing in *Blepharisma* has only been reported once and has since not been reproducibly demonstrated (Sobierajska et al. 2011). Moreover, the MIC-enriched fraction from the *Blepharisma* cell-lysate suffers from a different form of MAC contamination, which is caused by the fragmentation of the long and ribbonous MAC in the lysate into smaller fragments. These fragments separate in the MIC-rich fraction of the sucrose gradient. The sequencing of the MIC-enriched fractions showed that only ~20% of the raw reads were MIC-specific and the rest originated from the MAC (Chapter 6), which was nevertheless sufficient for assembling the MIC-limited regions (Chapter 6). Conversely, this indicates that the proportion of MIC-reads in an un-fractionated, whole-cell lysate is likely to be even smaller and its influence on the genome assembly generated from genetic material obtained from whole cell-lysates can thus be considered negligible.

3.3.2. Short-read assemblies were produced for two strains of B. stoltei

A comparison of short-read SPAdes assemblies for the two strains of *B. stoltei* (ATCC 30299 and HT-IV, isolated from Aichi prefecture, Japan) showed that the HT-IV assembly is larger than that of ATCC 30299. This indicates that the HT-IV assembly might be heterozygous, and the allelic reads may not be reconcilable, leading to a larger assembly. This also indicates that the somatic genome of *B. stoltei* strain ATCC is homozygous, a condition which may have resulted from selfing events over the course of laboratory propagation in the 50 years since its isolation in 1968. The short-read SPAdes assemblies for *B. stoltei* ATCC exhibited a higher genome size than the long-read assemblies generated by Ra and Raven. Regions of low complexity, such as telomeres or other repeats, are difficult to resolve from the short reads. Such low complexity regions can be better resolved through long-reads, where the contiguity of the longer DNA sequences helps to establish the contiguity of local segments of the assembly, accounting for the difference in genome sizes between the two modes of assembly.

3.3.3. Structural gene prediction was performed using AUGUSTUS

Most non-model organisms require *ab initio* gene prediction models for structural annotation, since the current repertoire of gene prediction models listed for popular gene prediction tools like AUGUSTUS over-represents animal metazoans, fungi, plants and green and red algae. The non-standard genetic code of *Blepharisma*, where the canonical stop codon UGA codes for tryptophan, further excluded gene prediction tools like GeneMark (Lomsadze et al.

2005), and dependent pipelines like BRAKER (Hoff et al. 2016) and MAKER (Holt and Yandell 2011). The *ab initio* gene prediction for *Blepharisma* involved changing hard-coded settings for parameters required by AUGUSTUS, e.g. minimum intron length whose default value is 39, reflecting the expected length of the shortest spliceosomal eukaryotic introns, ~40 nucleotides (Rogozin et al. 2012), outside the ciliates. This parameter was set to 9 to accommodate the short introns found in *Blepharisma*. Such parameters are challenging for gene prediction programs, which are based on statistical approaches. Most programs perform well in intermediate cases, but struggle to identify features at extreme ends of the spectrum - such as these very short introns (Scalzitti et al. 2020). These limitations served as the inspiration for Intronarrator, a wrapper script developed by our group (https://github.com/Swart-lab/Intronarrator) that invokes AUGUSTUS after inferring intronic regions directly from transcriptomic data, thus allowing a more accurate prediction of coding regions (Chapter 5). These small introns in the genome hint at the possible miniaturization of other non-coding DNA features, as is often observed in eukaryotic genomes (Cavalier-Smith 2005). Gene prediction in *Blepharisma* adds further support to this observation, as we find the *Blepharisma* somatic genome to be relatively gene dense. The *Blepharisma* somatic genome contains about 600 genes per one million base pairs, with only short intervening intergenic regions, in comparison to about 6 genes per million base pairs in humans.

3.3.4. Functional gene prediction was performed with HMMER3, Pannzer2 and eggNOG

Functional annotation of the *B. stoltei* MAC genome performed using HMMER3, Pannzer2 and eggNOG allowed protein domains, gene names and gene description to be assigned to a sizeable fraction of the MAC genome. Functional annotation of non-model species has limitations arising from the lack of close homologs in popular databases. Domain annotation using PFAM relies on recognition of domains curated by the PFAM database, which are predominantly of bacterial and opisthokont origin. Similar limitations apply to Pannzer2, which uses the Uniprot database (Bateman et al. 2015) to infer homologs (Törönen, Medlar, and Holm 2018) and EggNOG, which also relies on a combination of Uniprot, RefSeq and Ensembl databases (Huerta-Cepas et al. 2019) for determining protein homology and classification of CDSs into groups of orthologous proteins. The performance of eggNOG in assigning gene names (18.6%) and gene descriptions (49%) was better than that of Pannzer2 (2.8% gene names, 23% gene descriptions), which might be a consequence of the broader sampling of

multiple databases performed by eggNOG in comparison to the single database used by Pannzer2. Finally, in addition to these limitations of functional annotation based on protein homology, there still remains the experimental verification of gene function, which must eventually inform and complement these genome annotations.

The work described in this chapter provided a valuable first look at the genomic attributes of *Blepharisma* and acted as the antecedent of the final genome assembly and annotation described in the following chapters.

3.4. Bibliography

Aeschlimann, Samuel H., Franziska Jönsson, Jan Postberg, Nicholas A. Stover, Robert L. Petera, Hans-Joachim Lipps, Mariusz Nowacki, and Estienne C. Swart. 2014. "The Draft Assembly of the Radically Organized *Stylonychia Lemnae* Macronuclear Genome." *Genome Biology and Evolution* 6 (7): 1707–23. https://doi.org/10.1093/gbe/evu139.

Arnaiz, Olivier, Nathalie Mathy, Céline Baudry, Sophie Malinsky, Jean-Marc Aury, Cyril Denby Wilkes, Olivier Garnier, et al. 2012. "The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences." Edited by Harmit S. Malik. *PLoS Genetics* 8 (10): e1002984. https://doi.org/10.1371/journal.pgen.1002984.

Aury, Jean-Marc, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M. Porcel, Béatrice Ségurens, et al. 2006. "Global Trends of Whole-Genome Duplications Revealed by the Ciliate *Paramecium* Tetraurelia." *Nature* 444 (7116): 171–78. https://doi.org/10.1038/nature05230.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Bateman, Alex, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Rolf Apweiler, Emanuele Alpi, Ricardo Antunes, et al. 2015. "UniProt: A Hub for Protein Information." *Nucleic Acids Research* 43 (Database issue): D204. https://doi.org/10.1093/NAR/GKU989.

Cavalier-Smith, Thomas. 2005. "Economy, Speed and Size Matter: Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion." *Annals of Botany* 95 (1): 147–75. https://doi.org/10.1093/AOB/MCI010.

Chen, Lixin, Pingfang Liu, Thomas C. Evans, and Laurence M. Ettwiller. 2017. "DNA Damage Is a Pervasive Cause of Sequencing Errors, Directly Confounding Variant Identification." *Science* 355 (6326): 752–56. https://doi.org/10.1126/SCIENCE.AAI8690/SUPPL_FILE/CHEN-SM.PDF.

Earl, Dent, Keith Bradnam, John St. John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, et al. 2011. "Assemblathon 1: A Competitive Assessment of de Novo Short Read Assembly Methods." *Genome Research* 21 (12): 2224. https://doi.org/10.1101/GR.126599.111.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7 (10). https://doi.org/10.1371/journal.pcbi.1002195.

Eisen, Jonathan A., Robert S. Coyne, Martin Wu, Dongying Wu, Mathangi Thiagarajan, Jennifer R. Wortman, Jonathan H. Badger, et al. 2006. "Macronuclear Genome Sequence of the Ciliate *Tetrahymena* Thermophila, a Model Eukaryote." *PLoS Biology* 4 (9): 1620–42. https://doi.org/10.1371/journal.pbio.0040286.

Hoff, Katharina J., Simone Lange, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. 2016. "BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1." *Bioinformatics* 32 (5): 767–69.

https://doi.org/10.1093/bioinformatics/btv661.

Hoff, Katharina J., and Mario Stanke. 2019. "Predicting Genes in Single Genomes with AUGUSTUS." *Current Protocols in Bioinformatics* 65 (1): 1–54. https://doi.org/10.1002/cpbi.57.

Holt, Carson, and Mark Yandell. 2011. "MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects." *BMC Bioinformatics* 12 (1): 1–14. https://doi.org/10.1186/1471-2105-12-491/FIGURES/5.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. "EggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses." *Nucleic Acids Research* 47 (D1): D309–14. https://doi.org/10.1093/NAR/GKY1085.

Hunt, Martin, Chris Newbold, Matthew Berriman, and Thomas D. Otto. 2014. "A Comprehensive Evaluation of Assembly Scaffolding Tools." *Genome Biology* 15 (3): 1–15. https://doi.org/10.1186/GB-2014-15-3-R42/TABLES/4.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. "Assembly of Long, Error-Prone Reads Using Repeat Graphs." *Nature Biotechnology 2019 37:5* 37 (5): 540–46. https://doi.org/10.1038/s41587-019-0072-8.

Liang, A., and K. Heckmann. 1993. "*Blepharisma* Uses UAA as a Termination Codon." *Naturwissenschaften* 80 (5): 225–26. https://doi.org/10.1007/BF01175738.

Lomsadze, Alexandre, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. 2005. "Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm." *Nucleic Acids Research* 33 (20): 6494–6506. https://doi.org/10.1093/nar/gki937.

Lozupone, Catherine A., Robin D. Knight, and Laura F. Landweber. 2001. "The Molecular Basis of Nuclear Genetic Code Change in Ciliates." *Current Biology* 11 (2): 65–74. https://doi.org/10.1016/S0960-9822(01)00028-8.

Repak, Arthur J. 1968. "*Blepharisma* Stoltei." *Journal of Protozoology* 15 (3): 407–12.

Rhoads, Anthony, and Kin Fai Au. 2015. "PacBio Sequencing and Its Applications." *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89. https://doi.org/10.1016/J.GPB.2015.08.002.

Rogozin, Igor B., Liran Carmel, Miklos Csuros, and Eugene V. Koonin. 2012. "Origin and Evolution of Spliceosomal Introns." *Biology Direct* 7 (1): 1–28. https://doi.org/10.1186/1745-6150-7-11/FIGURES/6.

Scalzitti, Nicolas, Anne Jeannin-Girardon, Pierre Collet, Olivier Poch, and Julie D. Thompson. 2020. "A Benchmark Study of Ab Initio Gene Prediction Methods in Diverse Eukaryotic Organisms." *BMC Genomics* 21 (1). https://doi.org/10.1186/S12864-020-6707-9.

Slabodnick, Mark M., J. Graham Ruby, Sarah B. Reiff, Estienne C. Swart, Sager Gosai, Sudhakaran Prabakaran, Ewa Witkowska, et al. 2017. "The Macronuclear Genome of *Stentor Coeruleus* Reveals Tiny Introns in a Giant Cell." *Current Biology* 27 (4): 569–75. https://doi.org/10.1016/j.cub.2016.12.057.

Sobierajska, Katarzyna, Ewa Joachimiak, Cezary Bregier, Stanisław Fabczak, and Hanna Fabczak. 2011. "Effect of Phosducin Silencing on the Photokinetic Motile Response of *Blepharisma* Japonicum." *Photochemical and Photobiological Sciences* 10 (1): 19–24. https://doi.org/10.1039/c0pp00221f.

Swart, Estienne C., John R. Bracht, Vincent Magrini, Patrick Minx, Xiao Chen, Yi Zhou, Jaspreet S. Khurana, et al. 2013. "The *Oxytricha* Trifallax Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes." Edited by Jonathan A. Eisen. *PLoS Biology* 11 (1): e1001473. https://doi.org/10.1371/journal.pbio.1001473.

Törönen, Petri, Alan Medlar, and Liisa Holm. 2018. "PANNZER2: A Rapid Functional Annotation Web Server." *Nucleic Acids Research* 46 (W1): W84–88. https://doi.org/10.1093/NAR/GKY350.

Vaser, Robert, and Mile Šikić. 2019. "Yet Another de Novo Genome Assembler." *BioRxiv*, June, 656306. https://doi.org/10.1101/656306.

———. 2021. "Time- and Memory-Efficient Genome Assembly with Raven." *Nature Computational Science 2021 1:5* 1 (5): 332–36. https://doi.org/10.1038/s43588-021-00073-4.

Venter, J. C., M. D. Adams, G. G. Sutton, A. R. Kerlavage, H. O. Smith, and M. Hunkapiller. 1998. "Shotgun Sequencing of the Human Genome." *Science* 280 (5369): 1540–42. https://doi.org/10.1126/SCIENCE.280.5369.1540/ASSET/0926C819-B450-4DDD-8E64-9D02C5A0633C/ASSETS/GRAPHIC/1540-1.GIF.

Wang, Ruan lin, Wei Miao, Wei Wang, Jie Xiong, and Ai hua Liang. 2018. "EOGD: The Euplotes Octocarinatus Genome Database." *BMC Genomics* 19 (1): 1–6. https://doi.org/10.1186/S12864-018-4445-Z/FIGURES/3.

Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, et al. 2019. "Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome." *Nature Biotechnology 2019 37:10* 37 (10): 1155–62. https://doi.org/10.1038/s41587-019-0217-9.

Chapter 4

Expression of genome reorganization gene homologs during sexual reproduction in *B. stoltei*

Minakshi Singh[1], Kwee Boon Brandon Seah[1], Christiane Emmerich[1], Aditi Singh[1], Christian Woehle[2], Bruno Huettel[2], Adam Byerly[3], Naomi Alexandra Stover[4], Mayumi Sugiura[5], Terue Harumoto[5], Estienne Carl Swart[1]

[1] Max Planck Institute for Biology, Tübingen, Germany

[2] Max Planck Genome Center Cologne, Max Planck Institute for Plant Breeding, Cologne, Germany

[3] Department of Computer Science and Information Systems, Bradley University, Peoria IL, USA

[4] Department of Biology, Bradley University, Peoria, IL, USA

[5] Nara Women's University, Nara, Japan

This chapter has been adapted from the BioRxiv preprint DOI: https://doi.org/10.1101/2021.12.14.471607.

I generated all the data, performed the preliminary genome assemblies and preliminary gene annotation (Chapter 3) leading up to the draft assembly of the somatic genome, performed functional annotation of the genome. I executed the developmental time series and analyzed the gene expression by integrating transcriptomic data, partially described in Chapter 4. I performed the phylogenetic analysis, gene expression analysis and annotation of the transposase protein domains.

Dr. Estienne Swart compiled the final version of the draft assembly of the somatic genome. Additional analysis of the somatic genome was performed by Dr.'s Kwee Boon Brandon Seah and Estienne Swart. Dr.'s Estienne Swart and Kwee Boon Brandon Seah reviewed and edited the manuscript. Details of all author contributions are listed in Appendix A.1.

4.1. Introduction

A fully annotated MAC genome for *B. stoltei* was produced from the precursors discussed in Chapter 3, and now joins the sequenced and annotated genomes of the model ciliates *Paramecium, Tetrahymena* and *Oxytricha*. A great deal is known about the mechanisms of genome reorganization in these model ciliates. The scanning model of IES excision (Mochizuki et al. 2002) proposes that sRNAs generated in a complex process during sexual reproduction are transported to the developing MAC, where they demarcate the IESs from the macronuclear-destined sequences. They guide the excision machinery, containing a principal domesticated transposase, to the correct regions of the genome for DNA excision. *Tetrahymena* has distinct domesticated transposases that excise different subsets of IESs, namely those that are predominantly imprecisely excised and intergenic (by Tpb2) (Cheng et al. 2010), versus those that are rare, precisely excised and intragenic (by Tpb1 and Tpb6) (Cheng et al. 2016; Feng et al. 2017). In *Paramecium*, IESs are predominantly intragenic (Arnaiz et al. 2012). IES excision is precise (Arnaiz et al. 2012) and is carried out by a heteromeric complex of PiggyMac and PiggyMac-like proteins (Bischerour et al. 2018; Dubois et al. 2017).

Participants in this process specific to genome reorganization in addition to the transposases are Dicer-like (Dcls) proteins, which are involved in sRNA generation; Piwi proteins, which transport the sRNA. In ciliates such as *Paramecium* and *Tetrahymena,* the shorter Dicer-like proteins (Dcls) are distinguished from longer Dicer proteins (Dcrs), which possess additional N-terminal domains and produce small RNAs involved in gene regulation, notably siRNAs (Sandoval et al. 2014). In the scanning model of MAC development in *Tetrahymena* and *Paramecium*, Dcls cooperate with Piwi proteins, converting long double-stranded RNA transcripts produced in the maternal MIC into "scan RNAs" (scnRNAs) (Mochizuki et al. 2002; Mochizuki and Gorovsky 2005; Sandoval et al. 2014; Lepère et al. 2009; Schoeberl et al. 2012; Noto and Mochizuki 2018). Piwi-bound scnRNAs are transported to the maternal MAC where a subtractive process takes place, leaving only scnRNAs complementary to the MIC-limited genome. The remaining scnRNAs are transported to the new, developing MAC, where they target MIC-limited regions for excision (Mochizuki et al. 2002; Mochizuki and Gorovsky 2005; Sandoval et al. 2014; Lepère et al. 2009; Schoeberl et al. 2012; Noto and Mochizuki 2018). Additionally, histone variants and histones-modifications on nucleosomes regulate access to the DNA during the process.

In *Tetrahymena*, IESs frequently contain transposons or are derived from them. These IESs are targeted for removal by sRNA machinery that is involved in depositing methylation marks on Histone 3 Lysine 9 (H3K9) and Histone 3 Lysine 27 (H3K27), in a process akin to heterochromatin formation except that the marked regions are excised entirely (Chalker, 2008; Liu et al., 2007). In *Paramecium*, a mechanism involving H3K9- and H3K27-trimethylation (H3K27me3) represses MIC genome-encoded transposable element gene expression. Experimental elimination of these marks leads to low efficiency of IES excision and lethal outcomes when new MAC genomes are produced (Frapporti et al., 2019). A particular histone variant (H3.4) present in polytene DNA was proposed to be the target of trimethylation, facilitating heterochromatinization and excision of IESs not protected by 27 nt macRNAs in the ciliate *Stylonychia* (Postberg et al., 2018). These instances illustrate the pivotal role that histone modifications and histone variant play in genome reorganization

With the annotated *B. stoltei* MAC genome and transcriptomic data collected over the course of sexual reproduction and MAC development, we can carry out a comparative study of genome reorganization in ciliates. This will allow us to identify proteins in *Blepharisma* that are homologs of key proteins known to play important roles in genome reorganization in other ciliates. Here, I present these homologs and their expression during development of the new MAC in *Blepharisma*.

4.2. Results

4.2.1. Morphological staging of conjugation time-course in *B. stoltei*

Complementary *B. stoltei* strains were treated with gamones of the opposite mating type, before mixing to initiate conjugation (Miyake et al., 1991; Sugiura et al., 2012). Gamone 1 was obtained as the cell-free fluid from mating-type 1 cells (Chapter 8, section 8.7) and was used to treat mating type 2 cells. Synthetic gamone 2 was used to treat mating type 1 cells. Meiosis begins around 2 h after conjugating cell pairs form and continues up to 18 h, when gametic nuclei generated by meiosis are exchanged (Chapter 1, Section 1.5). This is followed by karyogamy and mitotic multiplication of the zygotic nucleus at 22 hours. Around 26 h, new, developing primary MACs can be observed as large irregular bodies in the conjugating pairs (Figure 4.1). These nuclei mature into the new MACs of the exconjugant cell by 38 h, after which cell division generates two daughter cells. Smaller secondary MACs derived directly from MICs, and therefore without all the intermediate nuclear stages, can also be seen from 22 h, though these eventually disappear and give way to the primary MACs.

The proportion of cells in each developmental stage was estimated by fixing cells at the specified timepoints, counting the number of cells at each of those stages and classifying them according to their nuclear morphology (Figure 4.1). The trend evident in the bar chart shows that the cells progressively underwent developmental changes at different timepoints. Cells from this developmental time course were also isolated for mRNA and small RNA extraction, which was sequenced (RNA-seq). Samples for morphological staging and RNA-seq were taken at intervals from the time of mixing ("0 hour" time point) up to 38 hours (Chapter 8, section 8.8). The transcriptomic reads from each of the series of developmental timepoints were used to analyze the expression of various genes of interest across development.

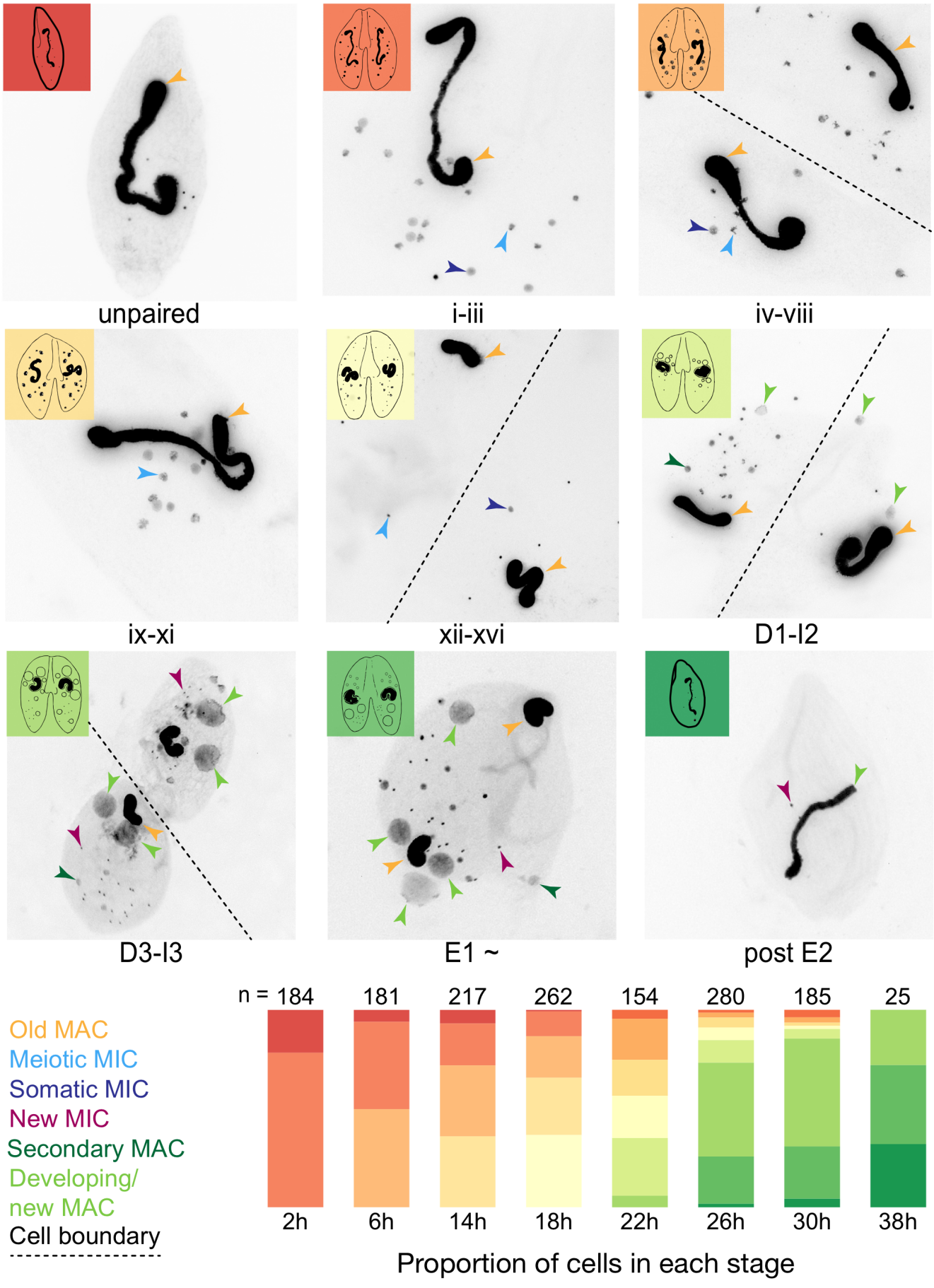

unpaired
i-iii
iv-viii
ix-xi
xii-xvi
D1-I2
D3-I3
E1 ~
post E2
Old MAC
Meiotic MIC
Somatic MIC
New MIC
Secondary MAC
Developing/
new MAC
Cell boundary
n = 184
181
217
262
154
280
185
25
2h
6h
14h
18h
22h
26h
30h
38h
Proportion of cells in each stage

Figure 4.1. Developmental staging of *B. stoltei* for RNA-seq. **Classification of nuclear morphology into stages is according to previous descriptions (Miyake et al., 1991).** Nuclear events occurring before and up to, but not including fusion of the gametic nuclei (syngamy) are classified into sixteen stages indicated by Roman numerals. These are the pre-gamic stages of conjugation, where the MICs undergo meiosis and the haploid products of meiotic MICs are exchanged between the conjugating cells. Stages after syngamy are classified into 10 stages as shown in Figure 1.5. Illustration of various cell stages adapted from (Suzuki, 1957)). Stacked bars show the proportion of cells at each time point at different stages of development. The number of cells inspected (n) is shown above each bar**.**

4.2.2. Transcriptomic analysis of the conjugation time-course in *B. stoltei*

To gain an overview of the molecular processes during *Blepharisma* genome editing, we examined gene expression trends across development of genes of interest.

4.2.2.1. Small RNA biogenesis machinery in B. stoltei

We found putative Dicer, Dicer-like and Piwi proteins encoded by the *B. stoltei* MAC genome (Figure 4.2). The single *B. stoltei* Dicer (Dcr) protein has the characteristic N-terminal Dicer domains followed by a pair of RNase III domains (PFAM domain Ribonuclease_3; PF00636). There are also three Dicer-like proteins (Dcl1-3), which consist of RNase III domains only. Dcl1 expression is upregulated shortly after conjugation begins and before onset of meiosis; Dcl2 and Dcl3 are upregulated from meiosis onwards, peaking during anlagen formation. *B. stoltei* also appears to have an additional truncated Dcr homolog (881 aa), a putative Dicer-derived protein (Dcrd), which lacks the RNase III domain portion found in the complete Dicer (Figure 4.2).

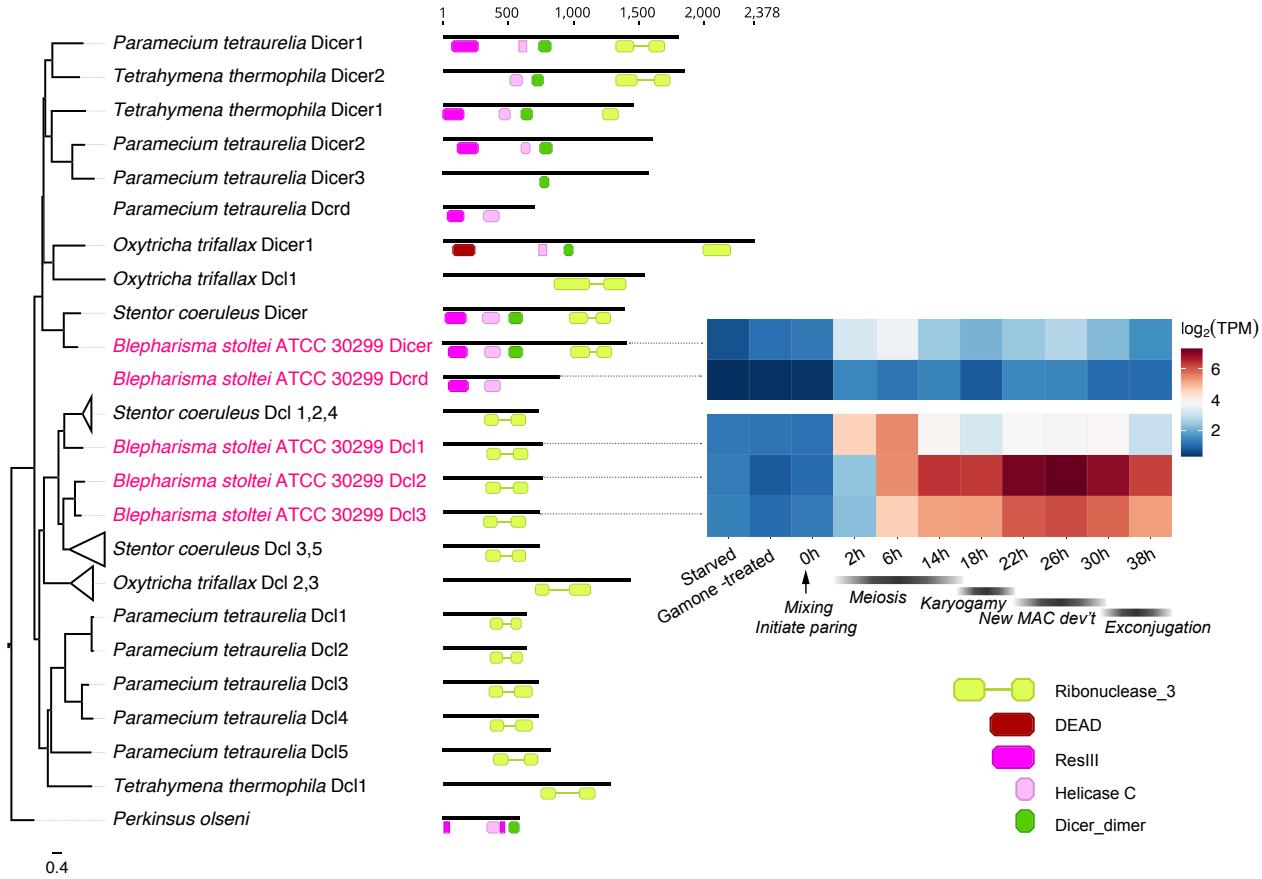


Figure 4.2. ResIII, Helicase_c and Ribonuclease_3 domains in *B. stoltei*. Phylogeny with PFAM domain architecture and gene expression heatmap for *B. stoltei*.

A comparative phylogenetic analysis of the *Blepharisma* Dicer and Dicer-like proteins with other ciliate Dicer and Dicer-proteins reveals that the *Blepharisma* Dicer shares ancestry with the *Stentor* Dicer protein, both of which form an outgroup to the clade of the Dicer proteins from *Paramecium*, *Tetrahymena* and *Oxytricha*. Similarly, the *Blepharisma* Dcl proteins are closely related to those of *Stentor* and more distantly related to the Dcl proteins of *Paramecium*, *Tetrahymena* and *Oxytricha*. The presence of the Dicer protein suggests the presence of an siRNA biogenesis mechanism in *Blepharisma,* while the Dcl proteins suggests the presence of sRNA biogenesis machinery involved in genome reorganization comparable to that observed in *Paramecium* and *Tetrahymena* (Sandoval et al. 2014).

We also found nine proteins with Piwi and PAZ domains (five of them with ArgoL domains) in the *B. stoltei* MAC genome. Two closely related *Blepharisma* Piwi paralogs, Biwi 7 and Biwi 8, are highly upregulated during meiosis and throughout subsequent development (Figure 4.3). These genes are both among the most highly expressed genes at 26 h, when the new MAC is forming.

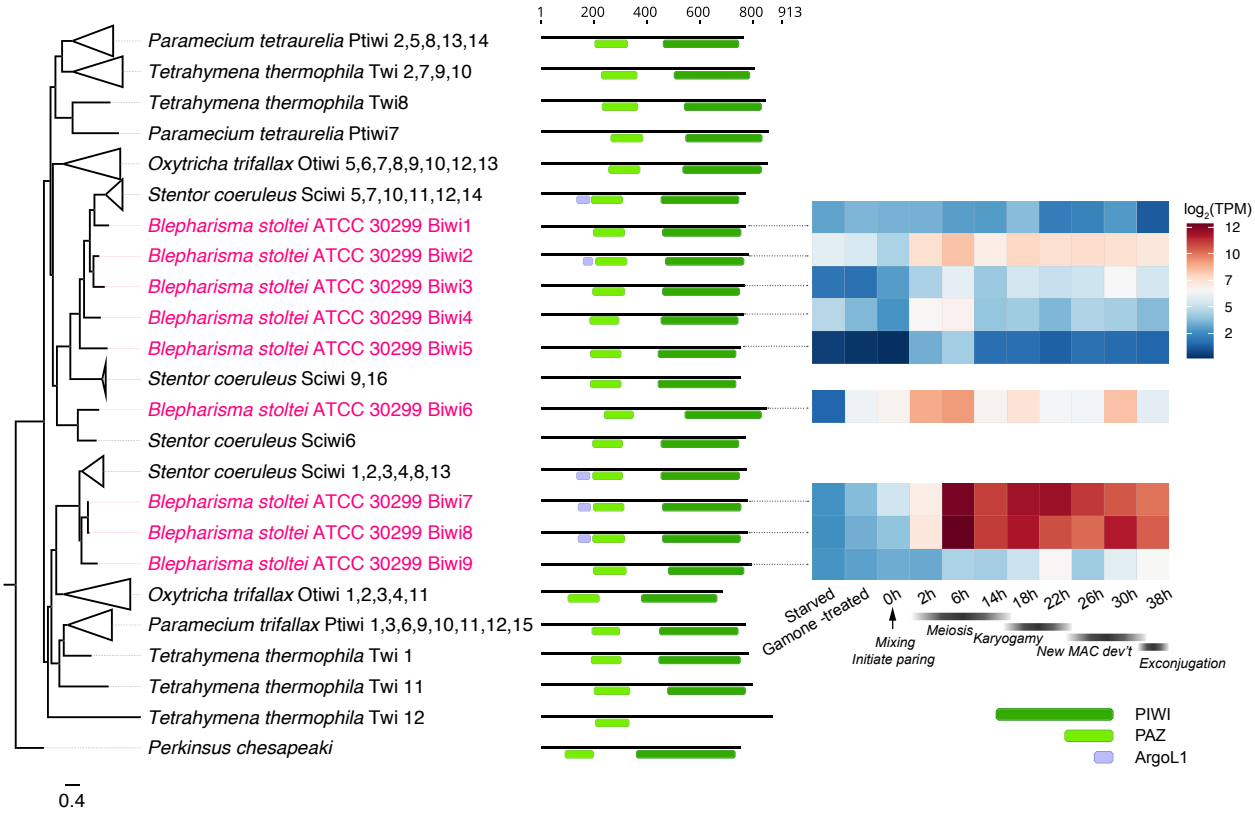


Figure 4.3.PIWI domain in *B. stoltei.* Phylogeny with PFAM domain architecture and gene expression heatmap for *B. stoltei*.

4.2.2.2. Development-specific histone variant upregulation

We annotated the four core histones, H2A, H2B, H3 and H4, in *B. stoltei* using curated domain from Histone DB (v2.0) to generate HMM (Hidden Markov Model) profiles and used hmmscan to annotate the histone candidate genes (Figure 4.4). We found eleven putative H2A, five H2B, eleven H3 and five H4 histone proteins. Histone H2A forms dimers with histone H2B and histone H3 forms dimers with histone H4 (Malik and Henikoff, 2003). The H2B and H4 histones are known to be more conserved in comparison to H2A and H3 across several eukaryotic lineages (Malik and Henikoff, 2003). The trend of greater diversity in homologs of H2A and H3 in other eukaryotic lineages is also preserved in *Blepharisma*, where we find twice as many H2A homologs as those of H2B and almost twice as many H3 homologs as those of H4.

While nine of the eleven H2A candidates are ~120 aa long and possess an H2A_C (H2A C-terminal) domain in addition to the identifiable H2A domain, the two remaining do not and are distinctly longer (Contig_7.g1593 ~300 and Contig_44.g977 ~700 aa). One of the five H2B proteins also diverges from the domain structure seen in the majority of the H2B candidates, since it possesses several TPR (Tetratricopeptide repeat) domains in addition to the H2B domain and is also much longer, ~600 aa compared with the 120 aa length of the others. A similar divergence from majority domain structure and length is also seen in two of the eleven H3 proteins, where the ~900 aa outliers include VIT (Vault protein inter-alpha-trypsin) and VWA_3 (Von Willebrand factor type A) domains in addition to the H3-specific domain. A slightly longer H4 candidate is also present among the set of putative H4 proteins. Awaiting biochemical verification, we suspect that the non-conforming candidates are not typical histones, and might be proteins with possessing histone-like domains which may be co-opted for other functions.

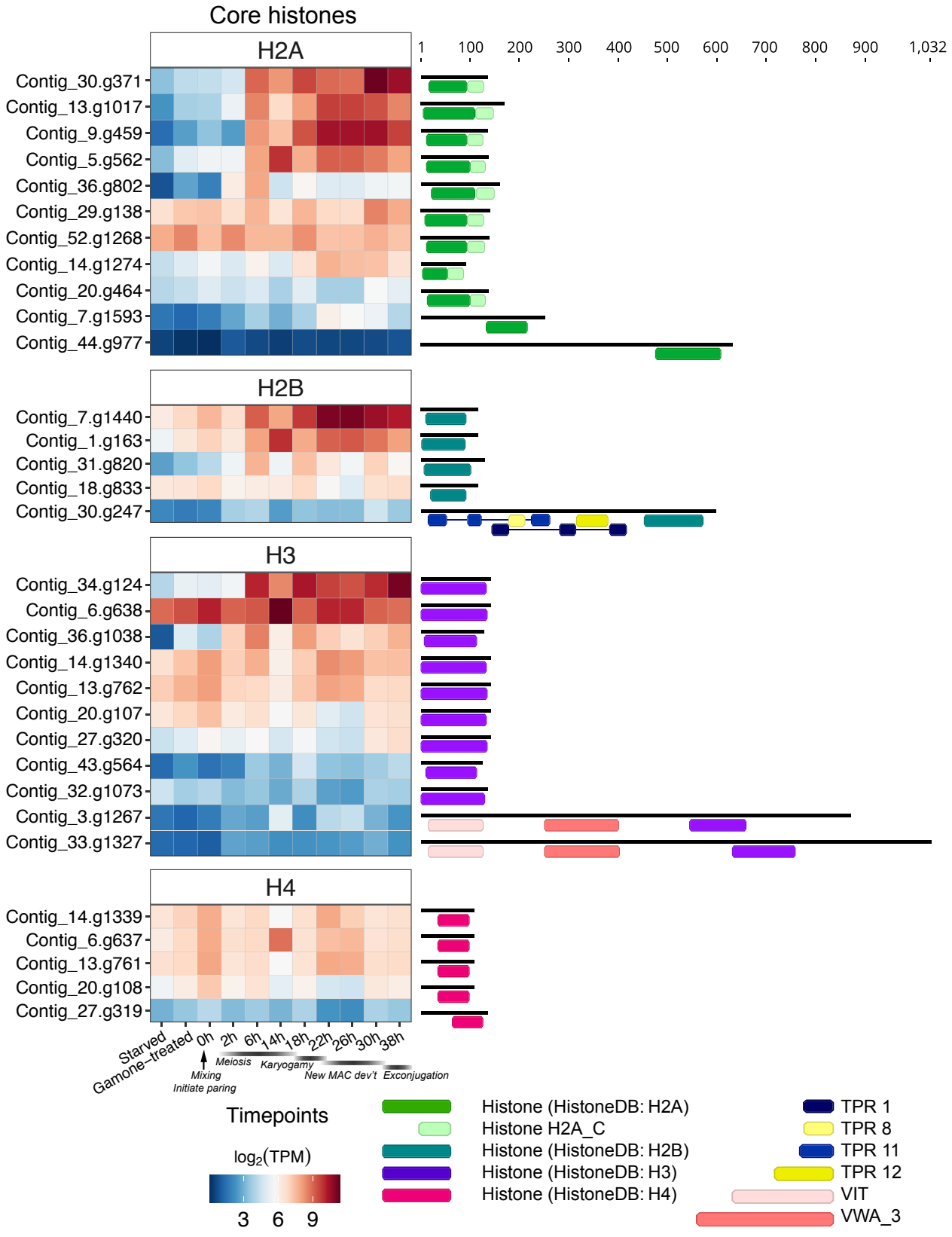


Figure 4.4. Histones and histone-domain-containing proteins in *Blepharisma*. Gene expression heatmaps are shown as in previous figures, clustered according to major histone type as classified according to HistoneDB domain models. Domains from PFAM and HistoneDB are shown to the right.

Since substantial upregulation of certain histone variants occurs during development in both *Oxytricha* and *Stylonychia*, including during the period of genome editing (Aeschlimann et al., 2014; Forcob et al., 2014; Postberg et al., 2018), we examined the patterns of expression during *Blepharisma* development. Among the *Blepharisma* histones, particular candidates of three of the core histones H2A, H2B and H3 are constitutively expressed at similar levels throughout the cycle of sexual reproduction, while others are upregulated at timepoints corresponding to different stages of meiosis (6 h and 14 h timepoints) and subsequently during new MAC development (Figure 4.4).

4.2.3. Transposase domains encoded in the *B. stoltei* somatic genome and their expression

The annotation of PFAM protein domains in predicted genes (CDSs) in the MAC also allowed us to search for and compare the presence of transposase domains in the *Blepharisma* MAC genome with other ciliate MAC genomes. The CDS regions of the *Blepharisma* somatic genome appear to encode several families of transposases (Figure 4.5), a feature also observed in other ciliates.

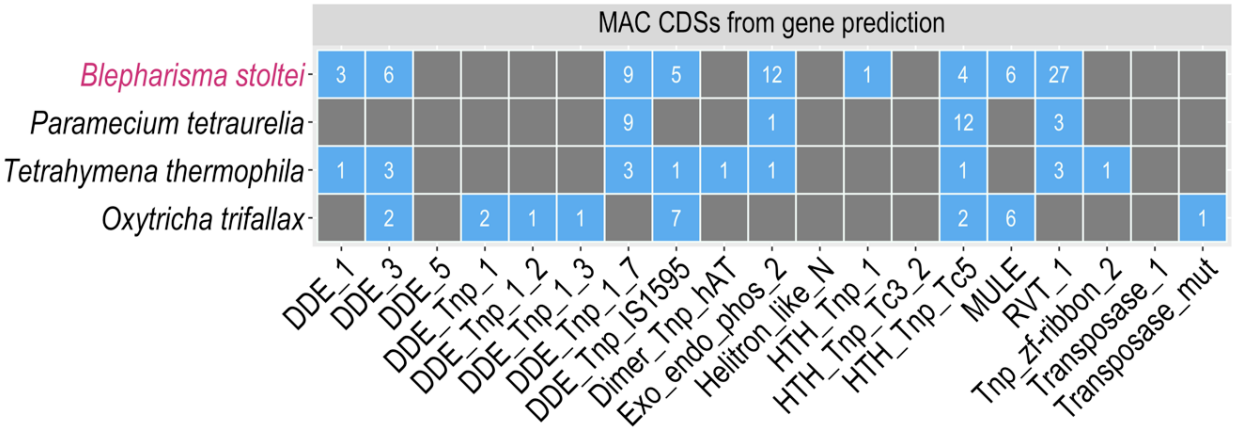


Figure 4.5. MAC genome-encoded transposases in ciliates. Presence/absence matrix of PFAM transposase domains detected in predicted MAC genome-encoded ciliate proteins.

Notable among the transposase domains present in the *Blepharisma* somatic genome is the DDE_Tnp_1_7 domain, which represents the PiggyBac family of transposases. This domain is also present in *Paramecium* and *Tetrahymena*, where the principal transposase involved in genome reorganization is known to be a PiggyBac transposase, called PiggyMac in *Paramecium* (Baudry et al. 2009), and *Tetrahymena* PiggyBac2 (Tpb2) in *Tetrahymena* (Cheng et al. 2010). Conspicuously, this domain is absent in *Oxytricha* (Swart et al. 2013), where the principal actors in genome reorganization are the germline-encoded Telomere Bearing Elements (TBEs), which encode a DDE_3 domain transposase (Nowacki et al. 2009).

Other transposase families such as those represented by the DDE_1 domain (Pogo/Tigger family transposases), the DDE_3 domain (Tc1/Mariner family transposases), the DDE_Tnp_IS1595 domain (Merlin family transposases) and the MULE domain (Mutator family transposases) are also present in the somatic genome of *Blepharisma*.

4.2.3.1. *PiggyBac family (DDE_Tnp_1_7) transposases in the MAC genome*

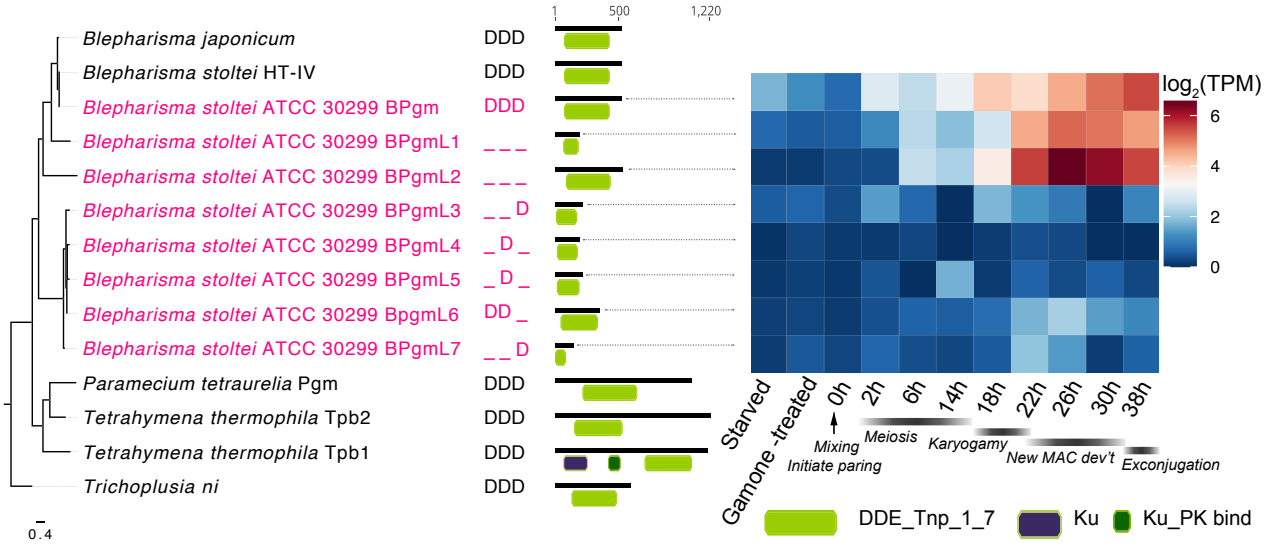


Figure 4.6. DDE_Tnp_1_7 domain in *B.stoltei.* **Phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*.** "Mixing" indicates when cells of the two complementary mating types were mixed. Outgroup: PiggyBac element from *Trichoplusia ni.* Catalytic residues: D- aspartate

Using HMMER (hmmscan) searches for the characteristic domain of PiggyBac transposases, DDE_Tnp_1_7 (PF13843), we found eight PiggyBac homologs in the *B. stoltei* MAC genome. Active transposons are characterized by terminal repeats on either side. However, none of the *Blepharisma* PiggyBac homologs were flanked by terminal repeats (identified by

RepeatModeler). We also found PiggyBac homologs in the MAC genomes of *B. stoltei* HT-IV and *B. japonicum* R1072, two other *Blepharisma* species we sequenced.

PiggyBac transposases possess three catalytic aspartate (D) residues, known as the DDD - catalytic triad. Reminiscent of *Paramecium tetraurelia*, which has ten PiggyMac homologs, but just one homolog with a complete catalytic triad (Bischerour et al. 2018), the DDD triad is preserved in just a single *Blepharisma* PiggyBac homolog (Figure 4.6; Contig_49.g1063), which we call the *Blepharisma* PiggyMac or BPgm. This gene is strongly upregulated during development from 22 to 38 h, when new MACs develop and IES excision is required (Figure 4.6).

PiggyMac homologs are also present in other heterotrich ciliates but have not yet been described because of genome assembly or annotation challenges. Using BPgm as a query sequence, we found convincing homologs containing the conserved catalytic DDD-motif in a genome assembly of the heterotrichous ciliate *Condylostoma magnum* (TBLASTN e-value 2e-24 to 2e-37). All the *C. magnum* PiggyMac homologs have a complete DDD-catalytic triad. While we failed to detect the DDE_Tnp_1_7 domain in predicted genes of the heterotrich *Stentor coeruleus*, we detected relatively weak adjacent TBLASTN matches split across two frames in its draft MAC genome (e-value 7e-15; SteCoe_contig_741 positions 6558-5475). After joining ORFs corresponding to this region and translating them, we obtained a more convincing DDE_Tnp_1_7 match with HMMER3 (e-value 2e-24). This either corresponds to a pseudogene or a poorly assembled genomic region. In addition, we searched for PiggyMac homologs in the MAC genome of the pathogenic oligohymenophorean ciliate *Ichthyophthirius multifiliis* (Coyne et al. 2011). We used the *T. thermophila* Tpb2 as a query for TBLASTN searches, since *Tetrahymena* belongs to the same ciliate sub-class as *Ichthyophthirus*. However the search returned no hits. A HMMER search using hmmscan with a six-frame translation of the *Ichthyophthirus* MAC genome against the PFAM-A database also did not return any matches with independent E-values (i-E-value) less than 1. We note that based on BUSCO analyses (Chapter 5, Supplementary Figure S5.1), the *Ichthyophthirus* genome appears to be less complete than other ciliates we examined. A better genome assembly will therefore be needed to confirm the absence or presence of PiggyBac homologs encoded in the *Ichthyophthirus* MAC genome.

4.2.3.2. *Pogo/Tigger family (DDE_1) transposases in the MAC genome*

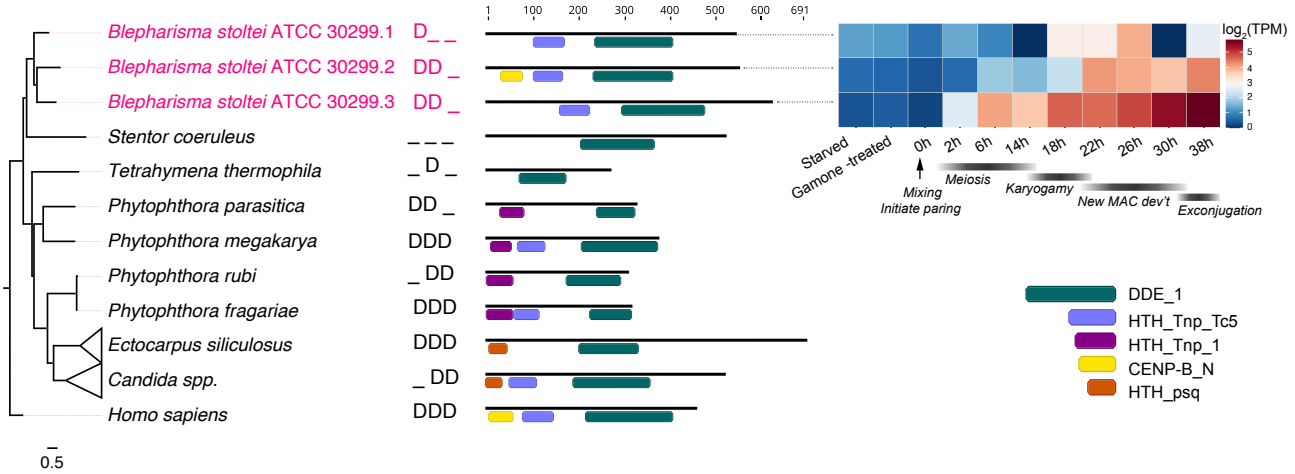


Figure 4.7. DDE_1 domain-containing proteins in *Blepharisma*. DDE_1 domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*.

Three *Blepharisma* MAC genome-encoded proteins possess the PFAM domain DDE_1 (PF03184; Figure 4.7). The PFAM domain HTH_Tnp_Tc5 (PF03221) occurs most commonly in combinations with this domain (5898 sequences; PFAM version 35). The CENP-B_N domain, one instance of which we see in the *Blepharisma* DDE_1 domain proteins, is characteristic of numerous transposases, notably those of the Tigger and Pogo families (Gao et al., 2020). Though pairwise sequence identity is low amongst the *Blepharisma* DDE_1 proteins (avg. 28.3%) in their multiple sequence alignment, the CENP-B_N domain in one of them appears to align reasonably well to corresponding regions in the two proteins lacking this domain, suggesting it deteriorated beyond the recognition capabilities of HMMER3 and the given PFAM domain model. BLASTp matches for all three proteins in GenBank are annotated either as Jerky or Tigger homologs (Jerky transposases belong to the Tigger transposase family (Gao et al., 2020)). Given that none of the *Blepharisma* MAC DDE_1 domain proteins appear to have a complete catalytic triad, it is unlikely they are involved in transposition or IES excision.

4.2.3.3. *Tc1/Mariner family (DDE_3) transposases in the MAC genome*

Six MAC-encoded transposases containing the DDE_3 domain (PF13358) are present in the *Blepharisma* MAC, all of which are substantially upregulated in MAC development and five of which possess the complete DDE catalytic triad characteristic of Tc1/Mariner transposases (Figure 4.8). All six *Blepharisma* DDE_3 genes have at least 150× HiFi read coverage, consistent with their presence in *bona fide* MAC DNA. Given that all but one of the *B. stoltei* paralogs

appear to possess a complete catalytic triad, there is a possibility that they may be involved in some IES excison.

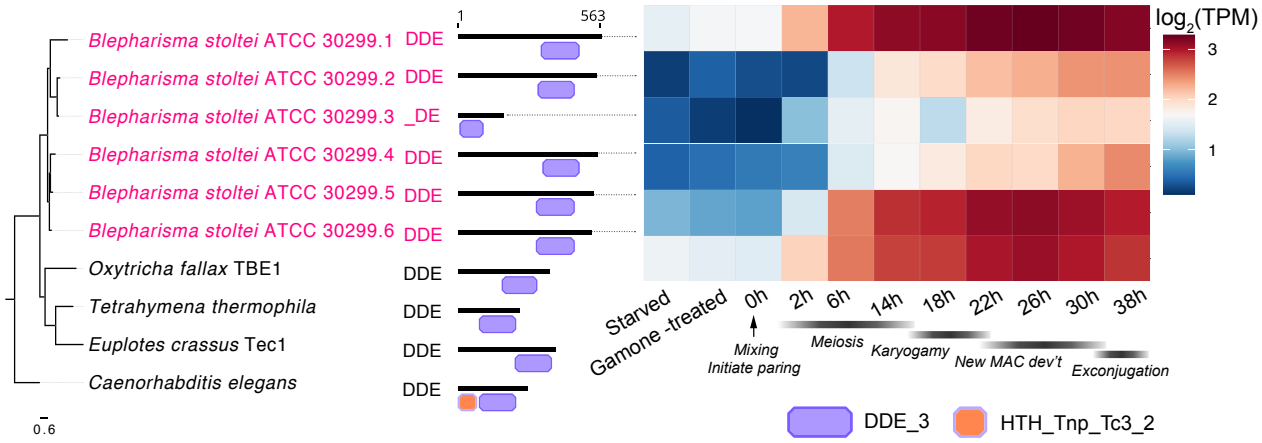


Figure 4.8. DDE_3 domain-containing proteins in *Blepharisma*. DDE_3 domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*.

4.2.3.4. *Merlin family (DDE_Tnp_IS1595) and Mutator family (MULE) transposases in the MAC genome*

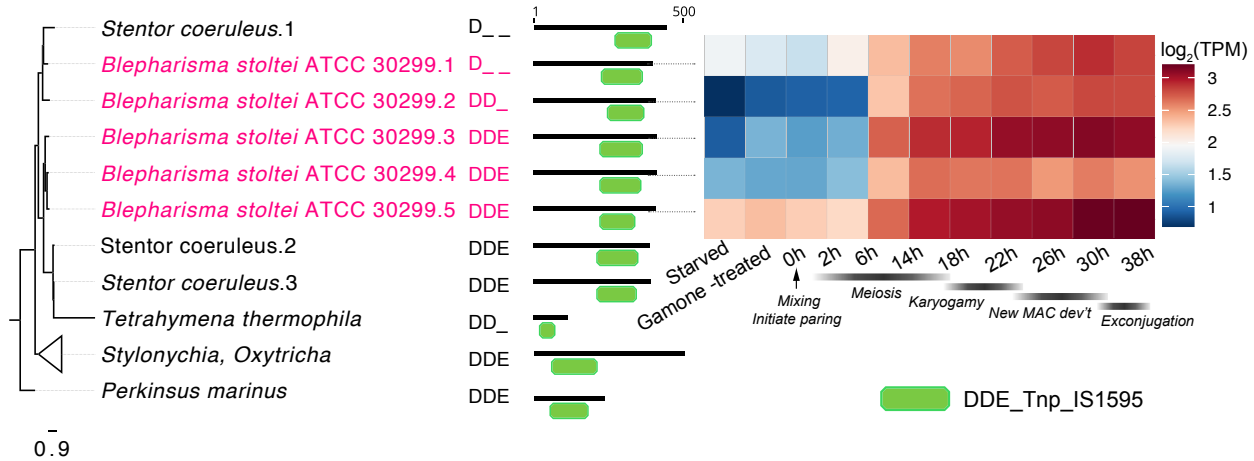


Figure 4.9. DDE_Tnp_IS1595 domain-containing proteins in *Blepharisma*. DDE_Tnp_IS1595 domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*.

We found five instances of the DDE_Tnp_IS1595 transposase (PFAM PF12762) domain and six instances of the MULE transposase (PFAM PF10551) domain in the *Blepharisma* MAC genome.

DDE_Tnp_IS1595 domains are characteristic of the Merlin transposon family and MULE is part of the Mutator transposon family (Yuan and Wessler, 2011). Their underlying

genes are upregulated during MAC development (Figure 4.9, Figure 4.10). Consistent with the notion of transposase domestication, the genes encoding DDE_Tnp_IS1595 and MULE proteins also appear to lack flanking transposon terminal inverted repeats. Additionally, members of both IS1595 and MULE transposases appear to have complete catalytic triads. Despite the presence of these transposases in the *Blepharisma* MAC and other ciliates (Figure 4.5) and their upregulation during MAC development, their role in genome reorganization remains undetermined.

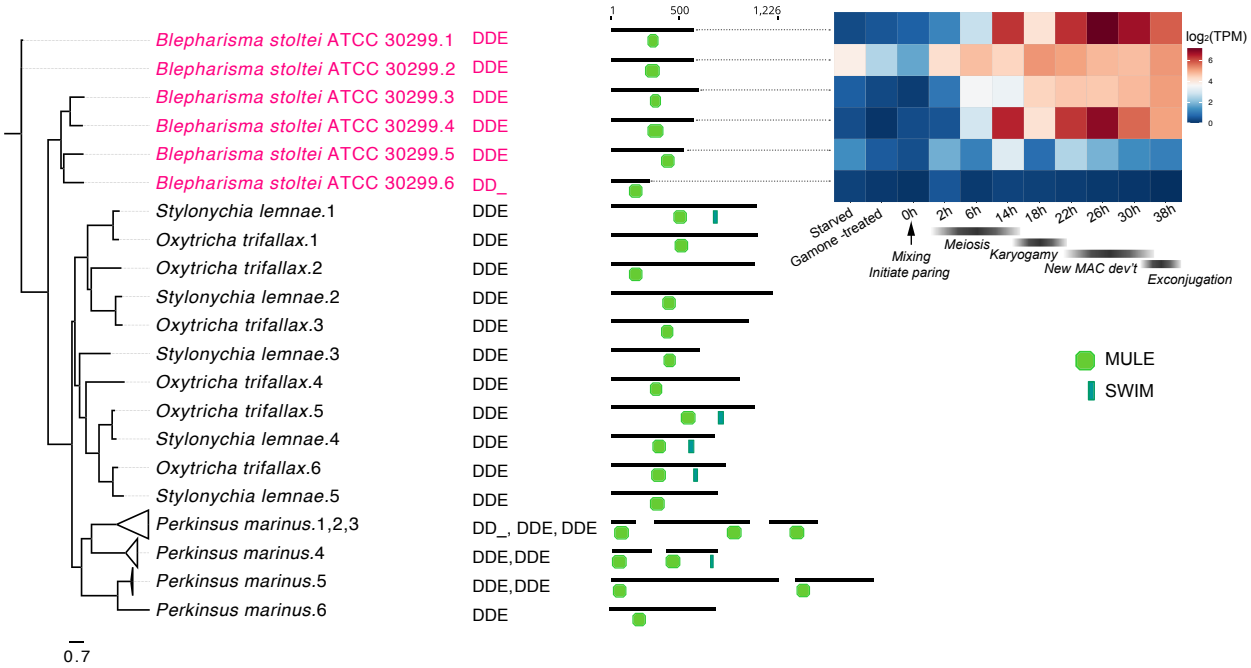


Figure 4.10. MULE domain-containing proteins in *Blepharisma*. MULE domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*.

4.3. Discussion

The ciliate somatic genome serves as a standard eukaryotic nucleus during the vegetative lifecycle of the cell. During the sexual cycle, however, the maternal somatic genome of the two cells participating in conjugation is either systematically degraded or diluted over successive cell divisions, while a new somatic genome arises from the newly formed zygotic germline nucleus. The developing somatic genome is thus the site of several key processes, such as DNA elimination and genome reorganization, which must be successfully executed to ensure the later vegetative health of the cell.

Genome reorganization has been characterized in model ciliates like *Paramecium*, *Tetrahymena* and *Oxytricha*. Its general mechanisms and participant protein identified in these model ciliates can be used for the comparative study of orthologous proteins in *Blepharisma*. Knowledge of the presence and expression of key proteins involved in genome reorganization is necessary prior to experimental investigation of their involvement in MAC development.

4.3.1. Small RNA biogenesis in ciliates is critical for genome reorganization

Development-specific proteins responsible for small RNA (sRNA) generation and transport play an important role in ciliate genome editing
(Chalker, Meyer, and Mochizuki 2013). We found Dicer, Dicer-derived and Dicer-like proteins in the *Blepharisma* MAC genome, which play an important role in sRNA generation. In *Paramecium,* two Dcls are co-expressed and cooperate to produce scnRNAs
(Sandoval et al. 2014), thus we predict that *Blepharisma* Dcl2 and Dcl3 may cooperate. In *Tetrahymena* and *Paramecium,* massive production of scnRNAs, using the Dcls and highly upregulated Piwis, initiates from meiotic nuclei. During development in *Blepharisma,* we observe a similar pattern of massive production of development-specific sRNAs (Figure 6.6), whose detailed analysis is reported in conjunction with the draft *B. stoltei* ATCC 30299 MIC genome (Seah, et. al, 2022, Chapter 6). Since *Blepharisma* species are distantly related to other ciliates whose sRNAs have been characterized, this suggests that an ancient, development-specific sRNA gene expression program may have been established in the ciliate common ancestor.

In ciliates, some Piwi proteins play a role in gene regulation in vegetative cells (Götz et al., 2016) while others are involved in genome editing (Bouhouche et al., 2011; Fang et al., 2012; Mochizuki et al., 2002). The high expression of the *Blepharisma* Piwis, Biwi 7, Biwi 8,

during sexual reproduction is a trend also observed in other ciliates. In *Stylonychia lemnae*, the massive upregulation of a Piwi homolog involved in genome editing was so conspicuous that it could be identified by subtractive hybridization of RNA (Fetzer et al. 2002). The ortholog of this gene in *Stylonychia lemnae*'s close relative, *Oxytricha trifallax*, is also one of the most highly transcribed and upregulated genes (Fang et al. 2012) during MAC development.

4.3.2. Histones in *Blepharisma*

Genome rearrangement in ciliates is influenced by processes that modify and regulate nucleosomes and consequently mediate the ability of IES-excision machinery to access the underlying DNA. The modification of core histones by acetylation and methylation is involved in allowing or repressing access to the DNA. The observed patterns of histone expression suggest that even the *Blepharisma* genome encodes variants that are likely to have a range of different functions, including in genome editing and likely also during DNA amplification in the developing new MAC. Histone H4, in contrast, appears to be expressed at relatively similar levels throughout conjugation. This constitutive expression of histone H4 is a characteristic shared among eukaryotes, which lack functional variants due their highly conserved constitution. This is a trait suggested to be favored by the greater necessity of this histone to maintain several protein-protein contacts with the other three histones (Malik and Henikoff, 2003).

Centromeric histones are involved in chromosome segregation in eukaryotes and have also been reported in the ciliates *Paramecium tetraurelia* and *Tetrahymena thermophila* (Cervantes et al. 2006). They are variants of Histone H3 (CenH3), which are longer and more divergent in sequence, and the ciliate candidates are comparable in their properties to other eukaryotic centromeric histones (Cervantes et al. 2006). We did not observe such histones in *Blepharisma*. While histone variants remain to be identified specifically in *Blepharisma,* the upregulated expression of certain H2A, H2B and H3 histones indicates that a parallel process may be at work in this ciliate.

4.3.3. *Blepharisma* possesses additional domesticated transposases whose roles await determination

All ciliate species have multiple MAC genome-encoded transposase families (Figure 4.5). Though upregulation of some of these homologs in model ciliates has been noted (Swart et al. 2013; Chen et al. 2014; Vogt and Mochizuki 2013), their roles remain to be

determined. In *Blepharisma,* the MAC genome encodes transposases with the PFAM domains "DDE_1", "DDE_3", "DDE_Tnp_IS1595" and "MULE" in addition to the PiggyBac homologs.

In *Blepharisma* and numerous other organisms, the DDE_1 domains co-occur with CENPB domains. Two such proteins represent entirely different proposed exaptations in mammals, where it acts as a centromere-binding protein, and fission yeast, where it acts as a regulatory protein
(Mojzita and Hohmann 2006; Hohmann 1993; Casola, Hucks, and Feschotte 2008). Given the great evolutionary distances involved, there is no reason to expect that the *Blepharisma* homologs possess either function. None of the three proteins with co-occurring DDE_1 and CENPB domains have a complete catalytic triad, making it unlikely that these are active transposases or IES excisases, though all three are noticeably upregulated during MAC development.

Of the six proteins with the PFAM domain DDE_3 encoded by *Blepharisma* MAC genes, five possess a complete catalytic triad. All the "DDE_3" protein genes are upregulated during conjugation in *B. stoltei*, peaking during new MAC development. The DDE_3 domain is also characteristic of DDE transposases encoded by the Telomere-Bearing Element transposons (TBEs) of *Oxytricha trifallax* (Williams et al., 1993; Witherspoon et al., 1997), which, despite being germline-limited, are proposed to be involved in IES excision (Nowacki et al., 2009). DDE_3-containing transposons, called Tec elements, are found in another spirotrichous ciliate, *Euplotes crassus*, but no role in genome editing has been established for these (Jahn et al., 1993). TBEs and Tec elements do not share obvious features, other than possessing an encoded protein belonging to the IS630-Tc1 transposase (super)-family (Doak et al., 1994).

A number of DDE_Tnp_IS1595 and MULE domain-containing proteins in *Blepharisma* have complete catalytic triads and also show pronounced upregulation during *Blepharisma* MAC development. Among other ciliates with draft MAC genomes, the IS1595- and MULE transposase-like domains have so far only been observed in the spirotrichous ciliates *Oxytricha* and *Stylonychia* (Figure 4.5) (Aeschlimann et al. 2014; Swart et al. 2013). Currently no particular functions have been demonstrated for these proteins in these ciliates, but their genes were substantially upregulated during their development (Swart et al. 2013; Chen et al. 2014).

The genes encoding all these transposase domains lack flanking terminal repeats characteristic of active transposons, suggesting they are further classes of domesticated

transposases. Many of these additional domesticated transposases have complete catalytic triads and are substantially upregulated during *Blepharisma* development, but it remains to be established whether they are capable of excision and, if so, whether this is precise. Should the additional *Blepharisma* domesticated transposases be still capable of excision, they might be involved in the excision of a subset of the intergenic IESs, though not in a precise form.

It is important to consider the upregulation of transposases during MAC developments, since it is crucial that the timing of IES excisase expression should coincide with the formation of the new MAC genome. However, equally important is the manner in which the excisase performs DNA elimination. Upon excision, classical cut-and-paste transposases in eukaryotes typically leave behind additional bases, notably including the target-site duplication arising where they were inserted, forming a "footprint" (van Luenen, Colloms, and Plasterk 1994). PiggyBac homologs, eight of which we found in *Blepharisma,* are unique in performing precise, "seamless" excision in eukaryotes (Elick, Bauser, and Fraser 1996). This conserves the number of bases at the site of transposon insertion after excision, a property rendering them popular for genetic engineering (Q. Chen et al. 2020). *Tetrahymena* Tpb2 is the only example of a PiggyBac homolog associated with imprecise excision (Cheng et al. 2010). Since intragenic IESs are abundant in *Blepharisma* (Chapter 6), like *Paramecium* and unlike *Tetrahymena*, it is essential that these are excised precisely. This imposes an important constraint on transposase candidates for the main IES excisase in *Blepharisma,* in spite of the presence of multiple transposases families upregulated during MAC development. The high expression of the *Blepharisma* PiggyBac homologs, coupled with the intragenic nature of IESs in *Blepharisma* (Chapter 6), strongly indicates that a PiggyBac homolog is the main IES excisase in *Blepharisma*. A detailed analysis of the PiggyBac homologs of *Blepharisma* is presented in Chapter 5.

The work presented so far indicates the presence of sRNA-generation and transport machinery i.e. homologs of Dicer, Dicer-like and Piwi protein and a full complement of histones in *Blepharisma,* in addition to several families of transposases. From these individual parts, it appears that the molecular toolkit present in *B. stoltei* contains all the key components stipulated by the scanning model of IES excision, where regions in the developing MAC genome are differentially marked for retention or excision through transposases by a specific family of sRNAs, which have been generated and transported in concert by the Dicer-like and Piwi proteins. This demonstrates that genome reorganization in *Blepharisma* proceeds through similar

mechanisms to the scanning model. This implies that the last ciliate common ancestor already possessed this machinery and that deviations from the scanning model, which may be discovered in the future in other ciliates are more likely to be modifications acquired by the separate lineages, as opposed to being traits common among ciliate lineages.

4.4. Bibliography

Aeschlimann, Samuel H, Franziska Jönsson, Jan Postberg, Nicholas A Stover, Robert L Petera, Hans-Joachim Lipps, Mariusz Nowacki, and Estienne C Swart. 2014. "The Draft Assembly of the Radically Organized *Stylonychia Lemnae* Macronuclear Genome." *Genome Biology and Evolution* 6 (7): 1707–23. https://doi.org/10.1093/gbe/evu139.

Arnaiz, Olivier, Nathalie Mathy, Céline Baudry, Sophie Malinsky, Jean-Marc Aury, Cyril Denby Wilkes, Olivier Garnier, et al. 2012. "The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences." *PLoS Genetics* 8 (10): e1002984. https://doi.org/10.1371/journal.pgen.1002984.

Baudry, Céline, Sophie Malinsky, Matthieu Restituito, Aurélie Kapusta, Sarah Rosa, Eric Meyer, and Mireille Bétermier. 2009. "PiggyMac, a Domesticated PiggyBac Transposase Involved in Programmed Genome Rearrangements in the Ciliate Paramecium Tetraurelia." *Genes & Development* 23 (21): 2478–83. https://doi.org/10.1101/gad.547309.

Bischerour, Julien, Simran Bhullar, Cyril Denby Wilkes, Vinciane Régnier, Nathalie Mathy, Emeline Dubois, Aditi Singh, et al. 2018. "Six Domesticated PiggyBac Transposases Together Carry out Programmed DNA Elimination in Paramecium." *ELife* 7 (September). https://doi.org/10.7554/eLife.37927.

Casola, Claudio, Donald Hucks, and Cédric Feschotte. 2008. "Convergent Domestication of Pogo-like Transposases into Centromere-Binding Proteins in Fission Yeast and Mammals." *Molecular Biology and Evolution* 25 (1): 29–41. https://doi.org/10.1093/molbev/msm221.

Cervantes, Marcella D, Xiaohui Xi, Danielle Vermaak, Meng-Chao Yao, and Harmit S Malik. 2006. "The CNA1 Histone of the Ciliate Tetrahymena Thermophila Is Essential for Chromosome Segregation in the Germline Micronucleus." *Molecular Biology of the Cell* 17 (1): 485–97. https://doi.org/10.1091/mbc.e05-07-0698.

Chalker, Douglas L, Eric Meyer, and Kazufumi Mochizuki. 2013. "Epigenetics of Ciliates." *Cold Spring Harbor Perspectives in Biology* 5 (12): a017764. https://doi.org/10.1101/cshperspect.a017764.

Cheng, Chao-Yin, Alexander Vogt, Kazufumi Mochizuki, and Meng-Chao Yao. 2010. "A Domesticated PiggyBac Transposase Plays Key Roles in Heterochromatin Dynamics and DNA Cleavage during Programmed DNA Deletion in Tetrahymena Thermophila." *Molecular Biology of the Cell* 21 (10): 1753–62. https://doi.org/10.1091/mbc.e09-12-1079.

Cheng, Chao-Yin, Janet M Young, Chih-Yi Gabriela Lin, Ju-Lan Chao, Harmit S Malik, and Meng-Chao Yao. 2016. "The PiggyBac Transposon-Derived Genes TPB1 and TPB6 Mediate Essential Transposon-like Excision during the Developmental Rearrangement of Key Genes in *Tetrahymena Thermophila*." *Genes & Development* 30 (24): 2724–36. https://doi.org/10.1101/gad.290460.116.

Chen, Qiujia, Wentian Luo, Ruth Ann Veach, Alison B Hickman, Matthew H Wilson, and Fred Dyda. 2020. "Structural Basis of Seamless Excision and Specific Targeting by PiggyBac

Transposase." *Nature Communications* 11 (1): 3446. https://doi.org/10.1038/s41467-020-17128-1.

Chen, Xiao, John R Bracht, Aaron David Goldman, Egor Dolzhenko, Derek M Clay, Estienne C Swart, David H Perlman, et al. 2014. "The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development." *Cell* 158 (5): 1187–98. https://doi.org/10.1016/j.cell.2014.07.034.

Coyne, Robert S, Linda Hannick, Dhanasekaran Shanmugam, Jessica B Hostetler, Daniel Brami, Vinita S Joardar, Justin Johnson, 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 Biology* 12 (10): R100. https://doi.org/10.1186/gb-2011-12-10-r100.

Doak, T G, F P Doerder, C L Jahn, and G Herrick. 1994. "A Proposed Superfamily of Transposase Genes: Transposon-like Elements in Ciliated Protozoa and a Common 'D35E' Motif." *Proceedings of the National Academy of Sciences of the United States of America* 91 (3): 942–46. https://doi.org/10.1073/pnas.91.3.942.

Dubois, Emeline, Nathalie Mathy, Vinciane Régnier, Julien Bischerour, Céline Baudry, Raphaëlle Trouslard, and Mireille Bétermier. 2017. "Multimerization Properties of PiggyMac, a Domesticated PiggyBac Transposase Involved in Programmed Genome Rearrangements." *Nucleic Acids Research* 45 (6): 3204–16. https://doi.org/10.1093/nar/gkw1359.

Elick, Teresa A., Christopher A. Bauser, and M J Fraser. 1996. "Excision of the PiggyBac Transposable Element in Vitro Is a Precise Event That Is Enhanced by the Expression of Its Encoded Transposase." *Genetica* 98 (1): 33–41. https://doi.org/10.1007/BF00120216.

Fang, Wenwen, Xing Wang, John R Bracht, Mariusz Nowacki, and Laura F Landweber. 2012. "Piwi-Interacting RNAs Protect DNA against Loss during *Oxytricha* Genome Rearrangement." *Cell* 151 (6): 1243–55. https://doi.org/10.1016/j.cell.2012.10.045.

Feng, Lifang, Guangying Wang, Eileen P Hamilton, Jie Xiong, Guanxiong Yan, Kai Chen, Xiao Chen, et al. 2017. "A Germline-Limited PiggyBac Transposase Gene Is Required for Precise Excision in *Tetrahymena* Genome Rearrangement." *Nucleic Acids Research* 45 (16): 9481–9502. https://doi.org/10.1093/nar/gkx652.

Fetzer, Christian P, Daniel J Hogan, and Hans J Lipps. 2002. "A PIWI Homolog Is One of the Proteins Expressed Exclusively during Macronuclear Development in the Ciliate Stylonychia Lemnae." *Nucleic Acids Research* 30 (20): 4380–86. https://doi.org/10.1093/nar/gkf579.

Forcob, Sakeh, Aneta Bulic, Franziska Jönsson, Hans J Lipps, and Jan Postberg. 2014. "Differential Expression of Histone H3 Genes and Selective Association of the Variant H3.7 with a Specific Sequence Class in Stylonychia Macronuclear Development." *Epigenetics & Chromatin* 7 (1): 4. https://doi.org/10.1186/1756-8935-7-4.

Hohmann, S. 1993. "Characterisation of PDC2, a Gene Necessary for High Level Expression of Pyruvate Decarboxylase Structural Genes in Saccharomyces Cerevisiae." *Molecular & General Genetics : MGG* 241 (5–6): 657–66. https://doi.org/10.1007/BF00279908.

Lepère, Gersende, Mariusz Nowacki, Vincent Serrano, Jean-François Gout, Gérard Guglielmi, Sandra Duharcourt, and Eric Meyer. 2009. "Silencing-Associated and Meiosis-Specific Small RNA Pathways in *Paramecium Tetraurelia*." *Nucleic Acids Research* 37 (3): 903–15. https://doi.org/10.1093/nar/gkn1018.

Luenen, H G van, S D Colloms, and R H Plasterk. 1994. "The Mechanism of Transposition of Tc3 in C. Elegans." *Cell* 79 (2): 293–301. https://doi.org/10.1016/0092-8674(94)90198-8.

Mochizuki, Kazufumi, Noah A Fine, Toshitaka Fujisawa, and Martin A Gorovsky. 2002. "Analysis of a Piwi-Related Gene Implicates Small RNAs in Genome Rearrangement in Tetrahymena." *Cell* 110 (6): 689–99. https://doi.org/10.1016/s0092-8674(02)00909-1.

Mochizuki, Kazufumi, and Martin A Gorovsky. 2005. "A Dicer-like Protein in *Tetrahymena* Has Distinct Functions in Genome Rearrangement, Chromosome Segregation, and Meiotic Prophase." *Genes & Development* 19 (1): 77–89. https://doi.org/10.1101/gad.1265105.

Mojzita, Dominik, and Stefan Hohmann. 2006. "Pdc2 Coordinates Expression of the THI Regulon in the Yeast Saccharomyces Cerevisiae." *Molecular Genetics and Genomics* 276 (2): 147–61. https://doi.org/10.1007/s00438-006-0130-z.

Noto, Tomoko, and Kazufumi Mochizuki. 2018. "Small RNA-Mediated Trans-Nuclear and Trans-Element Communications in *Tetrahymena* DNA Elimination." *Current Biology* 28 (12): 1938-1949.e5. https://doi.org/10.1016/j.cub.2018.04.071.

Nowacki, Mariusz, Brian P Higgins, Genevieve M Maquilan, Estienne C Swart, Thomas G Doak, and Laura F Landweber. 2009. "A Functional Role for Transposases in a Large Eukaryotic Genome." *Science* 324 (5929): 935–38. https://doi.org/10.1126/science.1170023.

Postberg, Jan, Franziska Jönsson, Patrick Philipp Weil, Aneta Bulic, Stefan Andreas Juranek, and Hans-Joachim Lipps. 2018. "27nt-RNAs Guide Histone Variant Deposition via 'RNA-Induced DNA Replication Interference' and Thus Transmit Parental Genome Partitioning in Stylonychia." *Epigenetics & Chromatin* 11 (1): 31. https://doi.org/10.1186/s13072-018-0201-5.

Sandoval, Pamela Y, Estienne C Swart, Miroslav Arambasic, and Mariusz Nowacki. 2014. "Functional Diversification of Dicer-like Proteins and Small RNAs Required for Genome Sculpting." *Developmental Cell* 28 (2): 174–88. https://doi.org/10.1016/j.devcel.2013.12.010.

Schoeberl, Ursula E, Henriette M Kurth, Tomoko Noto, and Kazufumi Mochizuki. 2012. "Biased Transcription and Selective Degradation of Small RNAs Shape the Pattern of DNA Elimination in *Tetrahymena*." *Genes & Development* 26 (15): 1729–42. https://doi.org/10.1101/gad.196493.112.

Swart, Estienne C, John R Bracht, Vincent Magrini, Patrick Minx, Xiao Chen, Yi Zhou, Jaspreet S Khurana, et al. 2013. "The *Oxytricha Trifallax* Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes." *PLoS Biology* 11 (1): e1001473. https://doi.org/10.1371/journal.pbio.1001473.

Vogt, Alexander, and Kazufumi Mochizuki. 2013. "A Domesticated PiggyBac Transposase Interacts with Heterochromatin and Catalyzes Reproducible DNA Elimination in

Tetrahymena." *PLoS Genetics* 9 (12): e1004032. https://doi.org/10.1371/journal.pgen.1004032.

Chapter 5

Genome editing excisase origins illuminated by somatic genome of *Blepharisma*

Minakshi Singh[1], Kwee Boon Brandon Seah[1], Christiane Emmerich[1], Aditi Singh[1], Christian Woehle[2], Bruno Huettel[2], Adam Byerly[3], Naomi Alexandra Stover[4], Mayumi Sugiura[5], Terue Harumoto[5], Estienne Carl Swart[1]

[1] Max Planck Institute for Biology, Tübingen, Germany

[2] Max Planck Genome Center Cologne, Max Planck Institute for Plant Breeding, Cologne, Germany

[3] Department of Computer Science and Information Systems, Bradley University, Peoria IL, USA

[4] Department of Biology, Bradley University, Peoria, IL, USA

[5] Nara Women's University, Nara, Japan

This section is adapted from the BioRxiv preprint DOI:
https://doi.org/10.1101/2021.12.14.471607.

I generated all the data, performed the preliminary genome assemblies and preliminary gene annotation (Chapter 3) leading up to the draft assembly of the somatic genome and performed functional annotation. I executed the developmental time series and analyzed the gene expression by integrating transcriptomic data, partially described in Chapter 4. I performed the phylogenetic analysis, gene expression analysis and annotation of the transposase protein domains.
Dr. Estienne Swart compiled the final version of the draft assembly of the somatic genome. Additional analysis of the genome was performed by Dr.'s Estienne Swart and Kwee Boon Brandon Seah. Dr.'s Estienne Swart and Kwee Boon Brandon Seah reviewed and edited the manuscript. Details of all author contributions are listed in Appendix A. 1.

5.1. Introduction

DNA excision in ciliates is a spectacular and widespread form of natural genome editing with profound consequences for what germline and somatic genomes mean (Arnaiz et al., 2012; Chen et al., 2014; Hamilton et al., 2016; Swart and Nowacki, 2015). Though the responsible processes are under active study, much remains to be learnt from these master DNA manipulators, including how and why this remarkable situation arose in them.

Knowledge of ciliate genome editing mechanisms is dominated by *Tetrahymena* and *Paramecium* (class Oligohymenophorea), with additional input from *Oxytricha*, *Stylonychia* and *Euplotes* (class Spirotrichea) (Chalker et al., 2013; Vogt et al., 2013). The remaining nine ciliate classes await detailed characterization. To advance investigation of natural genome editing and tackle questions about its origin we focused on the ciliate species *Blepharisma stoltei*. Together with its sister-class, Karyorelictea, the class Heterotrichea, to which this ciliate species belongs, represent the earliest branching ciliate lineages, more distantly related to current model ciliates than those models are to each other (Lynn, 2010). Furthermore, the genus *Blepharisma* exhibits distinctive alternative somatic nuclear developmental pathways, which have the potential to disentangle genome editing processes from indirect influences of preceding pathways.

As in model ciliates, we show in an accompanying paper (Chapter 6) that MIC-specific sequences are removed to form a functional *Blepharisma* MAC genome (Seah, et al. 2022). Like other ciliates, the resulting MAC genome appears to have been freed of mobile elements and other forms of junk DNA contained in the MIC genome (Klobutcher and Herrick, 1997). However, this situation is an oversimplification of the actual MAC genome content (Chapter 6). In the best studied ciliates, genome editing is thought to be coordinated or assisted by small RNAs (sRNAs) (Chalker et al., 2013). Specific MIC-limited DNA segments — internally eliminated sequences (IESs) — are excised by domesticated transposases (Arnaiz et al., 2012; Chalker et al., 2013; Klobutcher and Herrick, 1995; Prescott, 1994). Large scale genome-wide DNA amplification accompanies genome editing, producing thousands of copies in mature MACs of larger ciliate species (Klobutcher and Herrick, 1997; Prescott, 1994).

Here we provide essential somatic genome and transcriptomic resources for *B. stoltei*. From long-read sequencing, the *B. stoltei* MAC genome appears to be organized as numerous alternative minichromosomes. Among *Blepharisma*'s MAC-encoded transposase genes we identified were PiggyBac transposase homologs, which, thus far only reported in the distantly

related ciliates *Paramecium* and *Tetrahymena*. A few *Blepharisma* PiggyBac homologs are substantially upregulated in MAC development, including the main candidate IES excisase. Consistent with ancient origins of ciliate genome editing, *Blepharisma* shares pronounced development-specific upregulation of homologs known to be involved in this process. *Blepharisma* therefore represents an invaluable outgroup for investigations of genome editing evolution.

5.2. Results

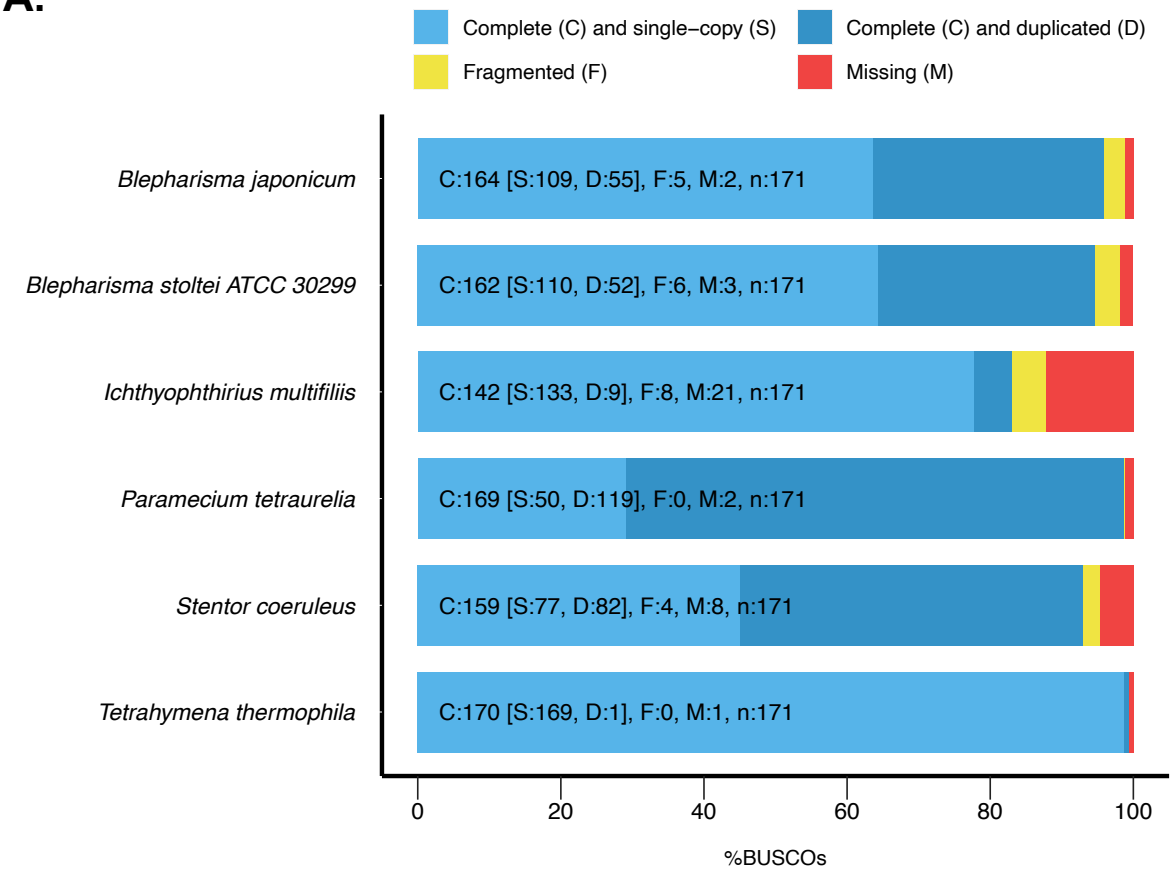
5.2.1. A compact somatic genome with a minichromosomal architecture

The draft *Blepharisma stoltei* ATCC 30299 MAC genome is compact (41 Mb) and AT rich (66%), like most sequenced ciliate MAC genomes (Figure 5.1A, Table S5.1, S5.2,). The genome is gene-dense (25,711 predicted genes), with short intergenic regions, tiny, predominantly 15 and 16 bp introns (Figure 5.2) and untranslated regions (UTRs) (Figure 5.3A). *B. stoltei* uses an alternative nuclear genetic code with UGA codons reassigned from stops to tryptophan (Figure 1B).

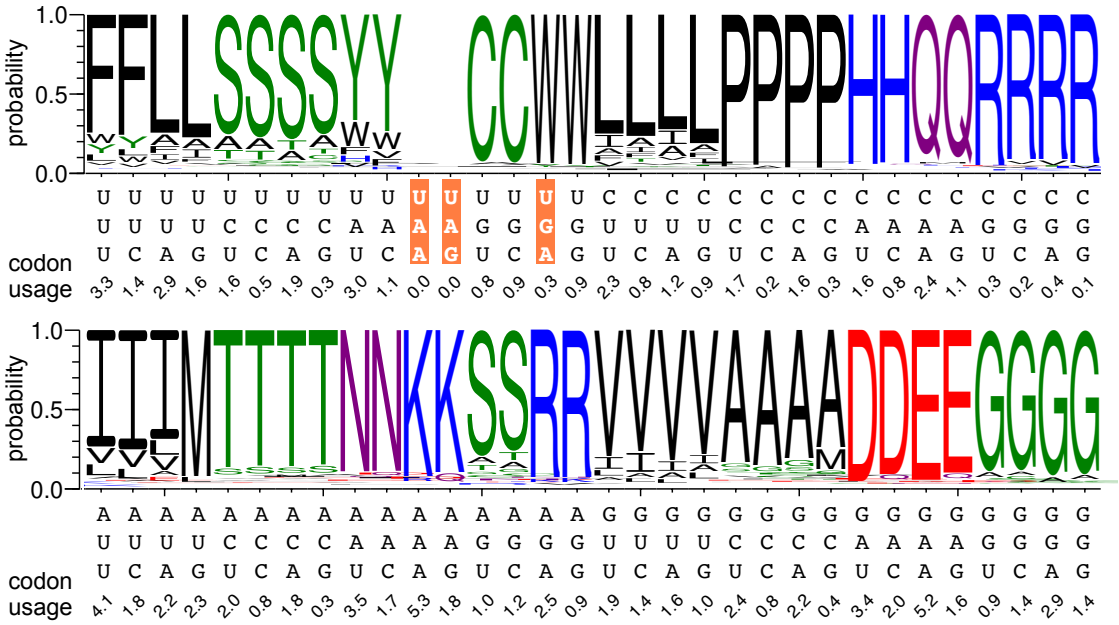
From joint variant calling of reads from strains ATCC 30299 and HT-IV, strain ATCC 30299 appears to be virtually homozygous, with only 1277 heterozygous single-nucleotide polymorphisms (SNPs) compared to 193725 in strain HT-IV (i.e., individual heterozygosity of 3.08×10^{-5} vs. 4.67×10^{-3} respectively). Low SNP levels were likely beneficial for overall genomic contiguity, since heterozygosity poses significant algorithmic challenges for assembly software (Chin et al., 2016). For brevity's sake, we refer to this genome as the *Blepharisma* MAC genome (and "*Blepharisma*" for the associated strain). Though the final assembly comprises 64 telomere-to-telomere sequences, MAC chromosome boundaries cannot be defined given the extensive natural fragmentation of the *Blepharisma* MAC genome (characterized in the upcoming sections), hence we simply refer to "contigs".

Figure 5.1. Analysis of assembly completeness and genetic code. A. Completeness of the *B. stoltei* ATCC 30299 MAC assembly was estimated by the percentage of BUSCOs found in the assembly with reference to the OrthoDB v10 alveolate database. The nature of the ortholog-matches is indicated by characters followed by counts: C (complete orthologs) - light blue, D (duplicated orthologs) - dark blue, F (fragmented orthologs) - yellow and M (missing orthologs) - red. B. Prediction for *B. stoltei* ATCC 30299 MAC genome by PORC; codons that are stops in the standard genetic code are highlighted in orange.

A.


Complete (C) and single−copy (S)
Complete (C) and duplicated (D)
Fragmented (F)
Missing (M)
Blepharisma japonicum
C:164 [S:109, D:55], F:5, M:2, n:171
Blepharisma stoltei ATCC 30299
C:162 [S:110, D:52], F:6, M:3, n:171
Ichthyophthirius multifiliis
C:142 [S:133, D:9], F:8, M:21, n:171
Paramecium tetraurelia
C:169 [S:50, D:119], F:0, M:2, n:171
Stentor coeruleus
C:159 [S:77, D:82], F:4, M:8, n:171
Tetrahymena thermophila
C:170 [S:169, D:1], F:0, M:1, n:171
0
20
40
60
80
100
%BUSCOs

B.


probability
1.0
0.5
0.0
codon usage

5.2.1.1. *Blepharisma has short spliceosomal introns*

Blepharisma introns are mostly (97%) 15 or 16 nucleotides (nt) long, like those of *Stentor* (Figure 5.2D). Though intron reduction (7389 introns predicted in the reference *B. stoltei* MAC genome, i.e., 0.29 introns per gene) is not as extreme as some other microbial eukaryotes, like Giardia lamblia (Roy et al. 2012), where almost all have been lost, both *Blepharisma* and *Stentor* have much fewer introns relative to other ciliates (e.g., intron densities of 1.6, 2.3 and 4.8 introns per gene in *Paramecium*, *Oxytricha* and *Tetrahymena*, respectively(Bondarenko and Gelfand 2016)) and to the putative, relatively intron-rich eukaryotic common ancestor (Csuros et al. 2011), along with their extreme length reduction.

Blepharisma 15 nt introns possess a characteristic branch-point "A", as would be expected in classical models of lariat formation during mRNA splicing (Figure 5.2C). 16 nt introns almost invariably have an "A" at either 10 or 11 nt downstream of the donor site (i.e., only one of 499 does not, but has "A" at 9 nt), although this is not obvious in the consensus sequence logo because the position is variable (Figure 5.2D). Similarly, 17 nt introns all possess "A" at 10-12 nt downstream of the donor site. Only a few intron bases, 5-8 and 12, of *Blepharisma*'s 15 nt introns are relatively unconstrained (Figure 5.2C). This leaves little room for the presence of any additional regulatory elements in the mRNA or underlying DNA.

In the final gene predictions, just over 1% of predicted *Blepharisma* introns lack canonical GT-AG boundaries (62 out of 4670 introns). Just under half of these (30) are 15 or 16 bp long and predominantly appear to represent true spliceosomal introns. The boundaries of two predicted introns with CT-AC boundaries (14 and 15 nt in length) resulted from misalignment of nucleotides in the mapped spliced reads at conventional GT-AG junctions. We found no evidence of minor spliceosomal RNAs (U11, U12, U4atac, and U6atac) using Infernal searches (E-value < 10).

Visual inspection of the mapped RNA-seq data to the non-canonical *Blepharisma* introns and predicted coding sequences suggests that the GC-AG, GT-GG and GG-AG introns are correct, i.e., lead to prediction of complete coding sequences downstream of their locations. Lower frequency alternative splicing may occur in some cases (e.g. Figure S5.4G), but these generate prematurely terminated coding sequences.

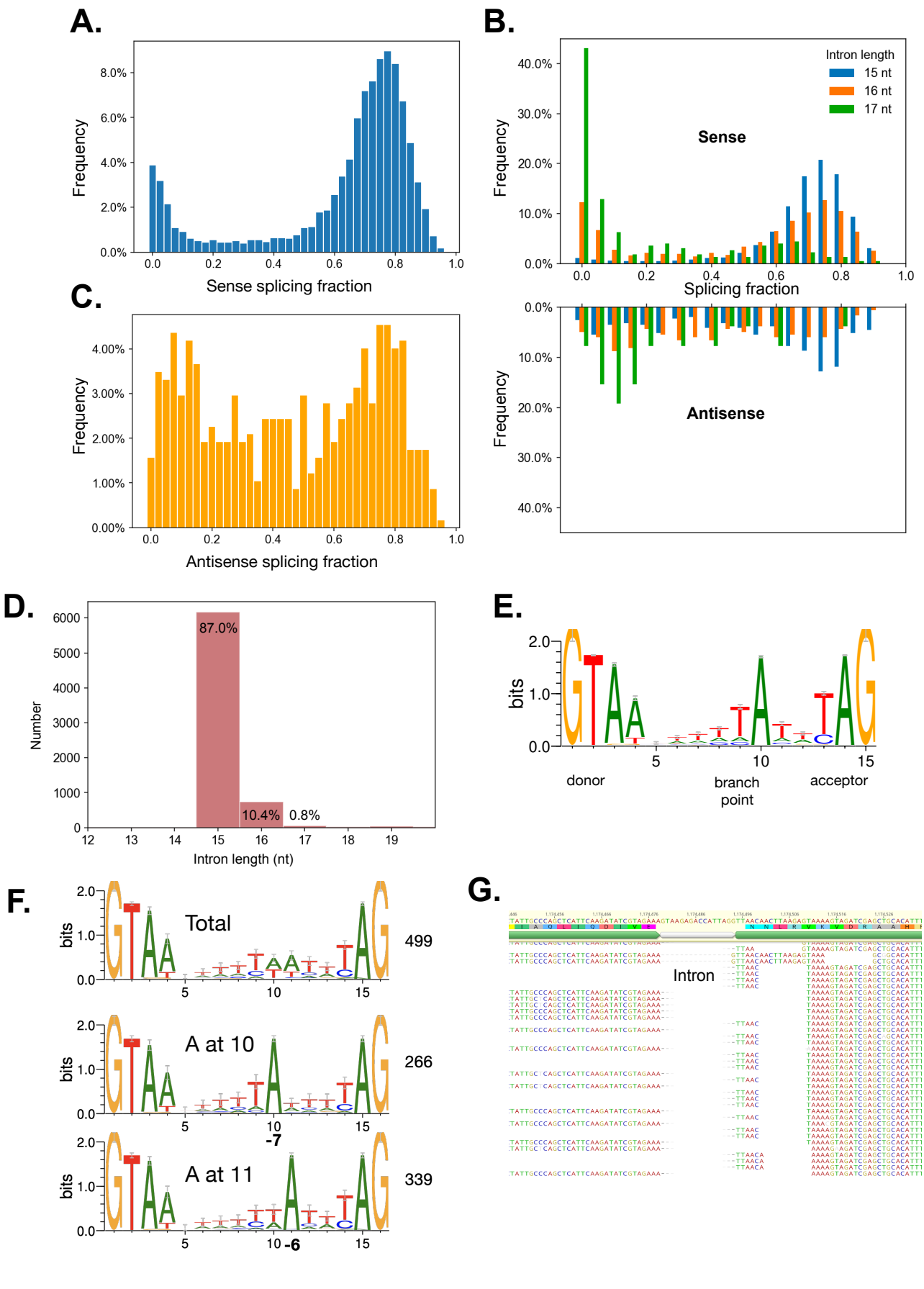

A.
Frequency
Sense splicing fraction
B.
Intron length
15 nt
16 nt
17 nt
Sense
Antisense
Frequency
Splicing fraction
C.
Antisense splicing fraction
D.
Number
87.0%
10.4%
0.8%
Intron length (nt)
E.
bits
donor
branch point
acceptor
F.
Total
A at 10
A at 11
499
266
339
G.
Intron

Figure 5.2. Intron splicing. A. Distribution of intron splicing fraction of candidate sense introns in the *B. stoltei* MAC genome. B. Distribution of intron splicing fractions of introns according to intron lengths. C. Distribution of intron splicing fraction of candidate antisense introns. D. Distribution of intron lengths from predicted genes. E. Sequence logos for 15 bp introns (splicing frequency > 0.5). F. Sequence logos for all predicted 16 nt introns, and 16 nt introns with "A" at either position -7 or -6 (counting from the 3' end). The number of introns underlying the logos are indicated to the right. G. Distribution of intron splicing fractions of introns according to intron lengths. H. Sample of RNA-seq reads mapped to a GT-GG intron from gene BSTOLATCC_MAC21551 (Contig_57.g761). Translation in alternative reading frames downstream of the predicted intron leads to premature stop codons soon after the intron.

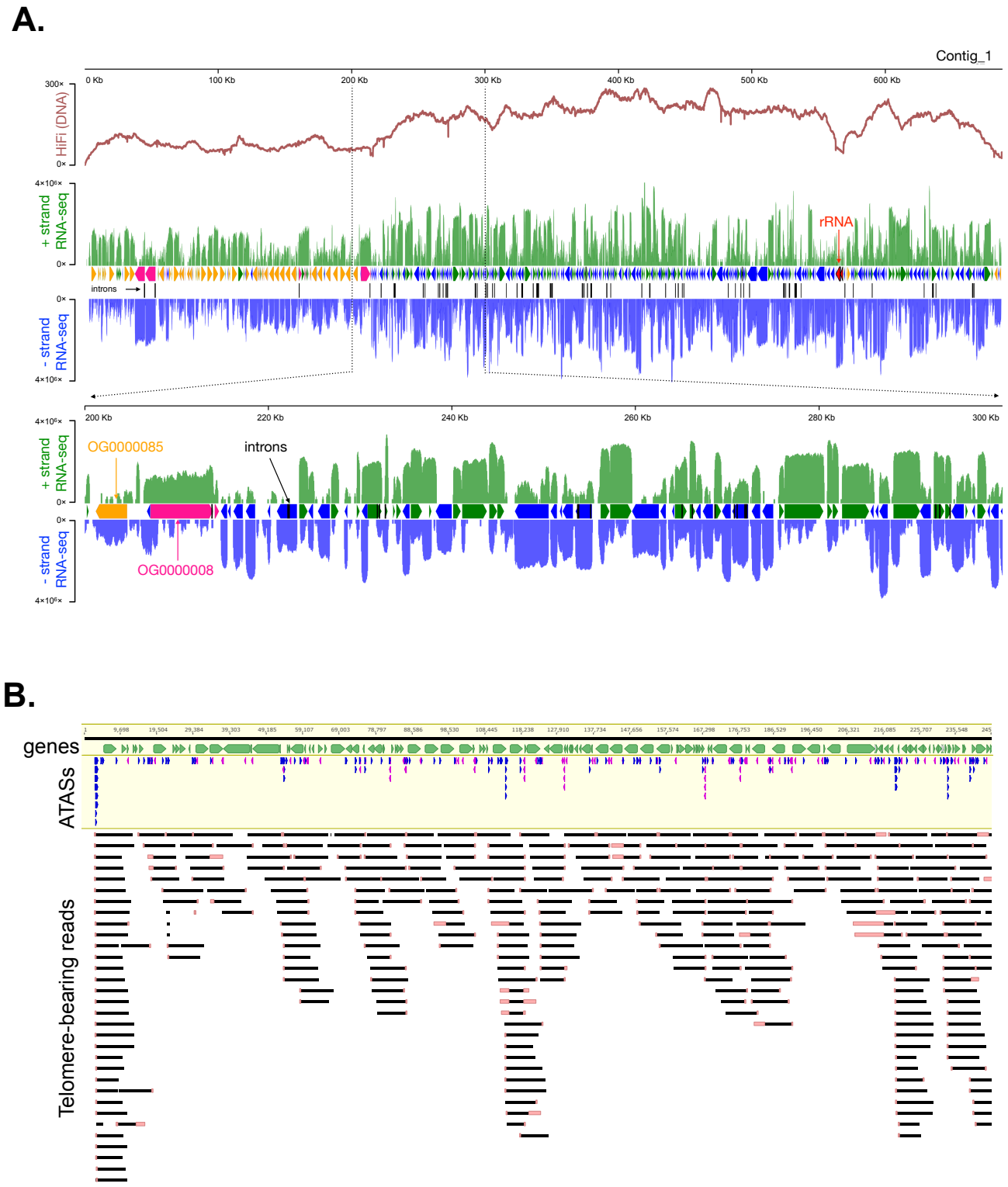


Figure 5.3. A gene-dense somatic genome with a minichromosomal architecture. A. HiFi (DNA) and RNA-seq coverage across a representative *B. stoltei* ATCC30299 MAC genome contig (Contig_1). Y scale is linear for HiFi reads and logarithmic (base 10) for RNA-seq. Plus strand (relative to the contig) RNA-seq coverage is green; minus strand RNA-seq coverage is blue. Between the RNA-seq coverage graphs each arrow represents a predicted gene. Two orthogroups classified by OrthoFinder are shown. B. Mapping of a subset telomere-containing HiFi reads to a *B. stoltei* MAC genome contig region, with alternative telomere addition sites (ATASs) shown by blue (5') or mauve (3') arrows. Pink bars at read ends indicate soft-masking, typically of telomeric repeats.

5.2.1.2. *Blepharisma has minichromosomal architecture in the somatic genome*

The basic telomere unit of *Blepharisma* is a permutation of CCCTAACA, like its heterotrich relative *Stentor coeruleus* (Slabodnick et al., 2017) (Figure 5.4C). Since a compelling candidate for a telomerase ncRNA (TERC) could not be found in either *Blepharisma* or *Stentor* using Infernal (Nawrocki et al., 2009) and RFAM models (RF00025 - ciliate TERC; RF00024 - vertebrate TERC), it was not possible to delimit the repeat ends. Heterotrichs may use a different or very divergent ncRNA. In contrast to the extremely short (20 bp) MAC telomeres of spirotrichs like *Oxytricha* with extreme MAC genome fragmentation (Swart et al., 2013), sequenced *Blepharisma* MAC telomeres are moderately long (Figure 5.4A), with a mode of 209 bp (~26 repeats of the 8 bp motif), extending to a few kilobases.

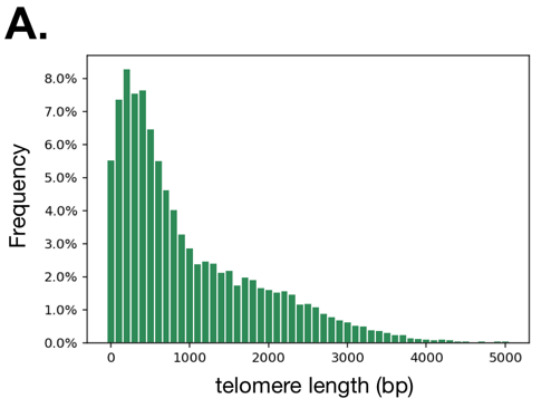
With a moderately strict definition of possessing at least three consecutive telomeric repeats, one in eight reads in the *Blepharisma* HiFi library were telomere-bearing. Telomeric reads are distributed across the entire genome (Figure 5.3B). Typically, a minority of mapped reads are telomere-bearing at individual internal positions, and so we term them alternative telomere addition sites (ATASs) (Figure 5.3B). We identified 46705 potential ATASs, the majority of which (38686) were represented by only one mapped HiFi read.

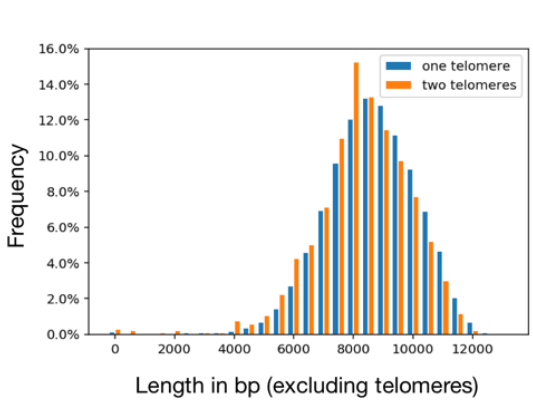
The expected distance between telomeres, and hence the average MAC DNA molecule length, is about 130 kb. This is consistent with the raw input MAC DNA lengths, which were mostly longer than 10 kb and as long as 1.5 Mb (Figure 3.4), and the small fraction (1.3%) of *Blepharisma*'s HiFi reads bound by telomeres on both ends. Excluding the length of the telomeres, telomere-bound reads may be as short as 4 kb (Figure 5.4B). Given the frequency of telomere-bearing reads, we expect many additional two-telomere DNA molecules longer than 12 kb, the approximate maximum length of the HiFi reads (Figure 5.4A). Since the lengths of the sequenced two-telomere DNA molecules on average imply that they encode multiple genes, we propose classifying them as "minichromosomes". This places them between the "nanochromosomes" of ciliates like *Oxytricha* and *Stylonychia*, which typically encode single genes and a few kilobases long (Aeschlimann et al., 2014; Swart et al., 2013), and *Paramecium tetraurelia* and *Tetrahymena thermophila* MAC chromosomes which are hundreds of kilobases to megabases long (Aury et al., 2006; Sheng et al., 2020; Zagulski et al., 2004). The *Paramecium bursaria* MAC genome is considerably more fragmented than those of other previously examined *Paramecium* species, and have thus also been classified as minichromosomes (Cheng et al., 2020).

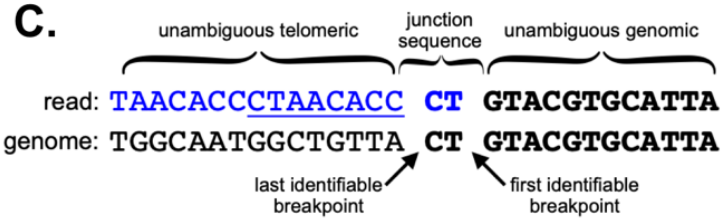
Beyond the first 2-5 bp corresponding to the junction sequences, the average base composition on the chromosome flanking ATAS junctions shows an asymmetrical bias (Figure 5.4D). From position +6 onwards there is an enrichment of T to about 40% and A to 35-39%, compared to the genome-wide frequencies of 33% each. At position +19 to +23, there is a slight decrease in T to 37-39%. AT values gradually decline back to about 35% each by position +150. Correspondingly, G and C are depleted downstream of ATAS junctions, dropping to a minimum of 8.6% and 11% respectively around position +37, compared to the genome-wide average of 17% each. AT enrichment and GC depletion upstream of ATAS junctions are less pronounced.

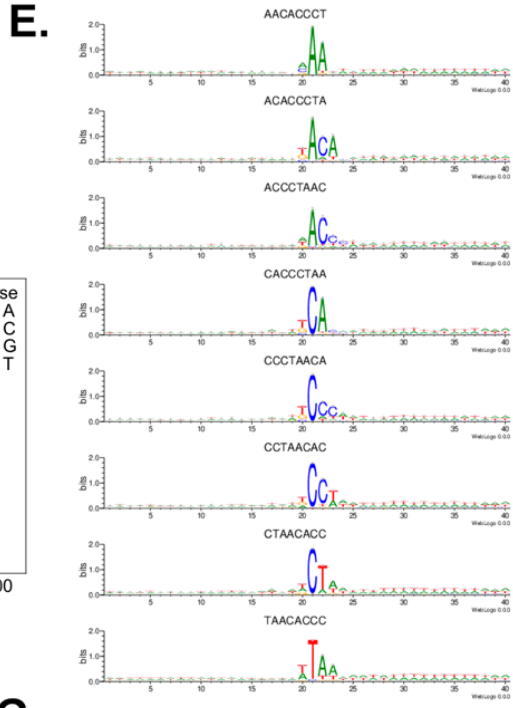
If breakage and chromosome healing were random, we would not expect such an asymmetry. This suggests that there is a nucleotide bias, whether in the initiation of breaks, telomere addition, or in the processing of breaks before telomere addition. However, we have not yet identified any conserved motif like the 15 bp chromosome breakage site (CBS) in *Tetrahymena* (Yao et al. 1990) nor a short 10-bp sequence periodicity in base composition like in *Oxytricha trifallax* (Cavalcanti et al. 2004). Therefore, telomere addition in *B. stoltei* appears to involve base-pairing of short segments of about 2 bp between the telomere and chromosome, with a bias centered on the "CT" in the telomere unit, and an asymmetrical preference for AT-rich sequences on the chromosomal side of the junction.

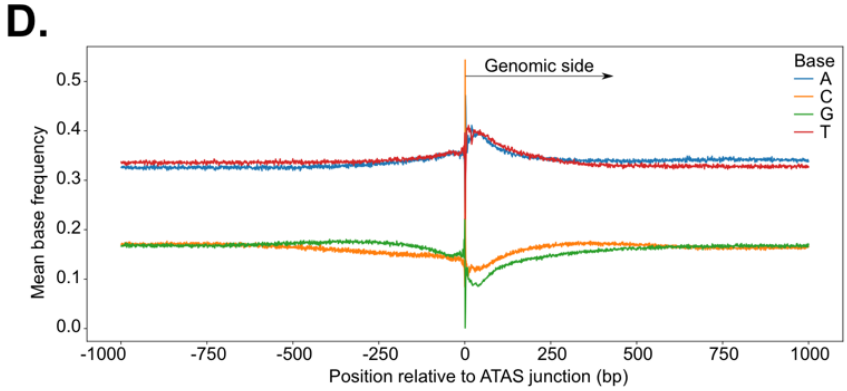
Figure 5.4. Properties of minichromosomes, telomeres, and alternative telomere addition sites. A. Length distribution of telomeres of telomere-bearing HiFi reads. B. Length distribution of HiFi reads delimited by telomeres. C. Diagram of a telomere-bearing read mapped onto genome reference at an ATAS. Sequence which is ambiguously chromosomal or telomeric is "junction sequence"; junction coordinate which maximizes telomere repeat length on the read is the "first identifiable breakpoint"; the coordinate maximizing alignment length to reference is the "last identifiable breakpoint". The last telomeric unit permutation at the last identifiable breakpoint is underlined (length 8 bp). D. Mean base frequencies in +/- 1 kbp flanking ATAS junctions. E. Sequence logos of chromosomal sequence at ATAS junctions, sorted by which permutation of the telomeric repeat is present (plot labels). Logos are aligned to the "last identifiable breakpoint" between positions 20 and 21; telomeric repeats on telomere-bearing reads begin to the left of the breakpoint. F. Frequencies of 2-mers in whole genome (blue), in telomeres (green), and at ATAS junctions (chromosomal side after last identifiable breakpoint, orange). G. Histogram of junction sequence lengths for ATASs in *B. stoltei*. H. Counts of each telomere repeat permutation at ATAS junctions (last identifiable breakpoint).

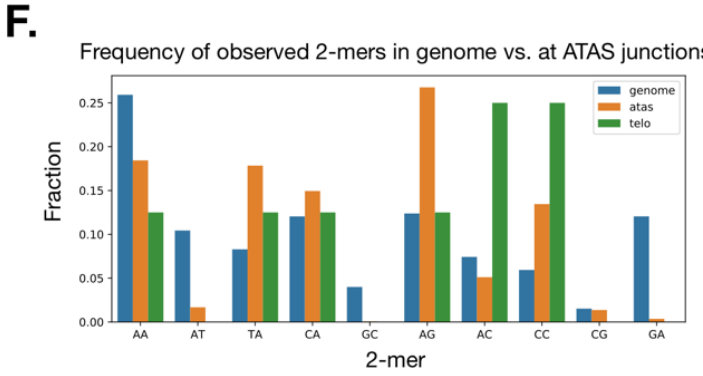

A.
Frequency
telomere length (bp)

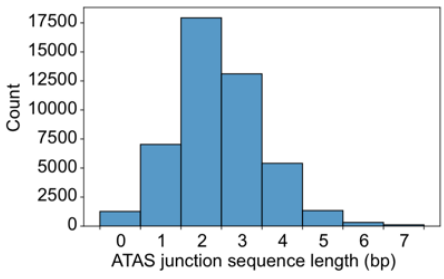

B.
one telomere
two telomeres
Frequency
Length in bp (excluding telomeres)

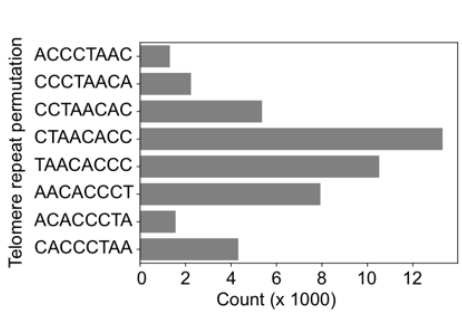

C.
unambiguous telomeric
junction sequence
unambiguous genomic
read: TAACACCCTAACACC CT GTACGTGCATTA
genome: TGGCAATGGCTGTTA CT GTACGTGCATTA
last identifiable breakpoint
first identifiable breakpoint


E.
AACACCCT
ACACCCTA
ACCCTAAC
CACCCTAA
CCCTAACA
CCTAACAC
CTAACACC
TAACACCC


D.
Genomic side
Base
A
C
G
T
Mean base frequency
Position relative to ATAS junction (bp)


F.
Frequency of observed 2-mers in genome vs. at ATAS junctions
genome
atas
telo
Fraction
AA
AT
TA
CA
GC
AG
AC
CC
CG
GA
2-mer


G.
Count
ATAS junction sequence length (bp)


H.
Telomere repeat permutation
ACCCTAAC
CCCTAACA
CCTAACAC
CTAACACC
TAACACCC
AACACCCT
ACACCCTA
CACCCTAA
Count (x 1000)

Most ATAS junctions in *B. stoltei* have an overlapping junction sequence, on average 2-3 bp long (Figure 5.4G). This can also be observed when separate sequence logos are drawn for each of the possible telomere repeat permutations observed at the ATAS junction (Figure 5.4E). Such a short overlap of a few base pairs between telomere repeat and chromosome sequence is similar to what has been observed in other organisms, such as 3-5 bp in yeast (Putnam et al. 2004) and 2-4 bp in humans (Morin 1991). This is in contrast to *Tetrahymena* where telomeres are often added to sites that have no homology to the telomere sequence (Wang and Blackburn 1997).

We hypothesized that the location of ATAS junctions in the genome might be randomly distributed and simply reflect the baseline sequence composition of the genome and/or the telomeres. To test this, we counted the frequency of 2-mers in the MAC genome (excluding telomeric regions) and in the telomere repeats, and compared them to the 2-mer frequencies observed at ATAS junctions (2 bp on chromosomal side of last identifiable breakpoints, Figure 5.4F). Sequence composition of the telomeres does have a strong influence, as 2-mers that are not represented in the telomeres (AT, GC, CG, GA) are poorly represented at ATAS junctions even though they may be frequent in the genome, e.g. GA, 12.0% in genome vs. 0.36% at ATAS; AT, 10.4% vs. 1.7%. However, 2-mer frequencies at ATAS junctions do not match frequencies in the telomeres closely either. For example, the 2-mer AG is about twice as frequent at ATAS junctions as compared to telomeres, and as compared to the genome generally. Instead, the telomere permutations at ATAS junctions are not uniformly distributed; the permutation CTAACACC is the most common, followed by its adjacent permutations TAACACCC and AACACCCT (using last identifiable breakpoints, Figure 5.4H). These would account for the three most common 2-mers at ATAS junctions: AG (canonical form of CT), AA, and TA.

5.2.2. PiggyBac transposases in the somatic and germline genomes of *Blepharisma*

We examined gene expression trends across development of the PiggyBac transposases found in the somatic genome of *Blepharisma* (Chapter 4). A detailed account of conjugation and the changes in nuclear morphology which accompany it is provided in Chapter 4 .

5.2.2.1. A single Blepharisma PiggyBac homolog has a complete catalytic triad

We detected distinct PFAM identifiers for transposase domains in *Blepharisma* (Figure 5.5A). Using HMMER searches with the domain characteristic of PiggyBac homologs, DDE_Tnp_1_7 (PF13843), we found eight homologs in the *B. stoltei* ATCC MAC genome and five additional ones within IESs, none of which were flanked by terminal repeats (identified by RepeatModeler). We also found PiggyBac homologs in the MAC genomes of *B. stoltei* HT-IV and *B. japonicum* R1072.

Reminiscent of *Paramecium tetraurelia*, which, among ten PiggyMac homologs, has just one homolog with a complete catalytic triad (Bischerour et al., 2018), the DDD triad is preserved in just a single *Blepahrisma* PiggyBac homolog (Figure 5.5B; Contig_49.g1063, BSTOLATCC_MAC17466). This gene is strongly upregulated during development from 22 to 38 h, when new MACs develop and IES excision is required (Figure 5.5B). In a multiple sequence alignment the canonical catalytic triad second aspartate of a lower-expressed, MIC-limited PiggyBac is offset by one amino acid (Supplemental data S5).

There are significant similarities in the basic properties of *Blepharisma* and *Paramecium* IESs, detailed in the *Blepharisma* MIC genome report (Chapter 6). Consequently, adopting the *Paramecium* nomenclature, we refer to the primary candidate IES excisase as *Blepharisma* PiggyMac (BPgm) and the other somatic homologs as BPgm-Likes (BPgmLs). By extension, we refer to their close relatives which are germline-limited as PiggyMics (Figure 5.5.B). Other than the PFAM DDE_Tnp_1_7 domain, three *Blepharisma* MAC genome-encoded PiggyBac homologs also possess a short, characteristic cysteine-rich domain (CRD) (Figure 5.5C), which is absent from the other BPgmLs and PiggyMics. PiggyBac CRDs have been classified into three different groups and are essential for *Paramecium* IES excision (Guérineau et al., 2021). In *Blepharisma*, the CRD consists of five cysteine residues arranged as CxxC-CxxCxxxxH-Cxxx(Y)H (where C, H, Y and x respectively denote cysteine, histidine, tyrosine and any other residue).

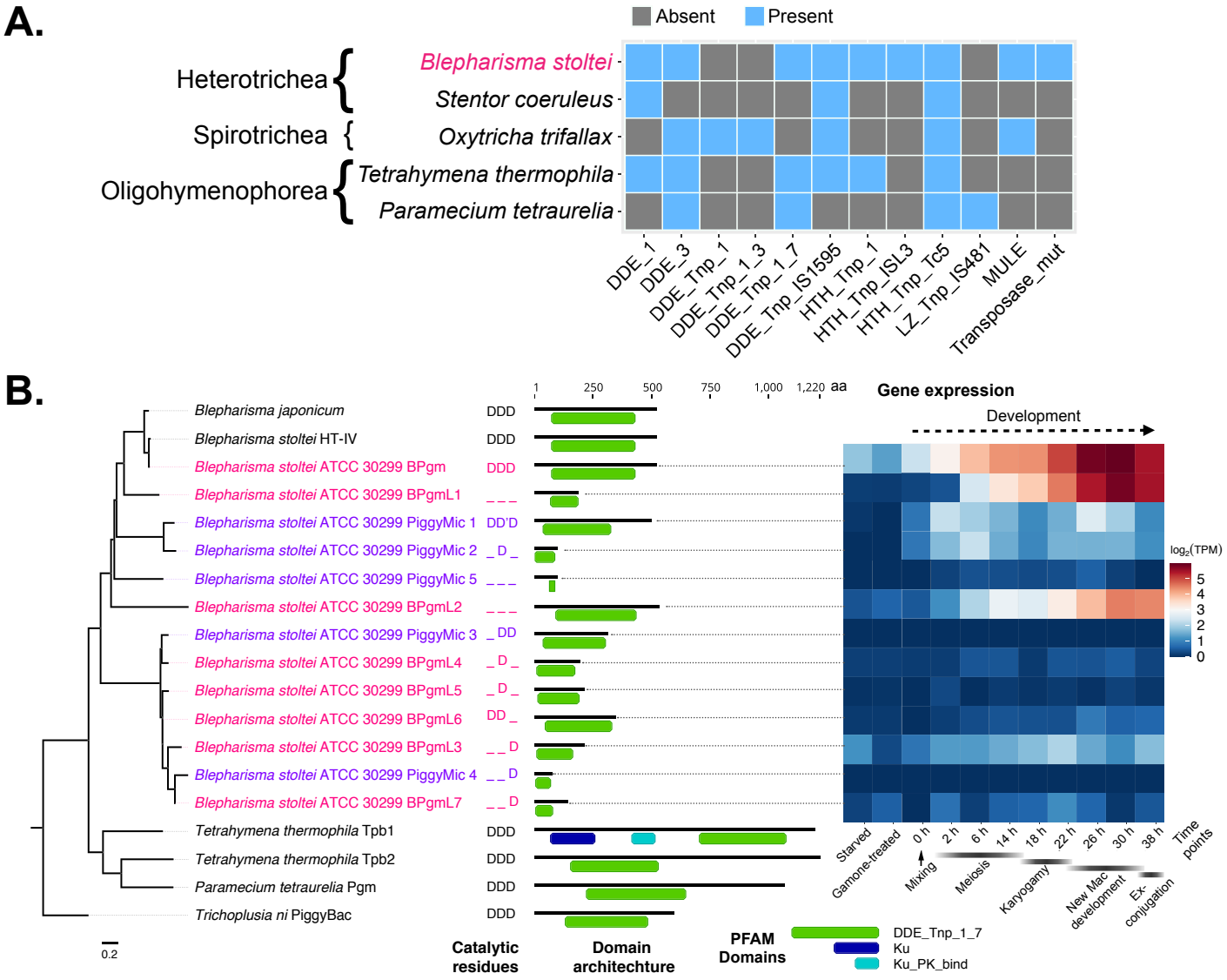

A.
Absent
Present
Heterotrichea
Spirotrichea
Oligohymenophorea
Blepharisma stoltei
Stentor coeruleus
Oxytricha trifallax
Tetrahymena thermophila
Paramecium tetraurelia
DDE_1
DDE_3
DDE_Tnp_1
DDE_Tnp_1_3
DDE_Tnp_1_7
DDE_Tnp_IS1595
HTH_Tnp_1
HTH_Tnp_ISL3
HTH_Tnp_Tc5
LZ_Tnp_IS481
MULE
Transposase_mut
B.
1
250
500
750
1,000
1,220 aa
Gene expression
Development
log2(TPM)
5
4
3
2
1
0
Blepharisma japonicum DDD
Blepharisma stoltei HT-IV DDD
Blepharisma stoltei ATCC 30299 BPgm DDD
Blepharisma stoltei ATCC 30299 BPgmL1 _ _ _
Blepharisma stoltei ATCC 30299 PiggyMic 1 DD'D
Blepharisma stoltei ATCC 30299 PiggyMic 2 _ D _
Blepharisma stoltei ATCC 30299 PiggyMic 5 _ _ _
Blepharisma stoltei ATCC 30299 BPgmL2 _ _ _
Blepharisma stoltei ATCC 30299 PiggyMic 3 _ DD
Blepharisma stoltei ATCC 30299 BPgmL4 _ D _
Blepharisma stoltei ATCC 30299 BPgmL5 _ D _
Blepharisma stoltei ATCC 30299 BPgmL6 DD _
Blepharisma stoltei ATCC 30299 BPgmL3 _ _ D
Blepharisma stoltei ATCC 30299 PiggyMic 4 _ _ D
Blepharisma stoltei ATCC 30299 BPgmL7 _ _ D
Tetrahymena thermophila Tpb1 DDD
Tetrahymena thermophila Tpb2 DDD
Paramecium tetraurelia Pgm DDD
Trichoplusia ni PiggyBac DDD
0.2
Starved
Gamone-treated
0 h
2 h
6 h
14 h
18 h
22 h
26 h
30 h
38 h
Time points
Mixing
Meiosis
Karyogamy
New Mac development
Ex-conjugation
Catalytic residues
Domain architechture
PFAM Domains
DDE_Tnp_1_7
Ku
Ku_PK_bind
C.
Tni
Hvi
Ago
Oni
Pny
Tru_Pigibaku
Cag
Har
Bmo
Blepharisma japonicum
Blepharisma stoltei HT-IV
Blepharisma stoltei ATCC 30299 BPgm
Blepharisma stoltei ATCC 30299 BPgmL2
Blepharisma stoltei ATCC 30299 BPgmL3
Condylostoma magnum Pgm1
Condylostoma magnum Pgm2
Condylostoma magnum Pgm3
Condylostoma magnum Pgm4
Condylostoma magnum Pgm5
Mlu_Piggybat
Hs-Pgbd2
Hs-Pgbd3
Hs-Pgbd4
PLE-wu
Pgm
PgmL1
PgmL2
PgmL3a
PgmL3b
PgmL3c
PgmL4a
PgmL4b
PgmL5a
PgmL5b
Tpb1
Tpb2
Tpb6
Tpb7
Lia5
Paramecium tetraurelia
Tetrahymena thermophila

Figure 5.5. MAC genome-encoded transposases in ciliates and properties of a putative *Blepharisma* IES excisase. A. Presence/absence matrix of PFAM transposase domains detected in predicted MAC genome-encoded ciliate proteins. Ciliate classes are indicated before the binomial species names. B. DDE_Tnp_1_7 domain phylogeny with PFAM domain architecture and gene expression heatmap for *Blepharisma*. "Mixing" indicates when cells of the two complementary mating types were mixed. Outgroup: PiggyBac element from *Trichoplusia ni*. Catalytic residues: D-aspartate, D'- aspartate residue with 1 aa translocation. C. Cysteine-rich domains of PiggyBac homologs. PBLE transposases: Ago (*Aphis gossypii*); Bmo (*Bombyx mori*); Cag (*Ctenoplusia agnata*); Har (*Helicoverpa armigera*); Hvi (*Heliothis virescens*); PB-Tni (*Trichoplusia ni*); Mlu (PiggyBat from *Myotis lucifugus*); PLE-wu (*Spodoptera frugiperda*). Domesticated PGBD transposases: Oni (*Oreochromis niloticus*); Pny (*Pundamilia nyererei*); Lia5, Tpb1, Tpb2, Tpb6 and Tpb7 (*Tetrahymena thermophila*); Pgm, PgmL1, PgmL2, PgmL3a/b/c, PgmL4a/b, PgmL5a/b (*Paramecium tetraurelia*); Tru (*Takifugu rubripes*); Pgbd2, Pgbd3 and Pgbd4 (*Homo sapiens*).

Two *Blepharisma* homologs possess this CRD without the penultimate tyrosine residue, while the third contains a tyrosine residue before the final histidine. This -YH feature towards the end of the CxxC-CxxCxxxxH-Cxxx(Y)H CRD is shared by all the PiggyBac homologs we found in *Condylostoma*, the bat PiggyBac-like element (PBLE) and human PiggyBac element-derived (PGBD) proteins PGBD2 and PGBD3. In contrast, PiggyBac homologs from *Paramecium* and *Tetrahymena* have a CRD with six cysteine residues arranged in the variants of the motif CxxC-CxxC-Cx{2-7}Cx{3,4}H, and group together with human PGDB4 and *Spodoptera frugiperda* PBLE (Figure 5.5C).

5.2.2.2. PiggyBac transposases are subject to purifying selection

Previous experiments involving individual or paired gene knockdowns of most of the ten *Paramecium tetraurelia* PiggyMac(-like) paralogs led to substantial IES retention, even though only one PiggyMac gene (Pgm) has the complete catalytic triad, indicating that all these proteins are functional (Bischerour et al., 2018). To examine functional constraints on *Paramecium* PiggyMac homologs we examined non-synonymous (d_N) to synonymous substitution rates (d_S), i.e. $\omega = d_N/d_S$, for pairwise codon sequence alignments using two closely related *Paramecium* species (*P. tetraurelia* and *P. octaurelia*). All d_N/d_S values for pairwise comparisons of each of the catalytically incomplete *P. tetraurelia* PgmLs versus the complete Pgm, were less than 1, ranging from 0.01 to 0.25 (Table S5.3). All d_N/d_S values for pairwise comparisons between *P. tetraurelia* and *P. octaurelia* PiggyBac orthologs were also substantially less than 1, ranging from 0.02 to 0.11 (Table S5.4). Since $d_N/d_S = 1$ indicates genes evolving neutrally (Yang and Nielsen, 2000), none of these genes are likely pseudogenes, and all appear subject to similar purifying selection.

Table 5.1. *Blepharisma* PiggyMac-like substitution rates. Reference gene: Contig_49.g1063

Gene ID	d_N/d_S	d_N	d_S
Contig_3.g998	0.0093	0.7106	76.4367
Contig_13.g879	0.0551	0.8871	16.0867
Contig_13.g927	0.0261	0.5547	21.2267
Contig_17.g391	0.0087	0.8394	96.9223
Contig_17.g392	0.0076	0.8195	107.6866
Contig_60.g827	0.1351	0.8401	6.2209
Contig_61.g932	0.0836	0.7727	9.2391
cORF_Contig_17. g3	0.0765	0.653	8.5395
cORF_Contig_17. g4/5	0.0068	0.5852	85.9998
cORF_Contig_21. g21	0.0697	0.4729	6.7874
cORF_Contig_39. g3	0.2763	1.2445	4.5036
cORF_Contig_39. g3/4	0.007	0.6817	97.5885

Only one of *Blepharisma*'s eight MAC and five MIC PiggyBac homologs has the complete, characteristic DDD triad necessary for catalysis. In pairwise comparisons of each of the MAC homologs with incomplete/missing triads versus the complete one d_N/d_S ranges from 0.0076 to 0.1351 (Table 5.1). The pairwise non-synonymous to synonymous substitution rates of the PiggyMics in comparison to the BPgm were also much less than 1 (range 0.007 to 0.2), indicating they are also subject to purifying selection. We detected PiggyBac homologs in two other heterotrichs, but not the oligohymenophorean *Ichthyophthirius multifiliis* ("Supplemental information").

5.2.3. PiggyBac transposases originated early in ciliate evolution

To determine whether the *Blepharisma* PiggyBac homologs share a common ciliate ancestor with the oligohymenophorean PiggyBacs, or whether they arose from independent acquisitions in major ciliate groups, we created a large phylogeny of PiggyBac homologs representative of putative domesticated transposases from *Blepharisma stoltei* ATCC 30299, *Condylostoma magnum*, *Paramecium* spp., *Tetrahymena thermophila*, as well as PiggyBac-like elements (PBLEs (Bouallègue et al., 2017)) from diverse eukaryotes (Figure 5.6). All the heterotrichous ciliates PiggyBac homologs, ie. BPgm, BPgmLs 1-7 and PiggyMics grouped together with the *Condylostoma* Pgms. The ciliate Pgms and PgmLs largely cluster as a single clade, with the exception of PiggyMic 5, which appears as a low-support outgroup to

opisthokont, archaeplastid and stramenopile PiggyBac-like elements. PiggyMic 5 has the shortest detected DDE_Tnp_1_7 domain (26 a.a.) and appeared poorly aligned relative to the other homologs.

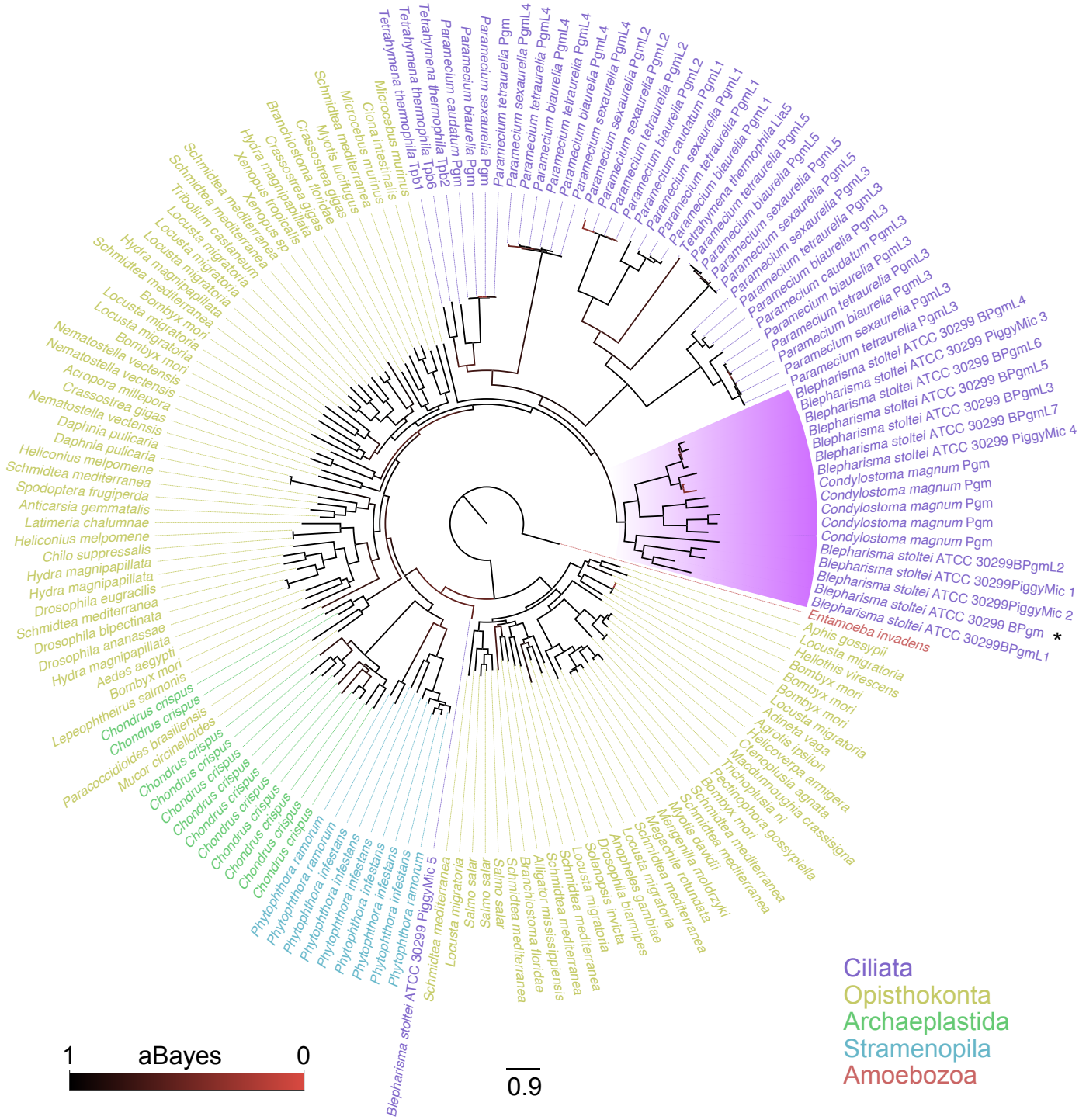


Figure 5.6. Phylogeny of ciliate PiggyBac homologs and eukaryotic PBLEs. Highlighted clade contains all PiggyBac homologs found in Heterotrichea, containing MAC and MIC-limited homologs of PiggyMac from *Blepharisma* and PiggyMac homologs of *Condylostoma magnum.* The tree is rooted at the PiggyBac-like element of *Entamoeba invadens.*

5.3. Discussion

5.3.1. The *Blepharisma* MAC genome is organized as minichromosomes

Alternative telomere addition sites in the MAC genome tend to be intergenic in model ciliates like *Oxytricha* trifallax (Swart et al., 2013). In *Blepharisma*, we found more intergenic ATASs (28309) than intragenic ones (18396). As intergenic regions only make up 10.1 Mb of the assembly, the intergenic frequency of ATASs is about five-fold higher (2.81 per 1 kb) than intragenic frequency (0.562 per 1 kb). The presence of intragenic ATASs raises the question how the cell tolerates or deals with mRNAs encoding partial proteins transcribed from 3' truncated genes. Since the sequence data was from a clonal population, it is not possible to tell how much ATAS variability there is within individual cells. However, it is conceivable that their positional variation in single cells reflects that of the population. In this case, together with redundancy from massive DNA amplification there would likely be sufficient intact copies of every gene.

Blepharisma, like the heterotirch *Stentor*, has predominantly 15 nt introns, making them the shortest spliceosomal introns in eukaryotes. *Blepharisma* appears to lack a minor spliceosome and minor spliceosomal introns. As far as we are aware no minor spliceosomal introns have been reported in any ciliates. Loss of minor spliceosomal machinery and introns, relative to the eukaryotic common ancestor, may be relatively common in alveolates including ciliates (Russell et al. 2006).

5.3.2. A PiggyBac is the main IES excisase in *Blepharisma*

In *Paramecium tetraurelia* and *Tetrahymena thermophila*, PiggyBac transposases are responsible for IES excision during genome editing (Baudry et al., 2009; Cheng et al., 2010). These transposases appear to have been domesticated, i.e., their genes are no longer contained in transposons but are encoded in the somatic genome where they play an essential genome development role (Baudry et al., 2009; Cheng et al., 2010). PiggyBac homologs typically have a DDD catalytic triad (Yuan and Wessler, 2011), which is preserved in *Paramecium* PiggyMac (Pgm) and *Tetrahymena* PiggyBac homologs Tpb1 and Tpb2 (Bischerour et al., 2018; Cheng et al., 2010). Among ciliates, domesticated PiggyBac transposases have so far only been reported in these model oligohymenophorean genera. Notably they have not been detected in either the MAC or MIC genome of the spirotrich *Oxytricha trifallax* (Chen et al., 2014; Swart et al., 2013).

The responsible IES excisases in the less-studied spirotrichs, *Oxytricha*, *Stylonychia* and *Euplotes*, are not as evident. *Oxytricha*'s TBE transposases are considered to be involved in IES excision, but are encoded by full-length germline-limited transposons and are absent from the MAC (Nowacki et al., 2009), unlike the primary, MAC genome-encoded IES excisase Tpb2 in *Tetrahymena,* and the *Paramecium* PiggyMacs and PiggyMac-likes. The pronounced developmental upregulation of numerous additional MAC- and MIC-encoded transposases in *Oxytricha* raises the possibility that transposases other than those of TBEs could also be involved in IES excision (Chen et al., 2014; Swart et al., 2013). Knowledge of IESs in other ciliates is sparse, primarily confined to the phyllopharyngean *Chilodonella uncinata* (Zufall and Katz, 2007; Zufall et al., 2012). As far as we are aware, no specific IES excisases have been proposed for them.

Since the oligohymenophorean PiggyBac homologs are clear IES excisases, we sought and found eight homologs of these genes in the *Blepharisma* MAC genome and five in the IESs. *Blepharisma* is the first ciliate genus aside from *Tetrahymena* and *Paramecium* in which such proteins have been reported. The *Blepharisma* PiggyBac homologs are distantly related to the PiggyBac homologs of both *Tetrahymena* and *Paramecium*. Additional searches revealed clear PiggyBac homologs in *Condylostoma magnum*, and a weaker pair of matches in *Stentor coeruleus*, suggesting that these are a common feature of heterotrich ciliates. A single *Blepharisma* PiggyBac homolog has a complete canonical DDD catalytic triad, and its gene is highly upregulated during MAC development. This is reminiscent of *Paramecium tetraurelia*, in which just one of the nine PiggyBac homologs, PiggyMac, has a complete DDD catalytic triad (Bischerour et al., 2018). As is characteristic of PiggyBac homologs, each of the three PiggyBac homologs in *Blepharisma*, *Paramecium* and *Tetrahymena* also has a C-terminal, cysteine-rich, zinc finger domain. The organization of the heterotrich PiggyBac homolog zinc finger domains is more similar to comparable domains of *Homo sapiens* PGBD2 and PGBD3 homologs than the zinc finger domains in *Paramecium* and *Tetrahymena* PiggyBac homologs.

In *Paramecium* aside from PiggyMac, all PiggyMac-likes have incomplete catalytic triads, and thus are likely catalytically inactive, but nevertheless their gene knockdowns lead to pronounced IES retention (Bischerour et al., 2018). It has therefore been proposed that the PiggyMac-likes may function as heteromeric multi-subunit complexes in conjunction with PiggyMac during DNA excision (Bischerour et al., 2018). On the other hand, cryo-EM structures available for moth PiggyBac transposase support a model in which these proteins

function as a homodimeric complex *in vitro* (Chen et al., 2020). Furthermore, the primary *Tetrahymena* PiggyBac, Tpb2, is able to perform cleavage *in vitro* alone (Cheng et al., 2010). In other eukaryotes, domesticated PiggyBacs without complete catalytic triads are thought to be retained due to co-option of their DNA-binding domains by the host genomes (Sarkar et al., 2003). One possibility for such purely DNA-binding transposase-derived proteins in ciliates could be in regulating the excision of DNA by the catalytically active transposases. Future experimental analyses of the BPgm and the BPgm-likes could aid in resolving the conundrums and understanding of possible interactions between catalytically active and inactive transposases.

Supplementary figures

Figure S5.1. *B. stoltei* ATCC30299 MAC genome orthogroups and assembly graph. Bandage (Wick et al., 2015) representation of Flye 2.8.1 assembly graph. Edges corresponding to contigs are colored by coverage (brightest pink = 160×, black=0×).

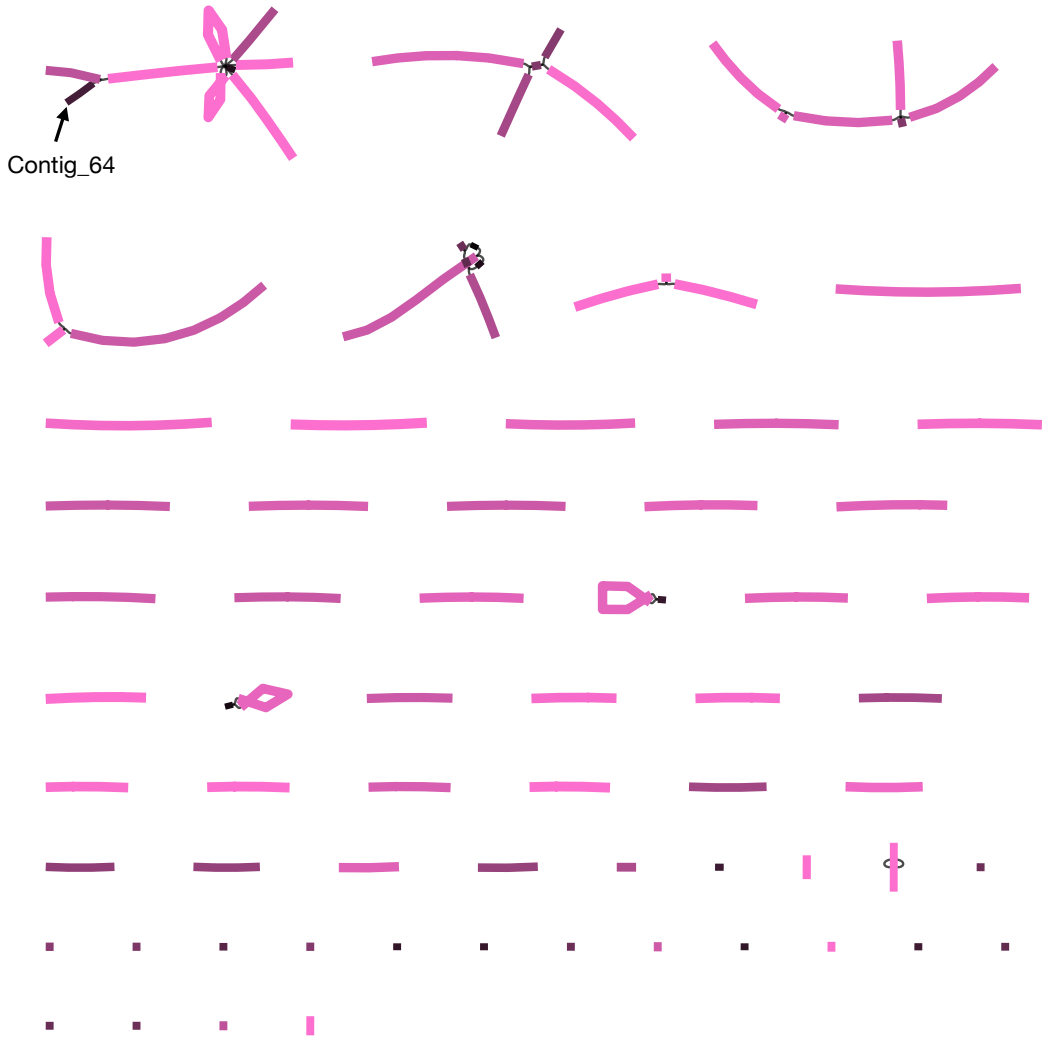
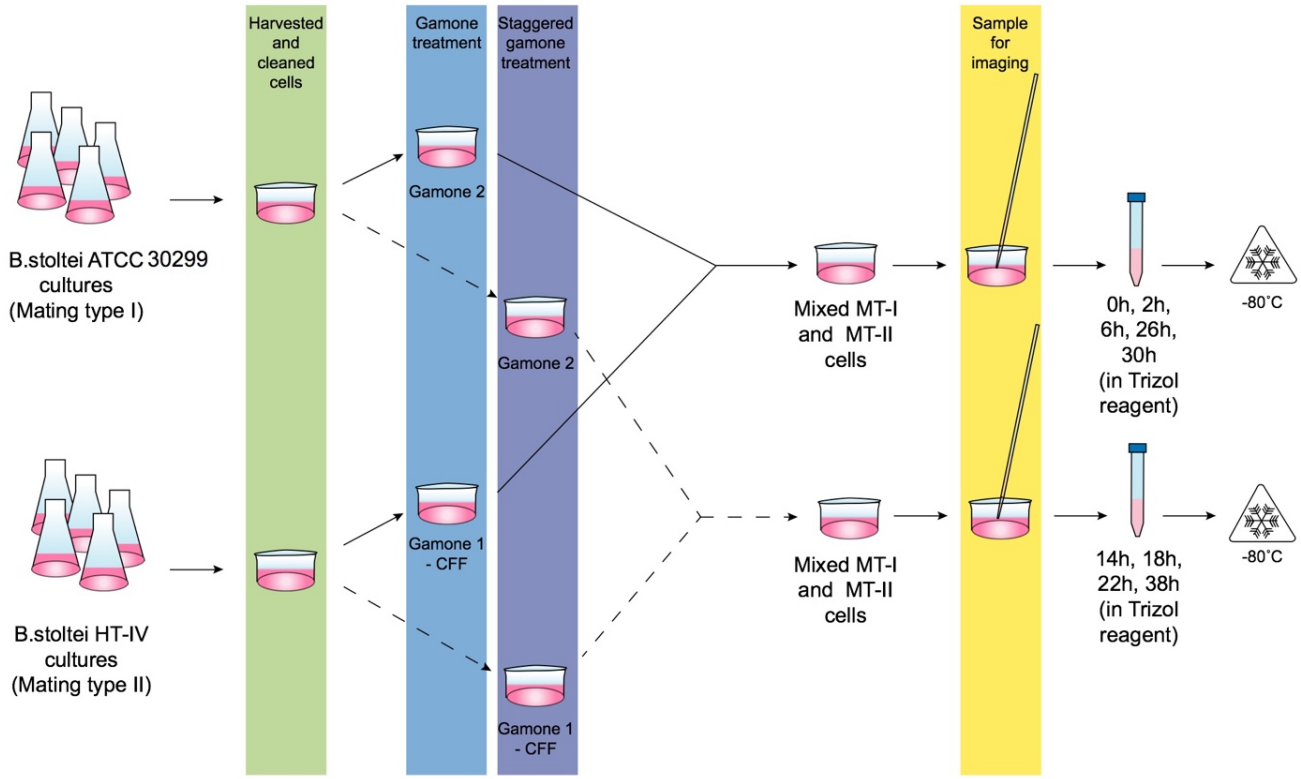


Figure S5.2. Experimental approach for conjugation RNA-seq time series. Complementary mating type strains of *Blepharisma stoltei* were harvested and cleaned by starving overnight. The cleaned cultures were treated in a time-staggered format, with gamones of the complementary mating type, where gamone 2 was a solution of the synthetic gamone 2 calcium salt and gamone 1 was provided as the cell-free fluid (CFF) harvested from mating-type I cells. Two sets of time-staggered gamone-treated cultures were used for the time series. Set I, indicated by the solid line, was mixed and used to observe and collect samples at 0 hours, 2 hours, 6 hours, 26 hours and 30 hours after mixing. Set II, indicated by the dashed lines, was mixed and used to observe and collect samples at 14 hours, 18 hours, 22 hours and 38 hours after mixing. Test tubes indicate Trizol samples prepared for RNA-extraction which were stored at -80 ˚C before processing. Cells collected for imaging were obtained shortly before the remainder were transferred into Trizol.



Supplementary tables

Table S5.1. Genome properties of the long-read assemblies for *B. stoltei*.

Assembly	Flye (v2.7) Replicate 1	Flye (v2.7) Replicate 2	Flye (v2.7) Combined	Flye (v2.8) Combined	Final assembly
Contigs	89	86	74	72	64 (excluding mitogenome)
Mean coverage (from flye.log)	76	70	145	145	NA
%GC	33.3	33.3	33.4	32.9	33.6
Longest contig (bp)	2036921	1188116	1541963	1608201	1514878
Assembly size (bp)	42701284	43066385	43062848	42982242	41464486
N50	738771	757357	799426	817639	795340
Two telomeres	38	37	36	16	64
One telomere	36	36	25	32	0
Zero telomeres	15	13	13	24	0

Table S5.2. Genome properties of model and non-model ciliates.

Species	Genome size (Mb)	Genome architecture	Genes (zygosity)	Codon reassignments
Blepharisma stoltei	41	Minichromosomes	25726(n)	UGA -> W
Stentor coeruleus	77[2]	?	31426[2](n)	Standard genetic code[2]
Paramecium tetraurelia	72[3]	Chromosomes[3], [4]	39642[3](n)	UAA, UAG -> Q[1]
Tetrahymena thermophila	103[5]	Chromosomes[6]	26258[5](n)	UAA, UAG -> Q[1]
Euplotes octocarinatus	88[7]	Nanochromosomes[8]	29076[7](n)	UGA -> C[9]
Stylonychia lemnae	52[10]	Nanochromosomes[10]	15102(n)[10]	UAA, UAG -> Q[1]
Oxytricha trifallax	50[11]	Nanochromosomes[11]	18400 (n)[11]	UAA, UAG -> Q[1]
Perkinsus olseni	63[12]	Chromosomes[12]	17342(4n)[12]	Standard genetic code[12]

References for Table S5.2

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

2. Slabodnick, M. M. *et al.* The Macronuclear Genome of *Stentor coeruleus* Reveals Tiny Introns in a Giant Cell. *Curr. Biol.* 27, 569–575 (2017).

3. Aury, J.-M. *et al.* Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444, 171–178 (2006).

4. Duret, L. *et al.* Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: A somatic view of the germline. *Genome Res.* 18, 585–596 (2008).

5. Sheng, Y. *et al.* The completed macronuclear genome of a model ciliate *Tetrahymena thermophila* and its application in genome scrambling and copy number analyses. *Sci. China Life Sci.* 63, 1534–1542 (2020).

6. Eisen, J. A. *et al.* Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* 4, 1620–1642 (2006).

7. Wang, R. lin, Miao, W., Wang, W., Xiong, J. & Liang, A. hua. EOGD: The Euplotes octocarinatus genome database. *BMC Genomics* 19, 1–6 (2018).

8. Ghosh, S., Jaraczewski, J. W., Klobutcher, L. A. & Jahn, C. L. Characterization of transcription initiation, translation initiation, and poly(A) addition sites in the gene-sized macronuclear DNA molecules of Euplotes. *Nucleic Acids Res.* 22, 214–221 (1994).

9. Meyer, F. *et al.* UGA is translated as cysteine in pheromone 3 of Euplotes octocarinatus. *Proc. Natl. Acad. Sci. U. S. A.* 88, 3758–3761 (1991).

10. Aeschlimann, S. H. *et al.* The Draft Assembly of the Radically Organized *Stylonychia lemnae* Macronuclear Genome. *Genome Biol. Evol.* 6, 1707–1723 (2014).

11. Swart, E. C. *et al.* The *Oxytricha trifallax* Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes. *PLoS Biol.* 11, e1001473 (2013).

12. Bogema, D. R. *et al.* Draft genomes of Perkinsus olseni and Perkinsus chesapeaki reveal polyploidy and regional differences in heterozygosity. *Genomics* 113, 677–688 (2021).

Table S5.3. Substitution rates between *Paramecium tetraurelia* PiggyMac-like genes and PiggyMac.

Gene abbreviation	*P. tetraurelia* gene ID	d_N/d_S	d_N	d_S
PGML2	PTET.51.1.G0380073	0.0773	1.1082	14.3409
PGML3a	PTET.51.1.G0010374	0.1245	1.0335	8.3021
PGML3b	PTET.51.1.G0080308	0.0404	1.1559	28.6183
PGML3c	PTET.51.1.G0020217	0.1508	1.0885	7.216
PGML4a	PTET.51.1.G0340197	0.1161	0.9593	8.2612
PGML4b	PTET.51.1.G0480099	0.2535	1.1062	4.3641
PGML5a	PTET.51.1.G0570051	0.0141	1.1514	81.7442
PGML5b	PTET.51.1.G0510172	0.0138	1.1642	84.3893

Reference gene: PGM - PTET.51.1.G0490162

Table S5.4. Substitution rates between *Paramecium tetraurelia* and *Paramecium octaurelia* PiggyMac and PiggyMac-likes

Gene abbreviation	*P. tetraurelia* gene ID	*P. octaurelia* gene ID	d_N/d_S	d_N	d_S
PGM	PTET.51.1.G0490162	POCT.K8.1.G71800002770580243	0.0234	0.0073	0.3106
PGML2	PTET.51.1.G0380073	POCT.K8.1.G71800002770130227	0.0180	0.0045	0.2507
PGML3a	PTET.51.1.G0010374	POCT.K8.1.G71800002770510320	0.0229	0.0082	0.3600
PGML3b	PTET.51.1.G0080308	POCT.K8.1.G71800002770810134	0.0818	0.0245	0.2993
PGML3c	PTET.51.1.G0020217	POCT.K8.1.G71800002770610330	0.1052	0.0365	0.3469
PGML4a	PTET.51.1.G0340197 –	POCT.K8.1.G71800002770180100	0.0425	0.0139	0.3262
PGML4b	PTET.51.1.G0480099	POCT.K8.1.G71800002770140101	0.0627	0.0153	0.2445
PGML5a	PTET.51.1.G0570051	POCT.K8.1.G71800002770010048	0.0393	0.0110	0.2800
PGML5b	PTET.51.1.G0510172	POCT.K8.1.G71800002769800173	0.0596	0.0123	0.2071

Observed d_N/d_S values for orthologous pairs of PiggyMac and PiggyMac-like proteins from *P. tetraurelia* and *P. octaurelia*.(PGML1, 95.1% nucleotide identity, PGML3c, 89.7% nucleotide identity)

5.4. Bibliography

Aeschlimann, Samuel H, Franziska Jönsson, Jan Postberg, Nicholas A Stover, Robert L Petera, Hans-Joachim Lipps, Mariusz Nowacki, and Estienne C Swart. 2014. "The Draft Assembly of the Radically Organized *Stylonychia Lemnae* Macronuclear Genome." *Genome Biology and Evolution* 6 (7): 1707–23. https://doi.org/10.1093/gbe/evu139.

Arnaiz, Olivier, Nathalie Mathy, Céline Baudry, Sophie Malinsky, Jean-Marc Aury, Cyril Denby Wilkes, Olivier Garnier, et al. 2012. "The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences." *PLoS Genetics* 8 (10): e1002984. https://doi.org/10.1371/journal.pgen.1002984.

Aury, Jean-Marc, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M Porcel, Béatrice Ségurens, et al. 2006. "Global Trends of Whole-Genome Duplications Revealed by the Ciliate *Paramecium Tetraurelia*." *Nature* 444 (7116): 171–78. https://doi.org/10.1038/nature05230.

Baudry, Céline, Sophie Malinsky, Matthieu Restituito, Aurélie Kapusta, Sarah Rosa, Eric Meyer, and Mireille Bétermier. 2009. "PiggyMac, a Domesticated PiggyBac Transposase Involved in Programmed Genome Rearrangements in the Ciliate Paramecium Tetraurelia." *Genes & Development* 23 (21): 2478–83. https://doi.org/10.1101/gad.547309.

Bischerour, Julien, Simran Bhullar, Cyril Denby Wilkes, Vinciane Régnier, Nathalie Mathy, Emeline Dubois, Aditi Singh, et al. 2018. "Six Domesticated PiggyBac Transposases Together Carry out Programmed DNA Elimination in Paramecium." *ELife* 7 (September). https://doi.org/10.7554/eLife.37927.

Bondarenko, Vladyslav S, and Mikhail S Gelfand. 2016. "Evolution of the Exon-Intron Structure in Ciliate Genomes." *Plos One* 11 (9): e0161476. https://doi.org/10.1371/journal.pone.0161476.

Bouallègue, Maryem, Jacques-Deric Rouault, Aurélie Hua-Van, Mohamed Makni, and Pierre Capy. 2017. "Molecular Evolution of PiggyBac Superfamily: From Selfishness to Domestication." *Genome Biology and Evolution* 9 (2): 323–39. https://doi.org/10.1093/gbe/evw292.

Cavalcanti, Andre R O, Diane M Dunn, Robert Weiss, Glenn Herrick, Laura F Landweber, and Thomas G Doak. 2004. "Sequence Features of Oxytricha Trifallax (Class Spirotrichea) Macronuclear Telomeric and Subtelomeric Sequences." *Protist* 155 (3): 311–22. https://doi.org/10.1078/1434461041844196.

Chalker, Douglas L, Eric Meyer, and Kazufumi Mochizuki. 2013. "Epigenetics of Ciliates." *Cold Spring Harbor Perspectives in Biology* 5 (12): a017764. https://doi.org/10.1101/cshperspect.a017764.

Cheng, Chao-Yin, Alexander Vogt, Kazufumi Mochizuki, and Meng-Chao Yao. 2010. "A Domesticated PiggyBac Transposase Plays Key Roles in Heterochromatin Dynamics and

DNA Cleavage during Programmed DNA Deletion in Tetrahymena Thermophila." *Molecular Biology of the Cell* 21 (10): 1753–62. https://doi.org/10.1091/mbc.e09-12-1079.

Cheng, Yu-Hsuan, Chien-Fu Jeff Liu, Yen-Hsin Yu, Yu-Ting Jhou, Masahiro Fujishima, Isheng Jason Tsai, and Jun-Yi Leu. 2020. "Genome Plasticity in Paramecium Bursaria Revealed by Population Genomics." *BMC Biology* 18 (1): 180. https://doi.org/10.1186/s12915-020-00912-2.

Chen, Qiujia, Wentian Luo, Ruth Ann Veach, Alison B Hickman, Matthew H Wilson, and Fred Dyda. 2020. "Structural Basis of Seamless Excision and Specific Targeting by PiggyBac Transposase." *Nature Communications* 11 (1): 3446. https://doi.org/10.1038/s41467-020-17128-1.

Chen, Xiao, John R Bracht, Aaron David Goldman, Egor Dolzhenko, Derek M Clay, Estienne C Swart, David H Perlman, et al. 2014. "The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development." *Cell* 158 (5): 1187–98. https://doi.org/10.1016/j.cell.2014.07.034.

Chin, Chen-Shan, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. "Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing." *Nature Methods* 13 (12): 1050–54. https://doi.org/10.1038/nmeth.4035.

Csuros, Miklos, Igor B Rogozin, and Eugene V Koonin. 2011. "A Detailed History of Intron-Rich Eukaryotic Ancestors Inferred from a Global Survey of 100 Complete Genomes." *PLoS Computational Biology* 7 (9): e1002150. https://doi.org/10.1371/journal.pcbi.1002150.

Guérineau, Marc, Luiza Bessa, Séverine Moriau, Ewen Lescop, François Bontems, Nathalie Mathy, Eric Guittet, Julien Bischerour, Mireille Bétermier, and Nelly Morellet. 2021. "The Unusual Structure of the PiggyMac Cysteine-Rich Domain Reveals Zinc Finger Diversity in PiggyBac-Related Transposases." *Mobile DNA* 12 (1): 12. https://doi.org/10.1186/s13100-021-00240-4.

Hamilton, Eileen P, Aurélie Kapusta, Piroska E Huvos, Shelby L Bidwell, Nikhat Zafar, Haibao Tang, Michalis Hadjithomas, et al. 2016. "Structure of the Germline Genome of *Tetrahymena Thermophila* and Relationship to the Massively Rearranged Somatic Genome." *ELife* 5 (November). https://doi.org/10.7554/eLife.19090.

Klobutcher, L A, and G Herrick. 1995. "Consensus Inverted Terminal Repeat Sequence of *Paramecium* IESs: Resemblance to Termini of Tc1-Related and *Euplotes* Tec Transposons." *Nucleic Acids Research* 23 (11): 2006–13. https://doi.org/10.1093/nar/23.11.2006.

———. 1997. "Developmental Genome Reorganization in Ciliated Protozoa: The Transposon Link." *Progress in Nucleic Acid Research and Molecular Biology* 56: 1–62. https://doi.org/10.1016/S0079-6603(08)61001-6.

Lynn, Denis H. 2010. *The Ciliated Protozoa*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-8239-9.

Morin, G B. 1991. "Recognition of a Chromosome Truncation Site Associated with Alpha-Thalassaemia by Human Telomerase." *Nature* 353 (6343): 454–56. https://doi.org/10.1038/353454a0.

Nawrocki, Eric P, Diana L Kolbe, and Sean R Eddy. 2009. "Infernal 1.0: Inference of RNA Alignments." *Bioinformatics* 25 (10): 1335–37. https://doi.org/10.1093/bioinformatics/btp157.

Nowacki, Mariusz, Brian P Higgins, Genevieve M Maquilan, Estienne C Swart, Thomas G Doak, and Laura F Landweber. 2009. "A Functional Role for Transposases in a Large Eukaryotic Genome." *Science* 324 (5929): 935–38. https://doi.org/10.1126/science.1170023.

Prescott, D M. 1994. "The DNA of Ciliated Protozoa." *Microbiological Reviews* 58 (2): 233–67. https://doi.org/10.1128/mr.58.2.233-267.1994.

Putnam, Christopher D, Vincent Pennaneach, and Richard D Kolodner. 2004. "Chromosome Healing through Terminal Deletions Generated by de Novo Telomere Additions in Saccharomyces Cerevisiae." *Proceedings of the National Academy of Sciences of the United States of America* 101 (36): 13262–67. https://doi.org/10.1073/pnas.0405443101.

Roy, Scott W, Andrew J Hudson, Joella Joseph, Janet Yee, and Anthony G Russell. 2012. "Numerous Fragmented Spliceosomal Introns, AT-AC Splicing, and an Unusual Dynein Gene Expression Pathway in Giardia Lamblia." *Molecular Biology and Evolution* 29 (1): 43–49. https://doi.org/10.1093/molbev/msr063.

Russell, Anthony G, J Michael Charette, David F Spencer, and Michael W Gray. 2006. "An Early Evolutionary Origin for the Minor Spliceosome." *Nature* 443 (7113): 863–66. https://doi.org/10.1038/nature05228.

Sarkar, A, C Sim, Y S Hong, J R Hogan, M J Fraser, H M Robertson, and F H Collins. 2003. "Molecular Evolutionary Analysis of the Widespread PiggyBac Transposon Family and Related 'Domesticated' Sequences." *Molecular Genetics and Genomics* 270 (2): 173–80. https://doi.org/10.1007/s00438-003-0909-0.

Sheng, Yalan, Lili Duan, Ting Cheng, Yu Qiao, Naomi A Stover, and Shan Gao. 2020. "The Completed Macronuclear Genome of a Model Ciliate *Tetrahymena Thermophila* and Its Application in Genome Scrambling and Copy Number Analyses." *Science China. Life Sciences* 63 (10): 1534–42. https://doi.org/10.1007/s11427-020-1689-4.

Slabodnick, Mark M, J Graham Ruby, Sarah B Reiff, Estienne C Swart, Sager Gosai, Sudhakaran Prabakaran, Ewa Witkowska, et al. 2017. "The Macronuclear Genome of *Stentor Coeruleus* Reveals Tiny Introns in a Giant Cell." *Current Biology* 27 (4): 569–75. https://doi.org/10.1016/j.cub.2016.12.057.

Swart, Estienne C, John R Bracht, Vincent Magrini, Patrick Minx, Xiao Chen, Yi Zhou, Jaspreet S Khurana, et al. 2013. "The *Oxytricha Trifallax* Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes." *PLoS Biology* 11 (1): e1001473. https://doi.org/10.1371/journal.pbio.1001473.

Swart, Estienne C, and Mariusz Nowacki. 2015. "The Eukaryotic Way to Defend and Edit Genomes by SRNA-Targeted DNA Deletion." *Annals of the New York Academy of Sciences* 1341 (April): 106–14. https://doi.org/10.1111/nyas.12636.

Vogt, Alexander, Aaron David Goldman, Kazufumi Mochizuki, and Laura F Landweber. 2013. "Transposon Domestication versus Mutualism in Ciliate Genome Rearrangements." *PLoS Genetics* 9 (8): e1003659. https://doi.org/10.1371/journal.pgen.1003659.

Wang, H, and E H Blackburn. 1997. "De Novo Telomere Addition by Tetrahymena Telomerase in Vitro." *The EMBO Journal* 16 (4): 866–79. https://doi.org/10.1093/emboj/16.4.866.

Yang, Z, and R Nielsen. 2000. "Estimating Synonymous and Nonsynonymous Substitution Rates under Realistic Evolutionary Models." *Molecular Biology and Evolution* 17 (1): 32–43. https://doi.org/10.1093/oxfordjournals.molbev.a026236.

Yao, M C, C H Yao, and B Monks. 1990. "The Controlling Sequence for Site-Specific Chromosome Breakage in Tetrahymena." *Cell* 63 (4): 763–72. https://doi.org/10.1016/0092-8674(90)90142-2.

Yuan, Yao-Wu, and Susan R Wessler. 2011. "The Catalytic Domain of All Eukaryotic Cut-and-Paste Transposase Superfamilies." *Proceedings of the National Academy of Sciences of the United States of America* 108 (19): 7884–89. https://doi.org/10.1073/pnas.1104208108.

Zagulski, Marek, Jacek K Nowak, Anne Le Mouël, Mariusz Nowacki, Andrzej Migdalski, Robert Gromadka, Benjamin Noël, et al. 2004. "High Coding Density on the Largest Paramecium Tetraurelia Somatic Chromosome." *Current Biology* 14 (15): 1397–1404. https://doi.org/10.1016/j.cub.2004.07.029.

Zufall, Rebecca A, and Laura A Katz. 2007. "Micronuclear and Macronuclear Forms of Beta-Tubulin Genes in the Ciliate Chilodonella Uncinata Reveal Insights into Genome Processing and Protein Evolution." *The Journal of Eukaryotic Microbiology* 54 (3): 275–82. https://doi.org/10.1111/j.1550-7408.2007.00267.x.

Zufall, Rebecca A, Mariel Sturm, and Brian C Mahon. 2012. "Evolution of Germline-Limited Sequences in Two Populations of the Ciliate Chilodonella Uncinata." *Journal of Molecular Evolution* 74 (3–4): 140–46. https://doi.org/10.1007/s00239-012-9493-4.

Chapter 6

MITE infestation of germline accommodated by genome editing in *Blepharisma*

Brandon Kwee Boon Seah[1], Minakshi Singh[1], Christiane Emmerich[1], Aditi Singh[1], Christian Woehle[2], Bruno Huettel[2], Adam Byerly[3], Naomi Stover[4], Mayumi Sugiura[5], Terue Harumoto[5], Estienne Carl Swart[1]

[1] Max Planck Institute for Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany

[2] Max Planck Genome Center Cologne, Building B, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

[3] Department of Computer Science and Information Systems, Bradley University, Peoria IL, USA

[4] Department of Biology, Bradley University, Peoria IL, USA

[5] Department of Chemistry, Biology, and Environmental Sciences, Faculty of Science, Nara Women's University, Nara 630-8506, Japan

This chapter has been adapted from a BioRxiv preprint DOI:
https://doi.org/10.1101/2022.05.02.489906.

I generated all the data, performed gene annotation of the germline-limited regions of the germline genome, performed functional annotation of the germline-limited regions of the germline genome and analyzed the gene expression by integrating transcriptomic data, also generated by me. I contributed to the phylogenetic analysis depicted in Figure 6.4A and protein domain analysis shown in Figure 6.4E. I performed the gene expression analysis and analysis of the protein domains of the somatic genome-encoded and germline-limited transposases shown in Figure S6.4B. Additional analysis was performed by Dr.'s Estienne Swart and Kwee Boon Brandon Seah. Dr.'s Estienne Swart and Kwee Boon Brandon Seah reviewed and edited the original manuscript. Details of all author contributions are listed in Appendix A.1.

6.1. Introduction

Ciliates are microbial eukaryotes that maintain separate germline and somatic genomes in each cell, housed in two types of nuclei. During the sexual life cycle, germline micronuclei (MICs) develop into new somatic macronuclei (MACs) via a process of small RNA (sRNA)-assisted DNA elimination and DNA amplification. Most gene expression takes place from the resulting somatic MAC.

Germline-limited genome segments, called internally eliminated sequences (IESs), are excised during development from MIC to MAC. The MAC genome content is hence a subset of the germline MIC. Each taxon studied thus far has its own peculiarities. For example, *Paramecium* IESs are typically short, have unique sequence content, and are precisely excised, vs. those in *Tetrahymena* that are longer, more repetitive, and imprecisely excised (Hamilton et al. 2016; Arnaiz et al. 2012; L. Feng et al. 2017).

Ciliate IESs are thought to originate from cut-and-paste DNA transposons (Klobutcher and Herrick 1997) (Figure 6.1B), because: (i) 5'-TA-3' motifs at IES boundaries (*Euplotes*, *Paramecium*) resemble the terminal direct repeats of Tc1/Mariner-superfamily transposons (Klobutcher and Herrick 1995); (ii) transposon-derived "domesticated" excisases are used to remove IESs (Baudry et al. 2009; Cheng et al. 2010; Nowacki et al. 2009); and (iii) intact transposons encoding transposases are mostly germline-limited (Herrick et al. 1985; Jahn et al. 1993; Le Mouël et al. 2003; Arnaiz et al. 2012). Recently, non-autonomous mobile elements without transposases have been found in *Paramecium* spp. (Sellis et al. 2021). They resemble miniature inverted-repeat transposable elements (MITEs), but the autonomous counterparts of most of them, including the most abundant ones (thousands of copies), were not identified. MITEs, which are common in plants and animals, are deletion derivatives of Tc1/Mariner transposons, generally short (<500 bp), lacking coding sequences, and bounded by terminal repeats (Cedric Feschotte, Zhang, and Wessler 2002).

Developmental DNA elimination has been called "genome defense" because the process removes IESs, which not only derive from selfish genetic elements (transposons), but are often intragenic and hence deleterious if not removed (Yao, Fuller, and Xi 2003). The "defense" analogy was popularized due to parallels to other eukaryotes where small RNA-mediated DNA heterochromatinization is thought to suppress mobile element proliferation (Grewal and Jia 2007; Vogt and Mochizuki 2013; Coyne, Lhuillier-Akakpo, and Duharcourt 2012). Ciliates use

development-specific sRNAs to guide DNA elimination; in oligohymenophoreans, they mark sequences for elimination (Mochizuki et al. 2002; Yao, Fuller, and Xi 2003; Sandoval et al. 2014), whereas spirotrich sRNAs mark sequences to be retained (Zahler et al. 2012; Fang et al. 2012). Histone modifications are also required for elimination (Liu et al. 2007; Taverna, Coyne, and Allis 2002). However, this model has been questioned because sRNAs may not always be strictly necessary: in *Paramecium*, knockdown of key sRNA biogenesis enzymes had a smaller effect on shorter IESs, and were only weakly correlated with the more potent effects of knocking down the main IES excisase (Sandoval et al. 2014; Swart et al. 2014).

Other phenomena during genome editing can vary markedly between the few model ciliate species that have been studied in detail (reviews: (Coyne, Lhuillier-Akakpo, and Duharcourt 2012; Chalker, Meyer, and Mochizuki 2013; Rzeszutek, Maurer-Alcalá, and Nowacki 2020)). For example, germline chromosomes are fragmented into smaller somatic ones to some degree in all species, but spirotrichs produce extreme somatic "nanochromosomes" with only one or a few genes on average. "Unscrambling" of nonsequential MAC-destined sequences into the correct order in the somatic genome occurs frequently in some spirotrichs, e.g. *Oxytricha* and *Stylonychia* (Prescott and Greslin 1992), infrequently in *Tetrahymena* (Hamilton et al. 2016), and has not been reported in other ciliates (e.g. *Paramecium and Euplotes*). Furthermore, draft-quality germline genomes are available from only two out of eleven class-level taxa (following taxonomy of Lynn 2010): Oligohymenophorea (Hamilton et al. 2016; Sellis et al. 2021; Arnaiz et al. 2012; Guérin et al. 2017) and Spirotrichea (X. Chen et al. 2014).

Since it is not apparent which genome editing elements are common to all ciliates, we targeted the heterotrich *Blepharisma stoltei* (class Heterotrichea), whose last common ancestor with other ciliates with sequenced germline genomes is the last common ancestor of all ciliates (F. Gao and Katz 2014). *Blepharisma* has been a laboratory model for photobiology (Giese 1973) and mating factors (Kubota et al. 1973; Miyake and Beyer 1974; Miyake, Rivola, and Harumoto 1991; Sugiura and Harumoto 2001), so cultivated strains and protocols for inducing conjugation and development are available, and now too an accurate, highly contiguous draft somatic genome for *B. stoltei* (Singh et al. 2021). The somatic genome encodes a likely IES excisase, *Blepharisma* PiggyMac (BPgm), most closely related to the main IES excisases of *Paramecium* (PiggyMac) and *Tetrahymena* (Tpb2). Other somatic PiggyBac paralogs are also present but lack a complete "catalytic triad", similar to the situation in *Paramecium* (Bischerour et al. 2018). BPgm is

upregulated during new somatic MAC formation along with other development-specific genes, including homologs of sRNA biogenesis proteins implicated in genome editing (Singh et al. 2021).

In this study, we assembled a draft germline genome for *Blepharisma stoltei*. Through single molecule long read sequencing and a targeted assembly approach, we could fully assemble numerous IESs containing long, repetitive elements, which is not feasible with short read shotgun sequencing alone. We found numerous short (≤115 bp), precisely excised IESs with a periodic length distribution like in *Paramecium*. However, most IESs were longer (up to several kbp) and contained numerous repeat elements, including several "mobile IESs" where the IES corresponds to a complete repeat unit, and a Tc1/Mariner transposon whose non-autonomous MITE was also the most abundant repeat in the genome. We also identified small RNAs expressed during sexual development with characteristics of scnRNAs that guide DNA elimination in other ciliates. These results show common characteristics of germline-limited DNA in ciliates, but also illustrate how MITEs could be an intermediate stage in the origin and proliferation of IESs.

6.2. Results

6.2.1. Detection and targeted assembly of ca. forty thousand germline-limited IESs

We enriched germline micronuclei from *Blepharisma stoltei* strain ATCC 30299, and reconstructed 39799 IESs (13.2 Mbp, average coverage ~45x) scaffolded on the previously assembled 41 Mbp somatic genome (Chapter 5) (Singh et al. 2021), using a mapping and targeted assembly approach for PacBio long reads (Seah and Swart 2021). This MAC-scaffolded germline assembly is here referred to as the "MAC+IES" assembly. About 70% of all predicted IESs were intragenic (within coding sequences or introns), implying precise excision of IESs, as they would otherwise cause deleterious frameshifts. However, genes occupied 77% of the somatic assembly (excluding telomeres), so there was a small but statistically significant ($p = 3 \times 10^{-269}$) relative depletion of intragenic IESs.

6.2.2. A "hybrid" IES length distribution with periodic length peaks for short IESs

Most IESs were short (median 255 bp, mean 421 bp), but the distribution was long-tailed (90th percentile 603 bp, max 7251 bp). The length distribution was not unimodal, but had multiple peaks at specific length values (Figure 6.1A, Table S6.1), and could roughly be divided into two ranges: a "periodic" range from ~65 to 115 bp (11030 IESs), and a "non-periodic" range >115 bp (29646).

The "periodic" IES size range contained sharp peaks every 10 to 11 bp, similar to the periodicity of IESs in *Paramecium tetraurelia* (Arnaiz et al. 2012; Guérin et al. 2017). The first peak in *B. stoltei* was centered at 65 bp, compared to 28 bp in *P. tetraurelia*, and there was no "forbidden" peak. The most abundant "periodic" length peaks were at 72 bp and 110 bp. The "non-periodic" range (≥115 bp) contained isolated peaks at 153, 174, 228, and 389 bp, which has no obvious periodicity. Only 9701 IESs (total 1.36 Mbp) were contained within the size classes (width at half peak height) represented by the above peaks (both periodic and non-periodic) (Table S6.1), meaning that most IESs had lengths outside the peak values.

6.2.3. IESs are bounded by heterogeneous direct and inverted terminal repeats

In other ciliates, IES boundaries often have conserved terminal repeat motifs that could reflect excisase cut site preferences, or IES origins from specific classes of transposons. We therefore searched for both direct and inverted terminal repeats in *Blepharisma* IESs.

About three quarters of IESs (30212, 9.43 Mbp) were bounded by terminal direct repeats (TDRs) that contained the sequence TA ("TA-bound"). Other non-TA TDRs accounted for another 6566 (2.85 Mbp); the remainder were not TDR-bound, though some may be assembly errors (Figure 6.1A). *B. stoltei* genomes were AT-rich (somatic 33.5% GC, IESs 33.3% GC) like most ciliates, but the number of TA- and TDR-bound sequences was unlikely to be due to nucleotide composition alone (Figure 6.2A, 6.2B). The most common TDRs were simple alternations of T and A (TA, TAT/ATA, TATA), especially in IESs up to 228 bp (Figure 6.2C), with the exception of TAA/TTA (see below).

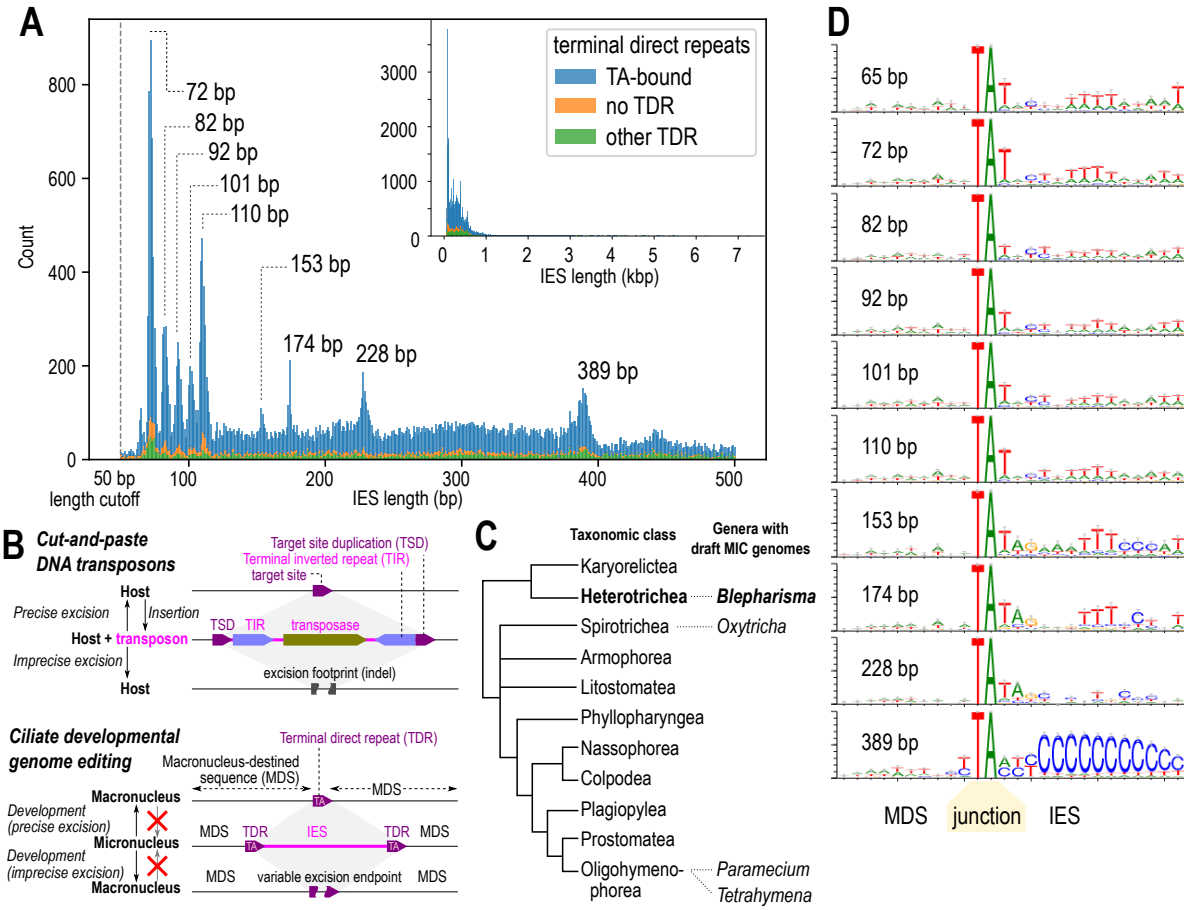
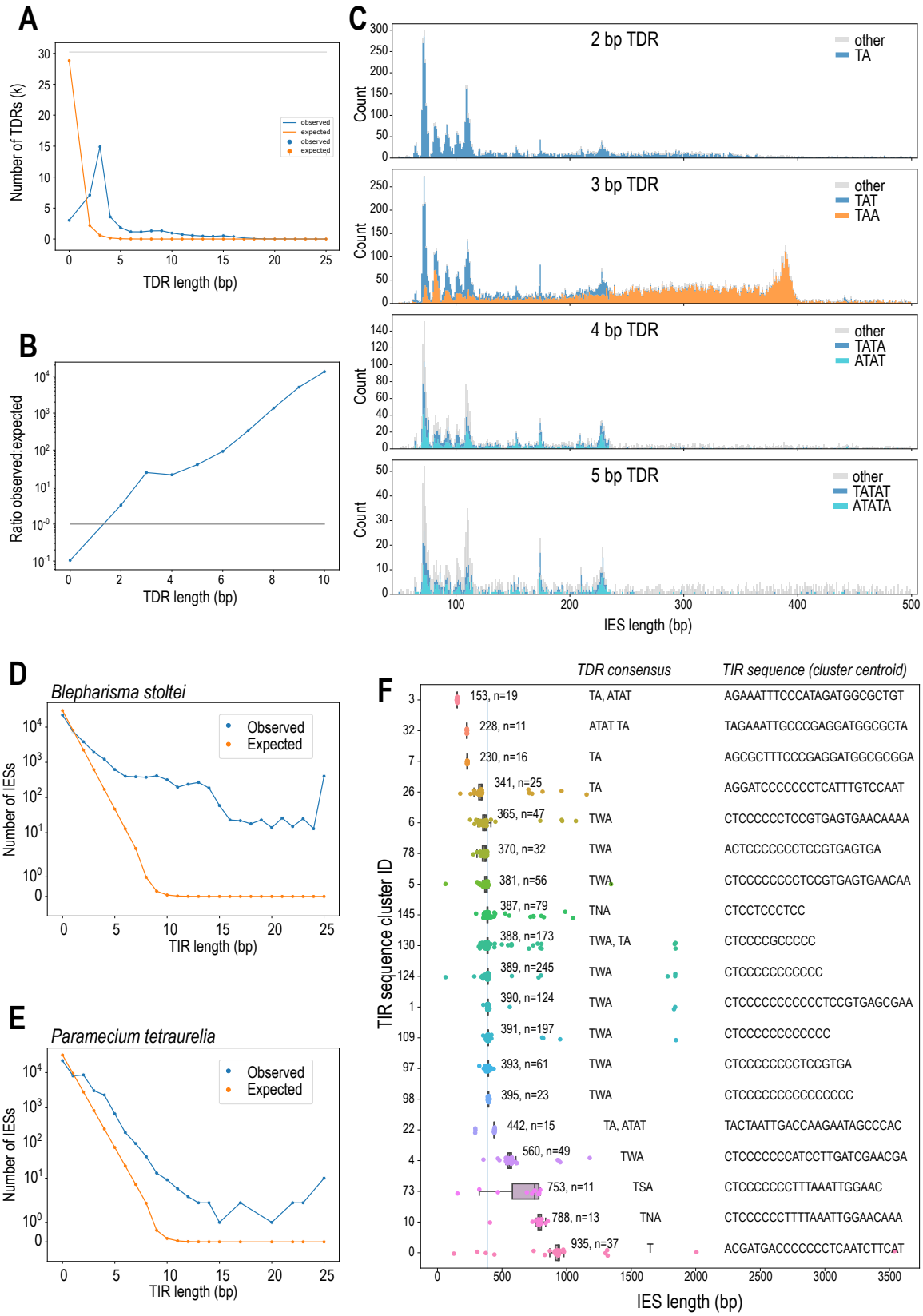


Figure 6.1. A "hybrid" IES length distribution with periodic length peaks for short IESs. (A) IES length histogram (0 to 500 bp (inset: full range), stacked bars for types of terminal direct repeats (TDRs) at IES boundaries: TA-bound (blue), no TDRs (orange), non-TA TDRs (green). Peaks for IES size classes discussed are marked; dashed vertical line indicates length cutoff (50 bp). (B) Comparison of cut-and-paste DNA transposons (above) and ciliate genome editing (below), showing parallels between target site duplications (TSD) of transposons and terminal direct repeats (TDRs) bounding IESs, and effects of precise vs. imprecise excision. MDS - macronuclear-destined sequence, TIR - terminal inverted repeat. (C) "Cryptic" IES length histogram, same color legend as panel A. (D) Sequence logos for MDS-IES junctions for TA-bound IESs of specific size classes, centered on the "TA" of the TDR.

Erroneous, low-frequency excision of MAC-destined sequences by the excision machinery ("cryptic" IESs) was also detected in MAC DNA libraries, with a slight peak at 72 bp (Figure 6.1C). Of 10048 cryptic IESs, 56% were TA-bound; TAA/TTA-bound IESs were also common, which suggests that the observed TDRs, including TAA/TTA, represented intrinsic cut site preferences of the domesticated excisase(s) (Figure 6.1C, Figure S6.1C to E).

Terminal inverted repeats (TIRs) at IES junctions were heterogeneous among IES size classes (Figure 6.1D, Figure 6.2F), and no single TIR motif was generally conserved across all *Blepharisma* IESs, unlike the 5'-TAYNR-3' motif of *Paramecium* IESs. Considering only TA-bound IESs, sequence logos of IES junctions for the "periodic" IESs had a weak consensus 5'-TAT rrn ttt t-3' (weakly conserved bases in lowercase). "Non-periodic" IESs had different signatures, e.g. ~153 and ~174 bp IESs had similar consensus 5'-TAT Agn nnT TT-3'. Despite their heterogeneity, TIRs were more common and longer than expected by chance, even with a strict criterion of no gaps or mismatches (Figure 6.2D to F). Sequence clustering of long (≥10 bp) TIRs showed distinct TIRs associated with specific IES lengths. Additionally, 376 palindromic IESs were identified, of

Figure 6.2. IESs are bounded by heterogeneous direct and inverted terminal repeats. (A) Numbers of terminal direct repeats (TDRs) per TDR length observed (blue) vs. number expected by random chance if bases were independently distributed (orange). (B) Ratio of observed to expected numbers of TDRs by length. (C) Length distributions of IESs containing TDRs of lengths 2, 3, 4, and 5 bp; the most abundant TDR sequences per TDR length are shown in color (sequences and their reverse complements are counted together, because TDRs could be encountered in either orientation, e.g. TAA/TTA), simple T/A alternations are in shades of blue. NB: plots in panel C have different vertical axis scales. (D) Observed IESs per terminal inverted repeat (TIR) length (blue line) vs. expected number by chance alone (orange). (E) Same as panel D but for *P. tetraurelia*. (F) Lengths (scatter-overlaid boxplot) of IESs containing long TIRs (≥10 bp), grouped by their TIR sequence (rows). Each TIR-cluster is annotated with the median IES length (bp), cluster size (n), TDR consensus sequence, and TIR representative sequence.


A
Number of TDRs (k)
TDR length (bp)
observed
expected
observed
expected
B
Ratio observed:expected
TDR length (bp)
C
2 bp TDR
other
TA
3 bp TDR
other
TAT
TAA
4 bp TDR
other
TATA
ATAT
5 bp TDR
other
TATAT
ATATA
Count
IES length (bp)
D
Blepharisma stoltei
Observed
Expected
Number of IESs
TIR length (bp)
E
Paramecium tetraurelia
Observed
Expected
Number of IESs
TIR length (bp)
F
TDR consensus
TIR sequence (cluster centroid)
153, n=19 TA, ATAT AGAAATTTCCCATAGATGGCGCTGT
228, n=11 ATAT TA TAGAAATTGCCCGAGGATGGCGCTA
230, n=16 TA AGCGCTTTCCCGAGGATGGCGCGGA
341, n=25 TA AGGATCCCCCCCTCATTTGTCCAAT
365, n=47 TWA CTCCCCCCTCCGTGAGTGAACAAAA
370, n=32 TWA ACTCCCCCCCTCCGTGAGTGA
381, n=56 TWA CTCCCCCCCCTCCGTGAGTGAACAA
387, n=79 TNA CTCCTCCCTCC
388, n=173 TWA, TA CTCCCCGCCCCC
389, n=245 TWA CTCCCCCCCCCCC
390, n=124 TWA CTCCCCCCCCCCCTCCGTGAGCGAA
391, n=197 TWA CTCCCCCCCCCCC
393, n=61 TWA CTCCCCCCCCTCCGTGA
395, n=23 TWA CTCCCCCCCCCCCCCCC
442, n=15 TA, ATAT TACTAATTGACCAAGAATAGCCCAC
560, n=49 TWA CTCCCCCCCATCCTTGATCGAACGA
753, n=11 TSA CTCCCCCCCTTTAAATTGGAAC
788, n=13 TNA CTCCCCCCTTTTAAATTGGAACAAA
935, n=37 T ACGATGACCCCCCCTCAATCTTCAT
TIR sequence cluster ID
IES length (bp)

which 153 (40.7%) fell within the same ~228 bp length peak, despite comprising several apparently unrelated palindrome sequences (Figure S6.2, Supplementary Information).

IESs in the ~389 bp sze peak had distinctive TDRs and TIRs, indicating that they are a family of mobile IESs, described further below (see "Pogo/Tigger-family transposon with abundant MITEs").

6.2.4. Repeat elements are abundant in long, non-periodic IESs

A quarter of the MAC+IES assembly (12.7 Mbp, 23.3%) was composed of interspersed repeats, however a much larger sequence fraction was repetitive in the germline-limited IESs (71.0%) than in the somatic genome (8.12%) (Figure 6.3A). The majority of sequence content contained in IESs ≥115 bp was annotated as repetitive, whereas the converse was true for shorter "periodic" IESs (Figure 6.3C), which parallels the pattern of short, unique IESs in *Paramecium* (Arnaiz et al. 2012).

Most interspersed repeats could not be classified to a known transposable element class by RepeatClassifier (Figure 6.3B, Table S6.2). The most abundant classifiable type was DNA/TcMar-Tc2, all of which actually belonged to a single repeat family rnd-1_family-1, followed by LINE/RTE-X. The most abundant repeat family, rnd-1_family-0, was unclassified and made up 21.2% (2.69 Mbp) of the total repetitive sequence. Repeat families rnd-1_family-0 and rnd-1_family-1 were related to each other and are discussed further below ("Pogo/Tigger-family transposon with abundant MITEs").

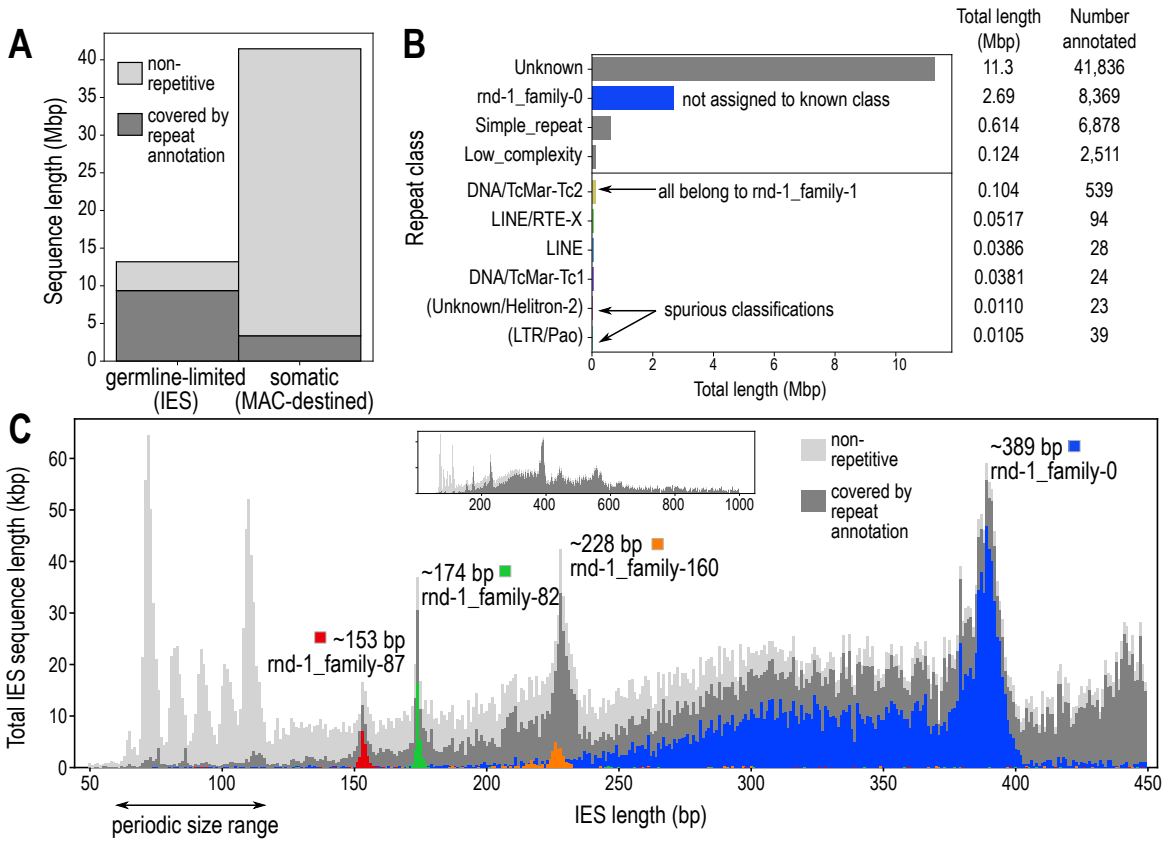


Figure 6.3. Repeat elements are abundant in long, non-periodic IESs. (A) Total sequence length annotated as interspersed repeats (dark) vs. non-repetitive (light), in germline vs. somatic parts of assembly. (B) Classification of repeat families by RepeatClassifier, and total annotated length per repeat class. (C) Total sequence length (vertical axis) per IES size class (horizontal axis), stacked plot of non-repetitive fraction (light) vs. interspersed repeats (dark), with the most abundant repeat families in the four non-periodic peaks overlaid in color. Inset: Distribution to 1000 bp.

Three non-periodic IES length peaks (153, 174, 389 bp) could be attributed to specific repeat families, suggesting that they proliferated recently (Table S6.3, Figure 6.3C, S6.3B). This was most pronounced for the ~389 bp peak, where 68.5% of the sequence content belonged to rnd-1_family-0, whereas about a quarter of the ~153 and ~174 bp peaks was composed of repeat families rnd-1_family-87 (palindromic) and rnd-1_family-82 respectively.

6.2.5. Germline-limited repeats include few autonomous transposons but many MITEs

Unlike *Tetrahymena* and *Oxytricha* where transposases are abundant in the germline-limited IESs but rare in the somatic genome, *Blepharisma* encoded only a few dozen identifiable transposase domains in either the germline-limited or somatic genomes. Cut-and-paste DNA transposase domains of the DDE/D superfamily identified in *Blepharisma* included DDE_1 and DDE_3 (Tc1/Mariner family), DDE_Tnp_1_7 (PiggyBac), DDE_Tnp_IS1595 (Merlin), and

MULE (Mutator) (Figure 6.4E, Table S6.4). Not all copies of DDE/D transposase domains in *Blepharisma* contained an intact catalytic triad, suggesting that some may be inactive fragments or pseudogenes. Nonetheless, domains with an intact triad were found in both germline-limited and somatic sequences. In general, the expression level of somatic genes was higher than the expression of the germline-limited ones (Figure S6.4B).

To identify intact transposon units, we examined the seven repeat families in the MAC+IES assembly that could be classified by RepeatClassifier (Figure 6.3B). Of these, only two were predominantly germline-limited and represented by more than one full-length copy, namely rnd-1_family-1 and rnd-1_family-73 (Table S6.5). They contained transposases that were distinct from those found in the MAC (Figure 6.4).

6.2.5.1. Pogo/Tigger-family transposon with abundant MITEs

Repeat elements of rnd-1_family-1 were bound by a ~30 bp terminal inverted repeat (TIR) 5'-CTC CCC CCC CCC CTC CGT GAG CGA ACA AAA-3' (poly-C run length was variable, possibly from assembly errors), and were flanked by a putative target site duplication (TSD) 5'-TAA-3' (or its reverse complement 5'-TTA-3') (Figure 6.4B). All thirty intact (≥95% of consensus length) copies of this family were found within IESs, and had high sequence identity (median 0.5% divergence from consensus).

The encoded transposase (~1200 bp) contained two domains characteristic of Pogo transposases from the Tc1/Mariner superfamily: a DDE/D superfamily endonuclease domain (DDE_1, Pfam PF03184) and a helix-turn-helix domain (HTH_Tnp_Tc5, Pfam PF03221) (B. Gao et al. 2020). The conserved acidic residues ("catalytic triad") characteristic of DDE/D transposases (Yuan and Wessler 2011) were also present, with the motif DD35D, i.e. all three residues were Asp, 35 a.a. between the second and third conserved Asp. Phylogeny of the DDE_1 domain placed the transposase in the Pogo/Tigger family, most closely related to the Tc2 subfamily and a sequence from the oyster *Crassostrea*, all of which also had the DD35D motif (Figure 6.4A). The transposase appeared to be germline-limited, with only ten partial Tblastn hits in the somatic MAC genome (seven on "cruft" contigs) mostly overlapping the HTH_Tnp_Tc5 domain (17 to 84 a.a., E-values 2.3×10^{-12} to 1.4×10^{-6}) and no matches to the DDE_1 domain. However, the TIR did

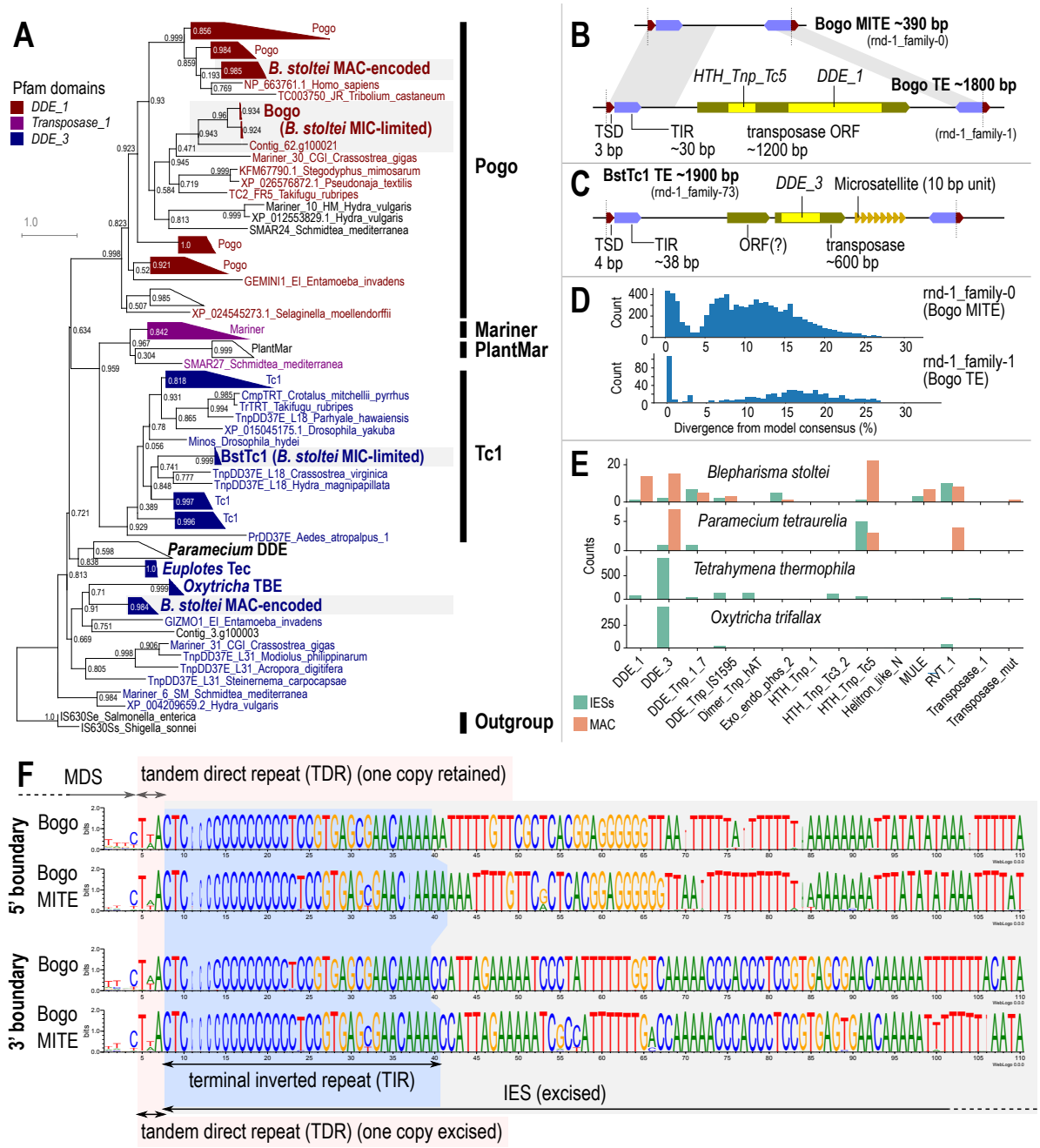


Figure 6.4. Germline-limited repeats include few autonomous transposons but many MITEs. (A) Phylogenetic tree of DDE/D domains for Tc1/Mariner superfamily, including *B. stoltei* germline-limited (Bogo and BstTc1) and somatic transposases. (B) Diagram of features in Bogo and BogoMITE; TSD - target site duplications, TIR - terminal inverted repeats, HTH_Tnp_Tc5, DDE_1 - conserved domains. (C) Diagram of features in BstTc1: DDE_3 - conserved domain. (D) Histograms of sequence divergence from repeat family consensus for copies of the Bogo and Bogo MITE repeat families annotated by RepeatMasker; for rnd-1_family-1, most low-divergence copies (<5% divergence) were short fragments, but all full-length copies were low-divergence. (E) Counts of transposase-related domains in different ciliates, comparing somatic (green) to germline-limited (orange).(F) Sequence logos for Bogo and BogoMITE repeat boundaries, aligned on the terminal inverted repeats (TIRs) and terminal direct repeats (TDRs). 3'-boundaries have been reverse complemented to show the TIRs. Sequence logos were generated from alignments of full-length, intact Bogo elements (>1.8 kbp) and BogoMITEs (between 385-395 bp), with columns comprising >90% gaps removed.

not match previously characterized TIR signatures for the Tc2, Fot, and Pogo subfamilies. A search of all *B. stoltei* IES sequences against HMMs for known DNA transposon TIRs in the Dfam database found only three matches with E-value < 0.01, none from the above subfamilies.

The same TIR and TSD were also found in another repeat family rnd-1_family-0, which was the most abundant repeat in the genome (Figure S6.4A), but these were short elements (consensus ~390 bp) without any predicted coding sequences. rnd-1_family-0 elements were often "mobile IESs" (Arnaiz et al. 2012): they constituted most of the ~389 bp IES size class (Figure 6.3C); the TSDs bounding the repeats (TAA/TTA) were the TDRs for most of these IESs (Figure 6.2C), and the C-rich TIR motif corresponded to the C-rich IES junctions (Figure 1D, Figure 2F). Copies of rnd-1_family-0 were also found nested in longer IESs, suggesting recent proliferation (Figure S6.3C). Degenerated or partial copies were found in shorter IESs (Figure 3C), with copies >5% divergence from consensus having median length 308 bp, vs. 388 bp for copies <5% divergence (Figure 6.4D).

Therefore, we interpreted rnd-1_family-1 as a new Pogo/Tigger transposon, with a non-autonomous derivative MITE, rnd-1_family-0. We propose the names Bogo for the transposon and BogoMITE for its MITE. Given their palindromic nature, both the rnd-1_family-87 and rnd-1_family-160 repeats may also be MITE IESs.

6.2.5.2. Tc1-family transposon with microsatellites

Another IES-limited repeat family, rnd-1_family-73, also contained a coding sequence for a DDE/D-type transposase. Twenty-two copies were "full-length" (>80% consensus) with low sequence divergence from the consensus (median 0.6%). A putative complete transposon bounded by a TSD 5'-TATA-3' and a 38 bp TIR 5'-GTA CCC CCC CCC TCG TTT GTC GCA TTT TCT AGT TTT TT-3' could be defined after manual curation of repeat boundaries (Figure 6.4C). Nine of these were mobile IESs, with the TSDs corresponding to the IES junctions. The remaining cases were nested in larger IESs alongside other repeat elements. Ten repeats also contained a microsatellite with ~5 to 42 copies of its 10 bp repeat unit 5'-GGG AAG GAC T-3' (Figure 6.4C) not found elsewhere in the genome. We propose the name BstTc1 for this putative transposon.

The transposase encoded in full-length copies of BstTc1 contained a conserved DDE/D superfamily domain (DDE_3, Pfam PF13358), phylogenetically affiliated to the Tc1 family although the exact placement is unclear, grouping with only moderate support with Tc1 elements from *Crassostrea* and *Hydra* (Figure 6.4A). Its catalytic triad motif DD34E differed from previously reported motifs for the Tc1 family, DD41D, DD37D or DD36E (Dupeyron et al. 2020), so it may be a novel subfamily.

6.2.6. Non-LTR retrotransposon sequences in both the somatic and germline genomes

Three retrotransposon repeat families in the MAC+IES assembly were classified by RepeatClassifier, i.e. "LINE" or "LINE/RTE-X" (Table S6.5). Two of these were more closely related with numerous very high identity sequences (>97%) (Figure 6.5A), suggesting recent radiation of two related retrotransposon elements, while the third was more divergent (Figure 6.5B; Supplementary Information). Unlike the Bogo and BstTC1-derived elements, more retrotransposon-derived sequences were detected in the *B. stoltei* somatic MAC genome than in assembled IESs (Figure 6.4E, Table S6.5). However genes in IESs may be undercounted because of (i) lower completeness of germline vs. somatic assembly; (ii) indels caused by the lower accuracy of the uncorrected long reads used to assemble IESs that prevent prediction; and (iii) shorter total length of IESs than somatic sequence. Consistent with them being true somatic sequences, mappings of error-corrected HiFi reads from a MAC-enrichment library spanned well into flanking regions (Figure 6.5C; Figure S6.5A, S6.5B). In each repeat family, some loci showed sharp dips in coverage suggesting partial excision as IESs while other loci did not (Figure S6.5B). In MAC-enriched DNA, coverage of such sequences is well above residual IES coverage (Figure S6.1B).

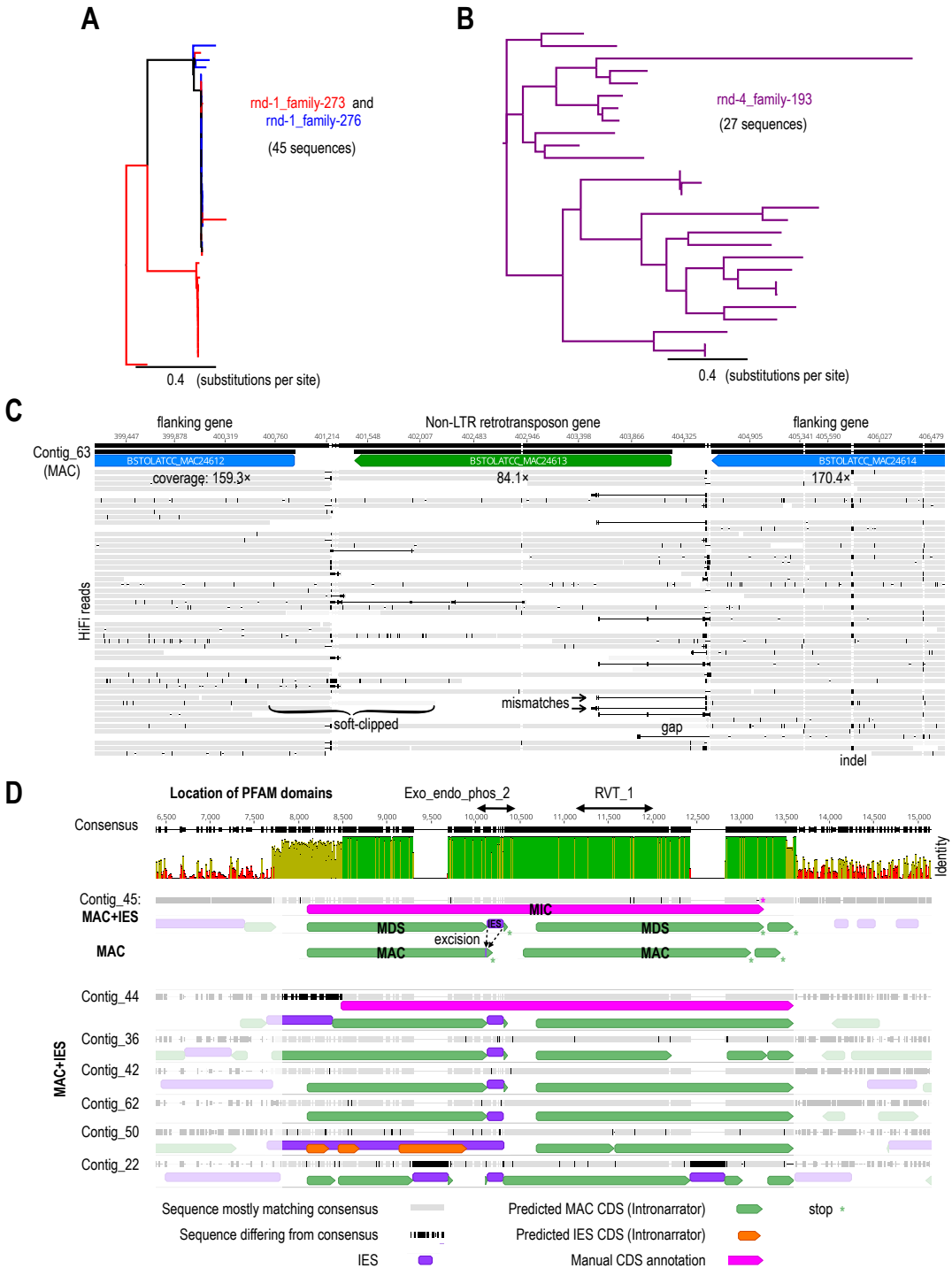


Figure 6.5. Non-LTR retrotransposon sequences in both somatic and germline genomes. (A) Phylogeny of rnd-1_family-273 and rnd-1_family-276 retrotransposon sequences. (B) Phylogeny of rnd-4_family-193 retrotransposon sequences. (C) Window of mapped HiFi reads from sucrose gradient-purified MACs (grey) spanning a retrotransposon gene with both an AP endonuclease domain and a reverse transcriptase domain (from rnd-4_family-193). Only sequence columns with < 90% gaps are shown. Soft clipped regions of reads able to align to flanking sequences are lighter gray. Mismatches and gaps are black. (D) Multiple sequence alignment of non-LTR retrotransposon copies from rnd-1_family-273. Schematic for consequences of IES excision (Contig_45). Identity scale: green=100%; gold=30-99.9%; red=0-29.9%.

Twenty-nine genes in the main somatic assembly encoded full or partial reverse transcriptase domains (RVT_1, Pfam PF00078) (Singh et al. 2021). The four longest retrotransposon genes also encode an N-terminal apurinic/apyrimidinic endonuclease (AP) endonuclease (Exo_endo_phos_2, Pfam PF14529) domain upstream of RVT_1. This domain pair is characteristic of some proteins from non-LTR retrotransposons/LINE-like transposable elements, e.g. the BS element from *Drosophila melanogaster* (UniProt Q95SX7) (Udomkit et al. 1995; Han 2010). In contrast to the development-specific upregulation of retrotransposon genes in *Tetrahymena* (Fillingham et al. 2004) and *Oxytricha* (X. Chen et al. 2014), expression of *Blepharisma* genes encoding proteins containing RVT_1 or Exo_endo_phos_2 domains was negligible in starved cells and throughout a post-conjugation developmental time series, for both germline-limited and somatic genes (Figure S6.4B) (Singh et al. 2021). The only exception was a somatic APEX1 protein homolog (BSTOLATCC_MAC3189). APEX1 is involved in DNA repair (Fritz 2000), and Blastp best matches of the *Blepharisma* protein to GenBank's NR database are other similarly annotated proteins.

Six retrotransposon-derived sequences from repeat family rnd-1_family-273 contained a central IES that encoded almost half the amino acids of an Exo_endo_phos_2 endonuclease domain (Figure 6.5D). Excision of the IES during development would thus knock out the endonuclease domain in the somatic version of the gene. Furthermore, the repeat units as a whole had >99% identity to each other over their ~4.1 kbp length, and were flanked by dissimilar sequences (Figure 6.5D). The similar length of these IESs (173 to 182 bp), their homologous location relative to the coding sequence, and their high sequence identity (>96%) all point to a replication of an ancestral retrotransposon which coincidentally contained a sequence recognized and excised as an IES. In two of these cases, the endonuclease and reverse transcriptase domains can be linked into a single reading frame when the IES is present (Figure 6.5D). None of *Blepharisma*'s putative domesticated transposases are anywhere near as abundant as the retrotransposon repeats in the somatic genome, let alone show signs of substantial recent replication.

6.2.7. Development-specific 24 nt small RNAs are likely scnRNAs in *Blepharisma stoltei*

Small RNA (sRNA) libraries were sequenced from a developmental time series, where two complementary mating types of *B. stoltei* (strains ATCC 30299 and HT-IV) were separately

gamone-treated and then mixed to initiate conjugation. Expression patterns of somatic genes from mRNA-seq and the morphological staging have been reported previously (Chapter 3) (Singh et al. 2021). Briefly: after mating types were mixed (0 h), cells paired, produced gametic nuclei by meiosis and exchanged them (2 to 18 h), followed by karyogamy (18 to 22 h) and development of the zygotic nuclei to new macronuclei (22 h onwards). At 38 h, about a third of observed cells were exconjugants.

The most abundant sRNA length classes were 22 and 24 nt, comprising 32% and 30% of the total reads respectively (Figure 6.6A), consistent with other ciliates, where Dicer-generated, mRNA-derived siRNAs employed in gene silencing are typically 21 or 22 nt long, whereas development-specific sRNAs are distinct and consistently ≥2 bp longer (Mochizuki et al. 2002; Lepère et al. 2009).

Developmental dynamics of the 24 nt *Blepharisma* sRNA resembled scnRNAs of other species. Coverage of 24 nt sRNAs mapping to all feature types initially increased from 2 to 6 h and appeared to plateau up to about 14 h. Coverage over IESs then continued to increase from 14 h to 22 h, reaching ~25 RPKM until the end of the experiment (38 h), whereas coverage declined over coding sequences (CDSs) and other genomic regions ("NON") after 14 h. The initial increase across all feature types coincided with meiotic stages iv to viii of (Miyake, Rivola, and Harumoto 1991) (Singh et al. 2021), whereas the divergence between IESs and the rest of the genome corresponded to the onset of karyogamy (Figure 6B). In contrast, 22 nt sRNAs were initially abundant (albeit with high variance) at CDS and NON regions but low (<1 RPKM) at IESs, and declined sharply to <5 RPKM in all features from 6 h onwards (Figure 6.6B).

Blepharisma 24 nt sRNAs had a strongly conserved 5'-U base preference, like scnRNAs in other ciliates (Lepère et al. 2009; Zahler et al. 2012; Mochizuki and Kurth 2013). For 24 nt sRNAs mapping to IESs, this 5'-U bias was consistent across all time points, except for a slight decrease at 6 h time point (Figure 6.6D, S6.6). Those mapping to CDSs initially displayed no biases, but from 6 h onwards displayed a 5'-U bias. We interpret this to mean that 24 nt sRNAs that mapped on IESs were predominantly scnRNAs at all time points, whereas those mapped to CDSs were initially siRNAs and other types of small RNAs, before being dominated by scnRNAs from 6 h onwards. In contrast, 22 nt sRNAs mapping to CDSs showed no strong biases for any base at any position, whereas 22 nt reads mapping to IESs had a small to moderate 5'-U bias

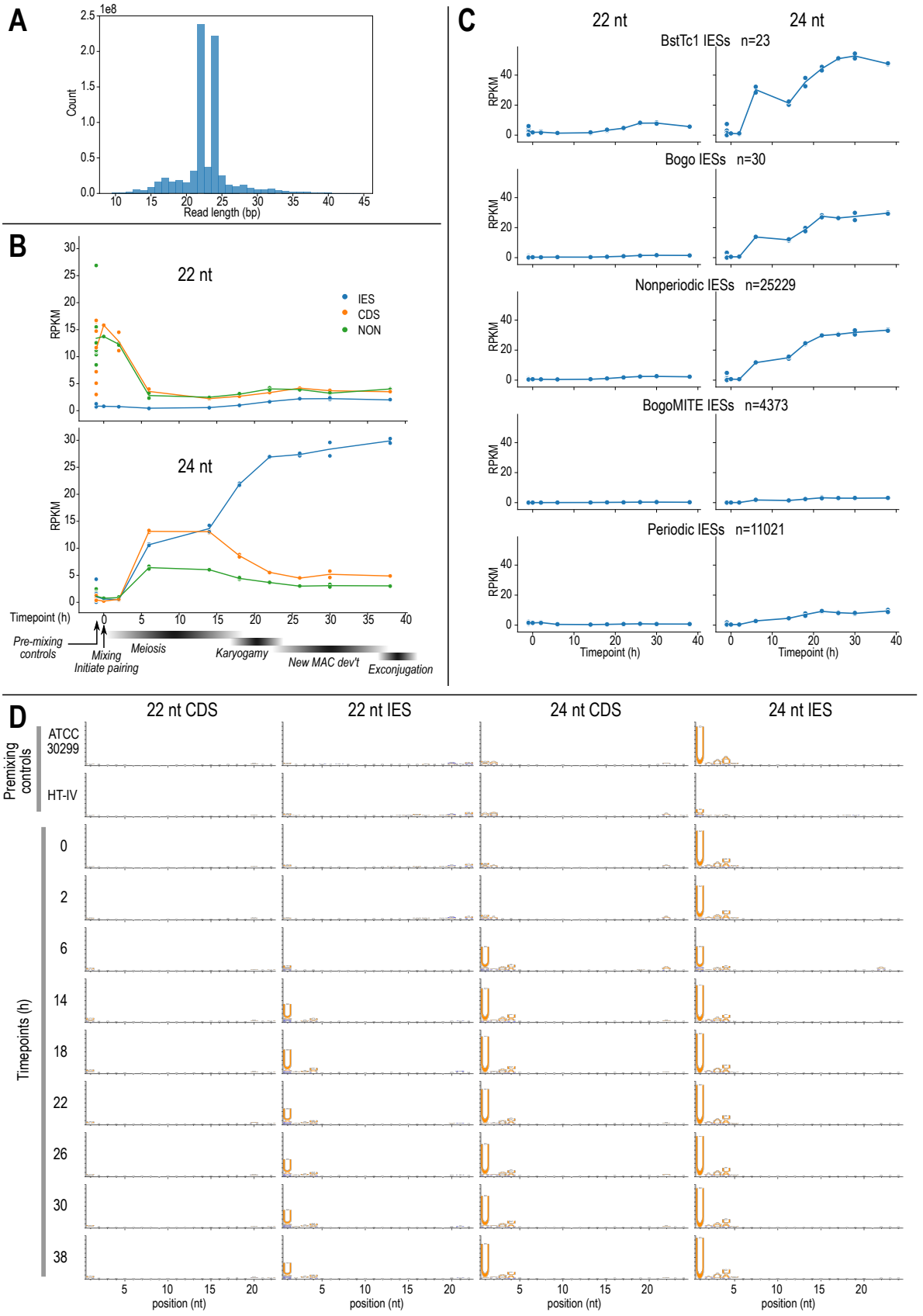
only from 6 h onwards. The 5'-U bias in 22 nt sRNAs mapping to IESs may represent true 22 nt scnRNAs, or fragments of originally 24 nt scnRNAs.

6.2.8. Putative scnRNAs have lower coverage over periodic IESs and BogoMITE IESs

The relative expression levels of the putative scnRNAs also differed between IES size classes. Based on the IES length distribution and repeat content, we divided IESs into five groups: (1) short "periodic" IESs (≤115 bp), (2) BogoMITE mobile IESs, because that was the most abundant repeat family, (3) IESs with full-length Bogo transposons, (4) IESs with full-length BstTc1 transposons, and (5) all other IESs ("non-periodic"). BogoMITE IESs and periodic IESs had lower scnRNA coverage (max ~5 and 10 RPKM respectively) compared with nonperiodic IESs (~30 RPKM). The former were comparable to or even lower than expression levels over non-IES features (Figure 6.6C). Nonetheless, scnRNA coverage of BogoMITE IESs and periodic IESs showed an initial increase then plateau, without the subsequent decline seen in non-IES regions. Bogo-containing IESs had similar scnRNA coverage to other non-periodic IESs, but BstTc1-containing IESs had higher coverage (Figure 6.6C).

Because of the repetitive sequence content in IESs and the short sRNA length, it is possible that the expression levels calculated could be affected by mis-mapping. We reason that such mismapping would not influence the results described above, because "periodic" IESs (group 1) had low repetitive content, whereas the transposon-containing IESs (groups 2, 3, 4) each represented a single repeat family so any mismappings would be contained within the same group and count towards the same RPKM value.

Figure 6.6. Development-specific 24 nt small RNAs are likely scnRNAs in *B. stoltei*. (A) Read length histogram for all sRNAs in the time series. (B) Relative expression (RPKM units, vertical axis) of 22 and 24 nt sRNAs mapping to different feature types across time series: blue - IES, orange - CDS, green - all other regions not annotated as IES or CDS (including UTRs and intergenic regions which are difficult to delimit exactly with available data). Timing of developmental stages inferred from morphology are labeled below (Chapter 3) (Singh et al. 2021). (C) Relative expression of 22 and 24 nt sRNAs mapping to different categories of IESs: containing full-length copies of BstTc1 and Bogo transposons, at least 90% covered by BogoMITE elements, IESs in the periodic length range (< 115 bp), and all other IESs ("non-periodic"). (D) Sequence logos for 22 and 24 sRNAs mapping to CDS and IES features in controls and different time points (rows).


A
1e8
Count
Read length (bp)
B
22 nt
24 nt
RPKM
IES
CDS
NON
Timepoint (h)
Pre-mixing controls
Mixing
Initiate pairing
Meiosis
Karyogamy
New MAC dev't
Exconjugation
C
22 nt
24 nt
BstTc1 IESs n=23
Bogo IESs n=30
Nonperiodic IESs n=25229
BogoMITE IESs n=4373
Periodic IESs n=11021
RPKM
Timepoint (h)
D
22 nt CDS
22 nt IES
24 nt CDS
24 nt IES
Premixing controls
ATCC 30299
HT-IV
Timepoints (h)
0
2
6
14
18
22
26
30
38
position (nt)

6.3. Discussion

The germline genome of *Blepharisma stoltei,* belonging to the earliest diverging lineage of ciliates sequenced to date, has similarities to established model species, especially the short, periodic IES lengths resembling those of *Paramecium*. It also provides fresh observations, notably recent proliferation of non-autonomous mobile elements (MITEs) that have autonomous counterparts in the same genome, and the finding of retroelements in the somatic genome.

6.3.1. Comparison to IESs in other ciliates

Most *Blepharisma* IESs are short, TA-bound, and intragenic, more similar to *Paramecium* than *Tetrahymena* or spirotrichs. The most striking parallel is the sharply periodic length distribution of short IESs with peaks every ~10 bp, coinciding with the DNA helical turn, implying that the *Blepharisma* excisase complex has analogous geometric constraints as proposed for *Paramecium* (Arnaiz et al. 2012). *Blepharisma* "periodic" IESs are longer on average and do not have a "missing" second peak, but the last peak (~110 bp) is still below the persistence length of DNA. In contrast, *Tetrahymena thermophila* has a continuous distribution (average length ~3 kbp) (Hamilton et al. 2016; Seah and Swart 2021), while *Oxytricha trifallax* non-scrambled IESs (length ~20 bp) have weak periodicity (X. Chen et al. 2014). Periodicity is consistent with a single primary IES excisase, rather than multiple excisase families, which would smooth the length distribution.

Longer, nonperiodic IESs of *Blepharisma* contain more repeats, including whole transposons, resembling *Tetrahymena* where 41.7% of high-confidence IESs comprise putative transposons (Hamilton et al. 2016), some of which can be grouped into families (Wuitschick et al. 2002; Fillingham et al. 2004). However, *Blepharisma*'s long IESs encode much fewer transposases, and their length distribution is not unimodal but long-tailed with distinct peaks representing individual abundant families. Germline-specific repeats and transposons across *Paramecium* spp. have recently been surveyed (Sellis et al. 2021), but were likely underestimated because such repeats are difficult to assemble from short-read data even with high coverage, as we saw with *Blepharisma* BogoMITE elements, (Supplementary Information, Figure S6.1A).

Consistent with the scnRNA turnover model, *Blepharisma* 24 nt sRNAs mapping to IESs increase during post-conjugation development more than those mapping to CDSs. In this model, RNA intermediates are produced from both IESs and MDSs (Malone et al. 2005;

Mochizuki and Gorovsky 2005), but those from MDSs are selectively degraded, allowing the remaining scnRNAs to mark IESs for excision. This complements our finding that MAC-encoded homologs to proteins involved in scnRNA biogenesis, Dicer-like (Dcl) and Piwi, are highly upregulated during development (Singh et al. 2021). Furthermore, higher coverage of *Blepharisma* scnRNAs in longer (presumably younger) IESs than in short (~older) periodic IESs mirrors the pattern in *Paramecium*, where younger IESs are more likely to require scnRNAs for efficient excision (Sellis et al. 2021; Lhuillier-Akakpo et al. 2014).

6.3.2. Are MITEs a missing link in the IBAF model?

The prevailing Invasion-Bloom-Abdication-Fade (IBAF) model for the evolution of IESs hypothesizes that they originate from cut-and-paste DNA transposons that invade and proliferate ("bloom") in the germline genome (Klobutcher and Herrick 1997). Transposon proliferation stops ("abdication") when its transposase is domesticated by a host promoter, releasing the transposons from purifying selection, whereupon their sequences erode by drift ("fade"). Depictions of the IBAF model usually show all the transposons expressing transposases during "bloom", i.e. functioning as autonomous transposons (Klobutcher and Herrick 1997; Y. Feng and Landweber 2021). This is reasonable for *Tetrahymena* and *Oxytricha*, which have hundreds of germline-encoded transposases that vastly outnumber those in the somatic genome (Table S6.4). However, *Blepharisma* only has a few dozen transposases in total, like *Paramecium*, although germline-limited transposases in the latter may also be more severely underestimated by short read sequencing.

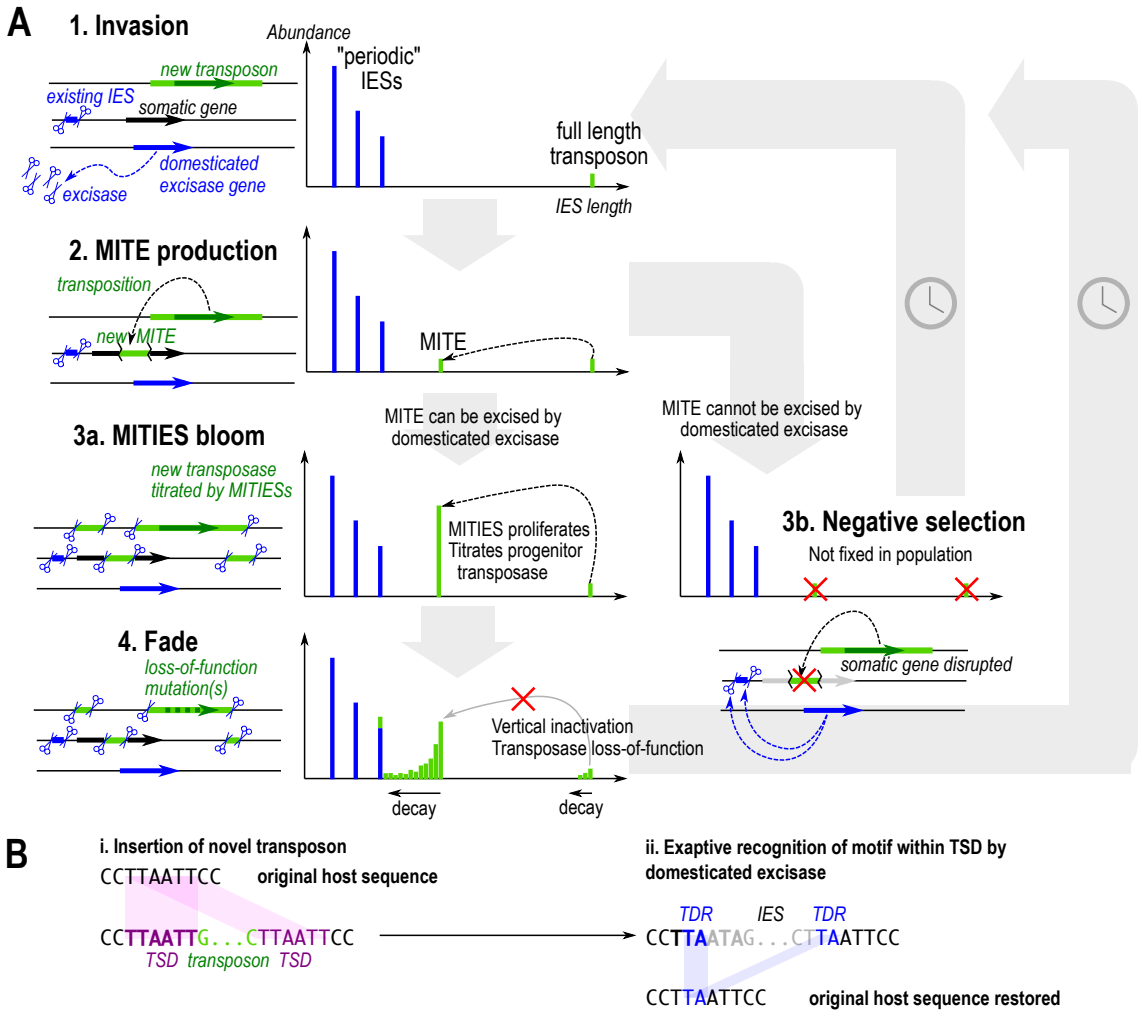


Figure 6.7. Model for transposon fixation as IESs in a ciliate genome with an existing domesticated excisase. (A) Graphs depict IES length distribution. (1) Invasion of germline genome by full length transposon (green); existing IESs (blue) are excised by domesticated excisase. (2) New transposon produces MITEs. (3a) If MITEs can be excised by domesticated excisase, they proliferate and titrate the progenitor transposase. (4) MITEs proliferation causes vertical inactivation of the full length transposon; loss of function stops production of new MITEs, leading to eventual decay. (3b) If the MITE cannot be excised by domesticated excisase, it is more likely to cause deleterious mutations upon insertion, and is therefore selected against and does not reach fixation. (B) If a transposon TSD contains a submotif that can be recognized by the domesticated excisase, it can theoretically be excised cleanly without leaving a "footprint", avoiding potential frameshift mutations.

This discrepancy can be resolved by taking MITEs into account. In *Blepharisma* this is best exemplified by the few copies of the autonomous Bogo transposon compared to thousands of non-autonomous BogoMITEs. The consistent lengths of BogoMITEs, their high sequence identity, and occasional nested insertion inside unrelated IESs are the clearest illustrations to date of recent MITE proliferation as mobile IESs. Bogo is also the first Pogo/Tigger transposon found

in a ciliate germline; this subfamily is known to be especially prone to MITE formation (C Feschotte and Mouchès 2000; Guermonprez, Loot, and Casacuberta 2008). The prevalence of IESs bound by terminal inverted repeats, including numerous palindromic IESs (Figure 6.2D, S6.2), also suggest that many more *Blepharisma* IESs are MITE derivatives.

In *Paramecium* spp., MITEs of the Thon and Merou transposons have been identified but only numbered about a dozen copies per genome, and their transposases belong to a different family within the DDE/D superfamily (Figure 6.4). The most abundant mobile IES family in *Paramecium*, FAM_2183, is probably a MITE but its autonomous counterpart was not found (Sellis et al. 2021). MITEs as intermediates in the transposon/IES life cycle can hence explain why *Blepharisma* and *Paramecium* have few MIC-encoded transposases compared to *Oxytricha* and *Tetrahymena*.

MITEs also provide a mechanism for transposon/IES proliferation to be self-limiting (Figure 6.7A). When MITEs outnumber the autonomous transposon, active transposase protein is more likely to bind to target sites in MITEs than the full length transposon ("titration"), hindering the replication of the autonomous version, which buys time for loss-of-function mutations to inactivate the transposases ("fade"). This "vertical inactivation" scenario (Hartl, Lohe, and Lozovskaya 1997) was already discussed in the original IBAF proposal (Klobutcher and Herrick 1997), but no plausible examples from ciliates were known at the time.

6.3.3. Is "genome defense" a flawed analogy?

The IBAF model also does not explain how ciliates can consistently and precisely excise novel mobile elements from different transposon families that invade the germline genome. The domesticated excisases of *Paramecium* (Baudry et al. 2009), *Tetrahymena* (Cheng et al. 2010), and *Blepharisma* (Singh et al. 2021) belong to the PiggyBac family. Except for *Tetrahymena* Tpb2, PiggyBacs are known to perform seamless excision, where the host sequence after transposon excision is identical to that before insertion (Q. Chen et al. 2020). This would make them the ideal progenitor for IESs that insert into coding sequences; indeed, PiggyBac transposons are also known to produce MITEs (Wang et al. 2010; Mitra et al. 2013). By extension, the first IESs probably originated from PiggyBac transposons. But what about subsequent invasions by other transposons that leave behind "scars" upon excision? Such imprecision would cause deleterious frameshift mutations in coding regions. How can they invade the germline genome and yet avoid deleterious effects?

Part of the answer lies in the "hijacking" model proposed from *Paramecium* (Sellis et al. 2021; Arnaiz et al. 2012), whereby the domestication of PiggyBac changed the dynamic for subsequent transposon invasions. New transposons would persist as IESs only if they also encode a seamless excisase, or if they can also be recognized and cut by the already-domesticated PiggyBac (exaptation). The latter favors the invasion of transposons that produce a TSD containing a submotif that can be recognized as a cut site by PiggyBac (Figure 6.7B). The similarity between boundaries of IESs and transposons would hence not be due to common origin or sequence evolution after IESs have fixed in the germline (Klobutcher and Herrick 1997), but rather because of selection for transposons whose TSDs already match the excision site preferences of domesticated PiggyBac. Analogous exaptation of TSDs for excision has been demonstrated in another context: the independent origin of introns from MITEs in at least two different eukaryotes, where one of the TSDs produced upon MITE insertion has been co-opted as an intron splice site (Huff, Zilberman, and Roy 2016). Cross-talk between different (albeit related) transposases for MITE transposition has also been documented (Cédric Feschotte et al. 2005).

We further argue that "genome defense" is a teleological expression that confuses cause and effect. Domesticated excisases actually help mobile elements to accumulate in the germline, because they shield them from selection by effectively excising them from the somatic genome. *Tetrahymena* is the exception that proves the rule: its domesticated excisase appears to be imprecise; correspondingly, most of its IESs are intergenic, because intragenic IESs have been efficiently removed by selection (L. Feng et al. 2017; Cheng et al. 2016). The origins of gene silencing by DNA methylation in vertebrates have also been reinterpreted with similar reasoning. Vertebrates have high levels of CpG methylation that inactivates transposons, which was thus proposed to "compensate for" transposon proliferation in eukaryotic genomes (Bestor 1990). When seen from a non-teleological perspective, it is precisely because CpG-mediated transposon inactivation is so effective, preventing exposure to selection, that transposons persist, leading to larger genomes (Zhou et al. 2020).

6.3.4. Why does the *Blepharisma* somatic genome have retrotransposon-derivatives?

Transposon-related sequences are typically germline-limited in other model ciliates, which was formerly interpreted as successful "genome defense" keeping them out of the somatic MAC genome (Fillingham et al. 2004; Guérin et al. 2017; Hamilton et al. 2016; Swart et al.

2013; X. Chen et al. 2014). We hence did not expect to find several retrotransposon-derived sequences in the *Blepharisma* MAC genome. Some show signs of partial excision or possible absence of the locus in part of the population, but plenty have uniform coverage typical of somatic sequences.

Recent retrotransposon proliferation in the soma, patchy distribution of different somatic transposase classes across ciliates (Table S6.4) (Singh et al. 2021), and recent horizontal acquisition of bacterial genes in *Blepharisma* (Swart et al., in prep.) all suggest that "genome defense" is at best leaky. We conjecture that if foreign DNA lacks suitable target sites recognized by the excisase, it might still be marked by scnRNAs but fail to be excised or be only partially excised (e.g. the IESs in Figure 6.5C). Nonetheless such DNA would still be deleterious if inserted intragenically.

Ciliate somatic MACs may be unable to repress mobile elements by heterochromatinization like germline MICs and other eukaryotic nuclei. In *Tetrahymena*, most MAC DNA is not associated with classical heterochromatin marks (Liu et al. 2007), while in *Paramecium* MACs, H3K27me3 is not associated with transcription repression, despite being a classic heterochromatin mark in multicellular eukaryotes (Drews et al. 2021). In such a permissive expression environment, selection against mobile elements that are not already excised as IESs may be especially effective, unless they are relatively inactive like the *Blepharisma* retroelements. On the other hand, regular *Blepharisma* stock culture passaging maintains a small effective population size, which would be expected to counteract selection against mobile element accumulation in the soma.

The genome defense model may lead one to dismiss IES retention in the somatic genome as excisase inefficiency or MIC contamination of the library, however, IES excision is not all-or-nothing but a continuum. Experimental evolution experiments in *Paramecium* suggest that IES retention variability is itself a plastic and evolvable trait with consequences for genotypic diversity (Catania, Rothering, and Vitali 2021; Vitali, Hagen, and Catania 2019). Assembly algorithms tend to eliminate repetitive and lower-coverage regions, which are characteristic of mobile elements and partially retained IESs, thus presenting an oversimplified, "pristine" view of somatic genomes. Accurate long read sequencing, haplotype-aware assemblers, and sequence graphs will all play a role in building a more realistic picture of genome heterogeneity.

6.3.5. Conclusion

Why have we credited developmental DNA elimination with defending the genome, when natural selection has been doing the hard work? Apart from technical biases during genome assembly, there is also sampling bias by using lab strains. These are often clonal and largely homozygous; if so, we would not observe accumulation of strongly deleterious foreign DNA that actually needs defending against, but only IESs that have reached fixation and that are already efficiently excised and non-deleterious. Purifying selection against deleterious IESs has had to be indirectly observed, e.g. in the lack of intragenic IESs in *Tetrahymena*, where excision is imprecise (Hamilton et al. 2016), and the statistical depletion of quasi-IES sequences in the *Paramecium* somatic genome (Swart et al. 2014). Similar evolutionary logic applies to the CRISPR defense systems of prokaryotes, where hidden fitness costs (autoimmunity) have been underestimated because those individuals are removed by selection (Stern et al. 2010), hence the phenomenon is easily misinterpreted as inheritance of acquired traits (Weiss 2015). Most studies on ciliate developmental DNA elimination to date have focussed on the underlying molecular mechanisms, but to understand its origins and evolution we should expand our view to diverse ciliates and their germline genomes from natural populations.

Supplementary figures

Figure S6.1. (**A**) Comparison of IES reconstructions from MIC-enrichment library sequenced with short reads by ParTIES (above) vs. from long reads by BleTIES (below). Main panels: IES length histograms up to 500 bp, insets: IES retention scores colored by TDR sequence type. Length peak at ~390 bp representing BogoMITE element is present in BleTIES reconstruction but not ParTIES. (**B**) IES retention scores from MAC-enrichment library sequenced with PacBio HiFi reads. (**C**) Retention scores of cryptic IESs from MAC read library, colored by TDR sequence type: TA-bound (blue), no TDR (orange), or a non-"TA" TDR (green). (**D**) Length distribution of cryptic IESs containing "TTA" or "TAA" in their TDR, detail <500 bp, inset detail <150 bp. (**E**) Sequence logos of TA-bound cryptic IES junctions centered on the TA motif, for all cryptic IESs (above) and the subset in the ~72 bp size class (below). (**F**) Mapping pileup at IES with TA-containing TDR. For aligned reads in panels E and F, dots: bases identical to reference, dashes: gaps relative to reference, red bar: read clipping. (**G**) Mapping pileup at IES with non-TA-containing TDR.

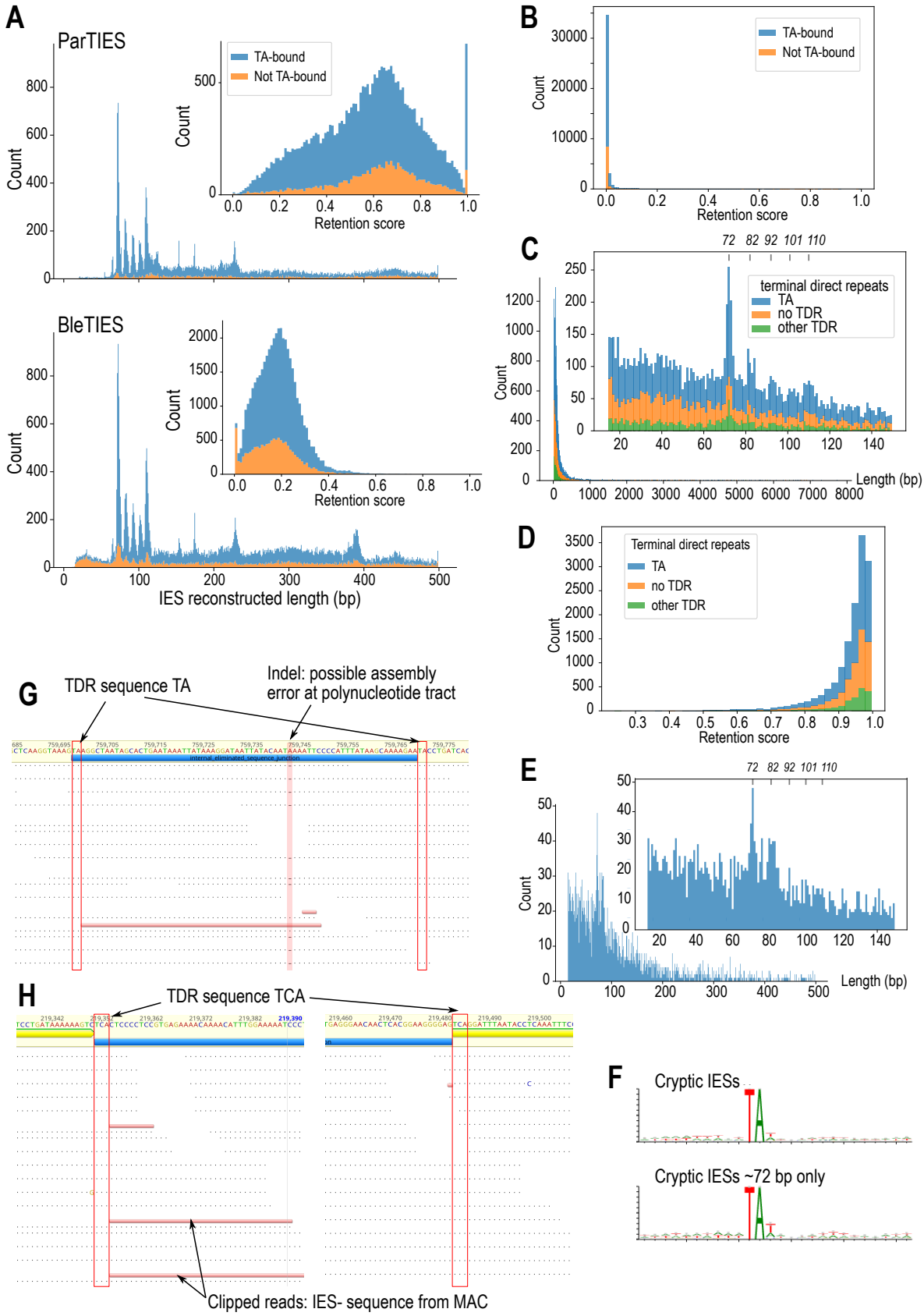

A
ParTIES
TA-bound
Not TA-bound
Count
Retention score
BleTIES
IES reconstructed length (bp)
B
TA-bound
Not TA-bound
Count
Retention score
C
72 82 92 101 110
terminal direct repeats
TA
no TDR
other TDR
Count
Length (bp)
D
Terminal direct repeats
TA
no TDR
other TDR
Count
Retention score
E
72 82 92 101 110
Count
Length (bp)
F
Cryptic IESs
Cryptic IESs ~72 bp only
G
TDR sequence TA
Indel: possible assembly error at polynucleotide tract
H
TDR sequence TCA
Clipped reads: IES- sequence from MAC

Figure S6.2. Strip plots of IES lengths for palindromic IESs (≥90% self-alignment identity), after they have been clustered by sequence identity (rows represent clusters). Each cluster is annotated with the median IES length and the cluster size. Insets: (A) Overall sequence length distribution histogram for all palindromic IESs. The most common length of palindromic IESs is ~230 bp. (B, C) Dendrogram of sequence distance and multiple sequence alignment of palindromic IESs with ~230 bp length to illustrate that they comprise several distinct, unrelated sequences.

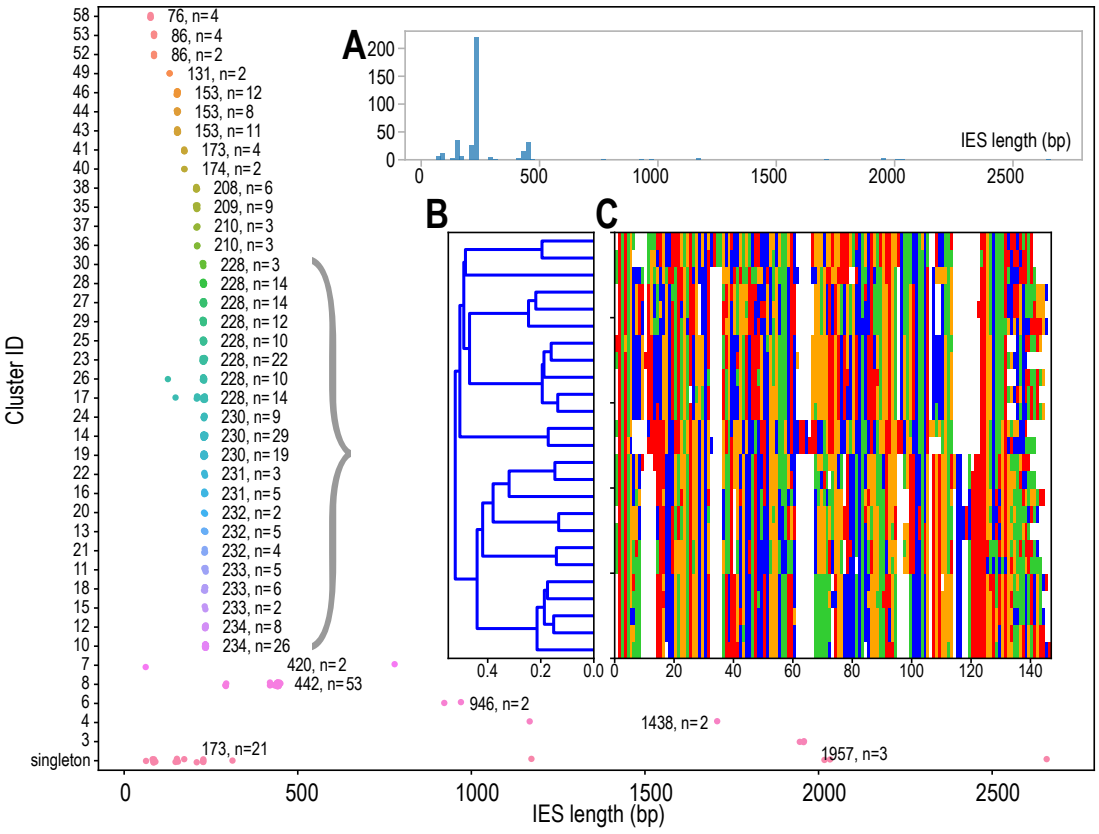
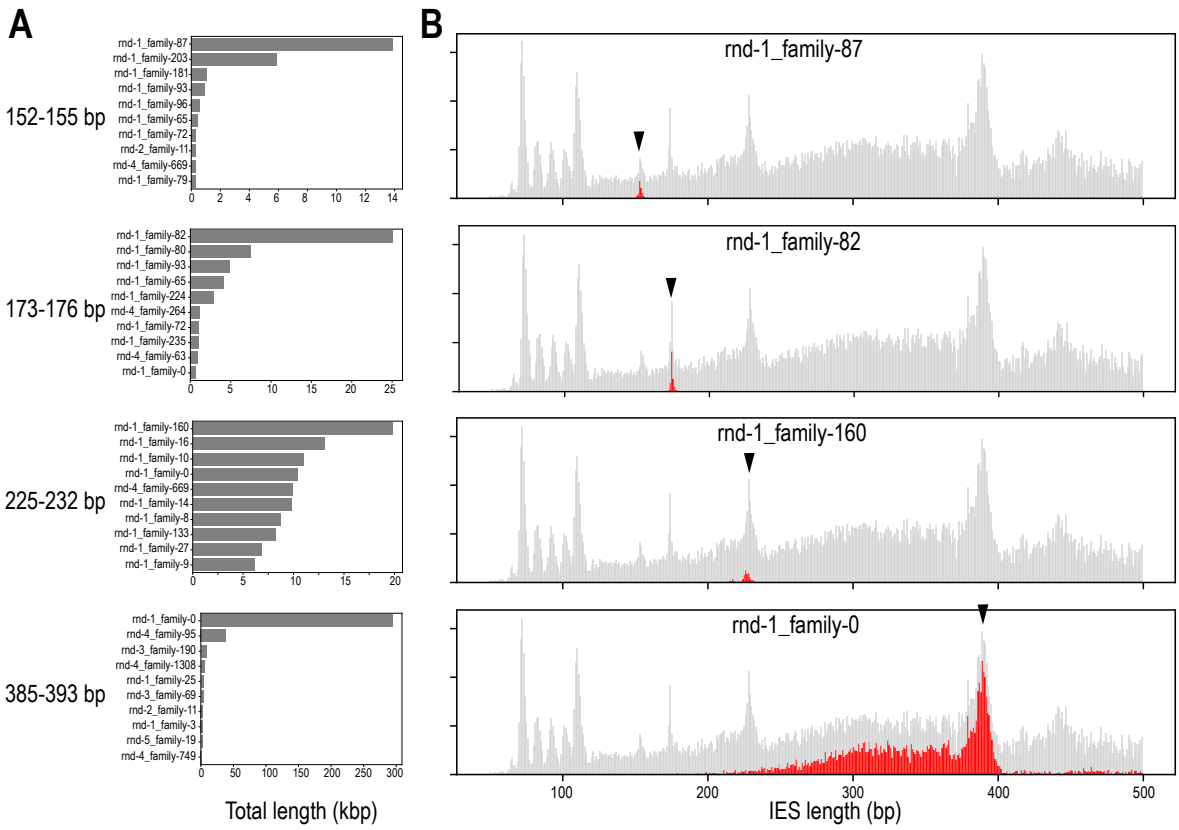
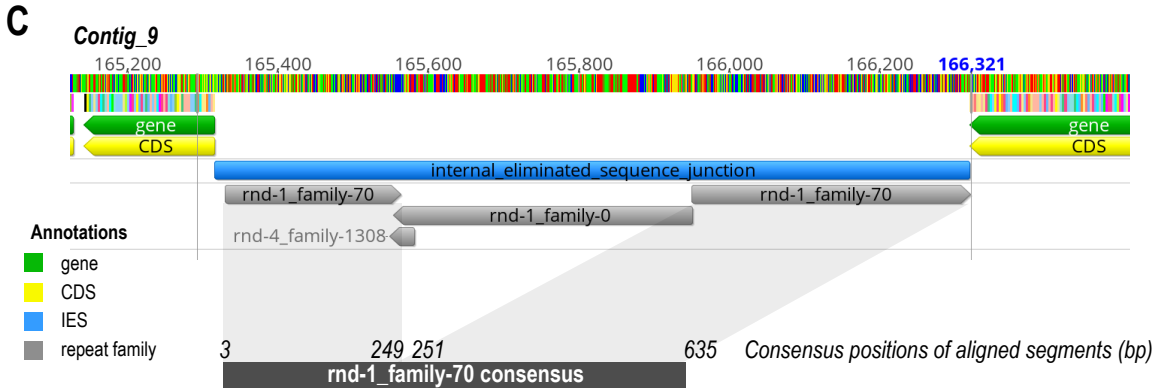


Figure S6.3. Most abundant repeat families in non-periodic IES size classes. (A) Total lengths (horizontal axis) of the top ten repeat families per IES size class (panel rows). (B) Top repeat family (by sequence length) for each IES size class (panel rows); the total length covered by that repeat family within IESs vs. the lengths of those IESs is shown in red, superimposed on the total sequence vs. IES length distribution of IESs in general (grey). Arrowheads mark centers of the size classes. (C) Examples of nested repeats within IESs. Nested elements can be recognized when the two outer repeat elements belong to the same family and align to consecutive parts of its family's consensus sequence, implying that the inner element has likely been inserted into the middle of an existing element. Coordinates of the split segments are relative to the repeat family consensus.


A
152-155 bp
rnd-1_family-87
rnd-1_family-203
rnd-1_family-181
rnd-1_family-93
rnd-1_family-96
rnd-1_family-65
rnd-1_family-72
rnd-2_family-11
rnd-4_family-669
rnd-1_family-79
0 2 4 6 8 10 12 14
173-176 bp
rnd-1_family-82
rnd-1_family-80
rnd-1_family-93
rnd-1_family-65
rnd-1_family-224
rnd-4_family-264
rnd-1_family-72
rnd-1_family-235
rnd-4_family-63
rnd-1_family-0
0 5 10 15 20 25
225-232 bp
rnd-1_family-160
rnd-1_family-16
rnd-1_family-10
rnd-1_family-0
rnd-4_family-669
rnd-1_family-14
rnd-1_family-8
rnd-1_family-133
rnd-1_family-27
rnd-1_family-9
0 5 10 15 20
385-393 bp
rnd-1_family-0
rnd-4_family-95
rnd-3_family-190
rnd-4_family-1308
rnd-1_family-25
rnd-3_family-69
rnd-2_family-11
rnd-1_family-3
rnd-5_family-19
rnd-4_family-749
0 50 100 150 200 250 300
Total length (kbp)
B
rnd-1_family-87
rnd-1_family-82
rnd-1_family-160
rnd-1_family-0
100 200 300 400 500
IES length (bp)


C
Contig_9
165,200 165,400 165,600 165,800 166,000 166,200 166,321
gene
CDS
gene
CDS
internal_eliminated_sequence_junction
rnd-1_family-70
rnd-1_family-70
rnd-1_family-0
rnd-4_family-1308
Annotations
gene
CDS
IES
repeat family
3 249 251 635 Consensus positions of aligned segments (bp)
rnd-1_family-70 consensus

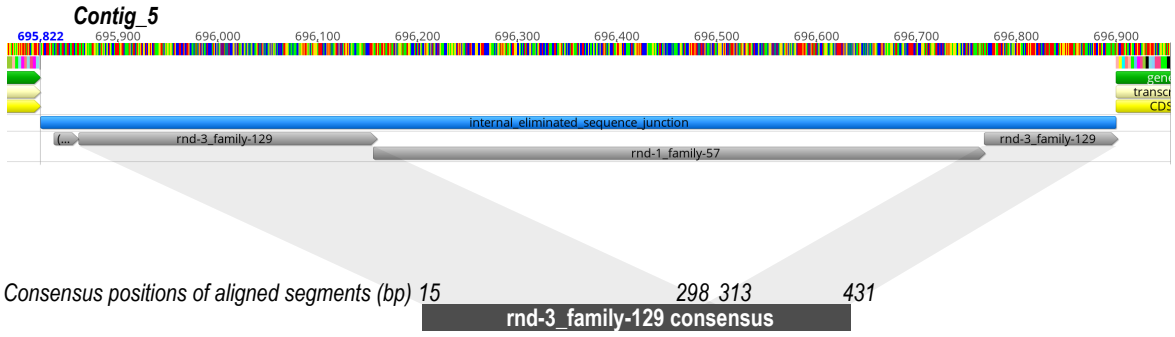

Contig_5
695,822 695,900 696,000 696,100 696,200 696,300 696,400 696,500 696,600 696,700 696,800 696,900
internal_eliminated_sequence_junction
rnd-3_family-129
rnd-3_family-129
rnd-1_family-57
Consensus positions of aligned segments (bp) 15 298 313 431
rnd-3_family-129 consensus

Figure S6.4. Expression of genes with transposase domains. Comparison of expression levels for MAC- vs. MIC-limited transposase-related domains across developmental time series; heatmap color scaled to log(transcripts per million). Domain architecture shown diagrammatically.

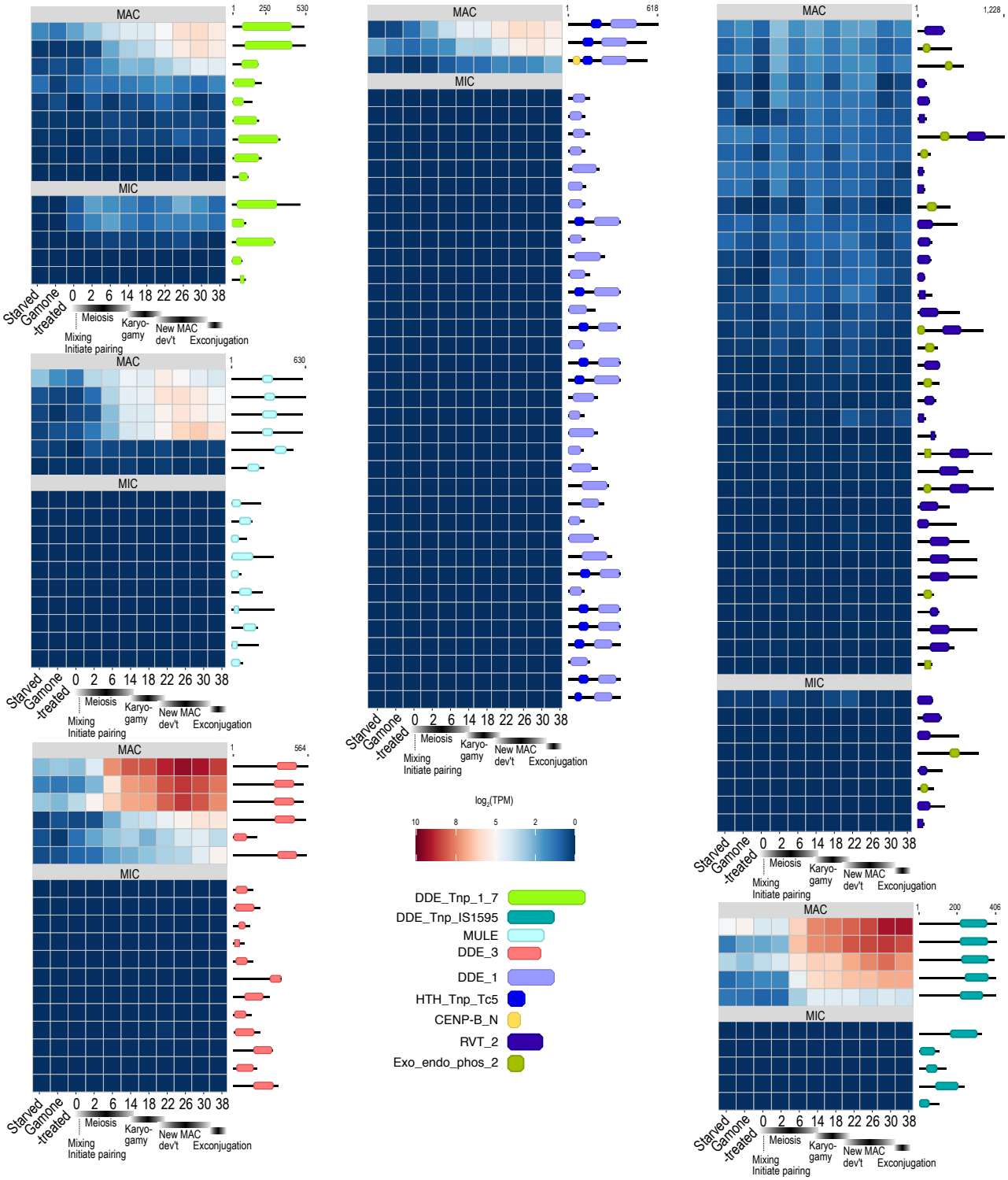


Figure S6.5. (**A**) As in Figure 6.5A. (**B**) As in Figure 6.5A. Inset shows coverage across the entire contig and position of the retrotransposon gene. (**C**) Alignment of MAC+IES and somatic genomic sequences for Contig_44 retroelement genes from Figure 6.5A, showing how excision of the central IES deletes part of the endonuclease domain and produces a premature stop codon.

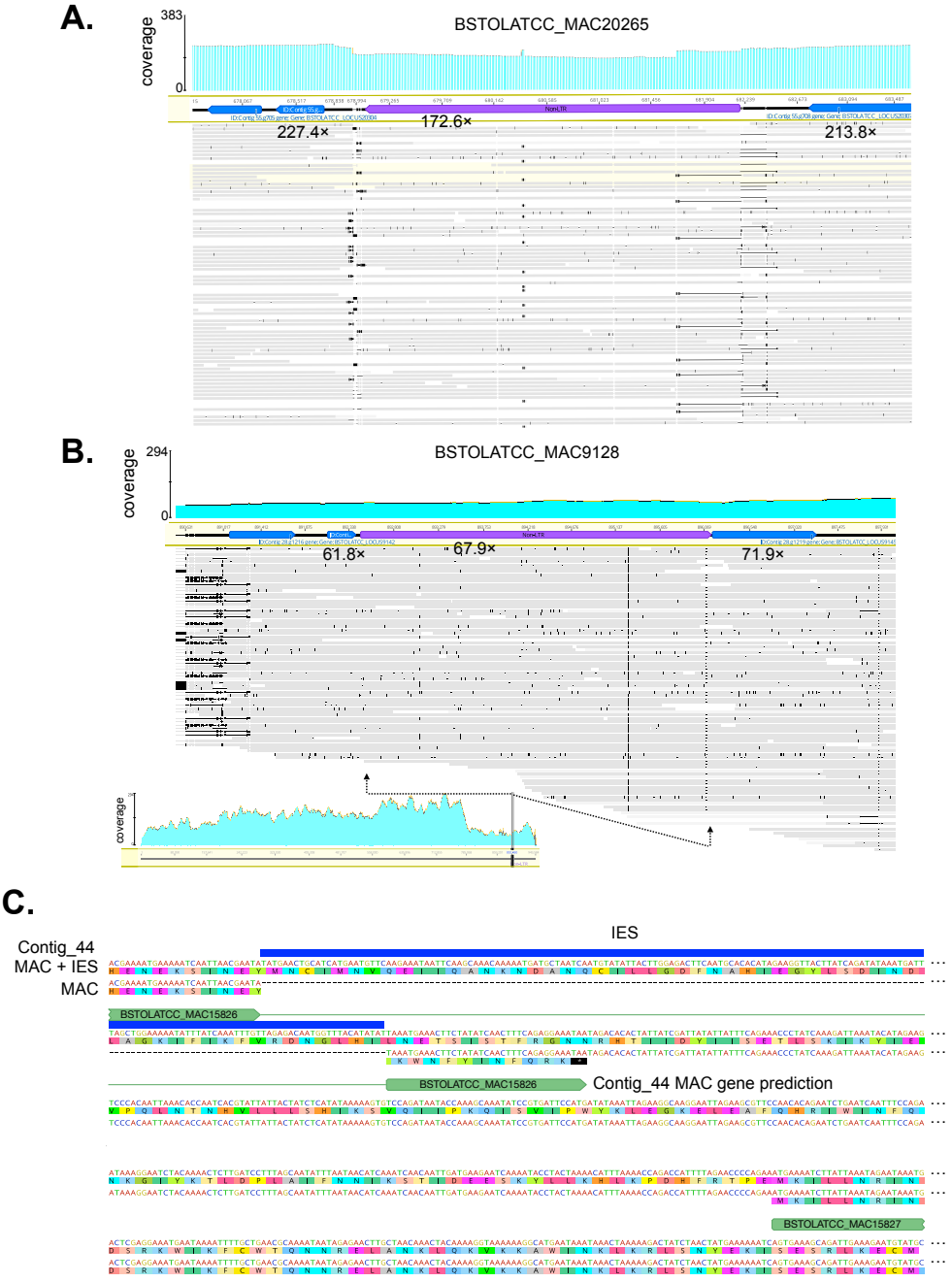
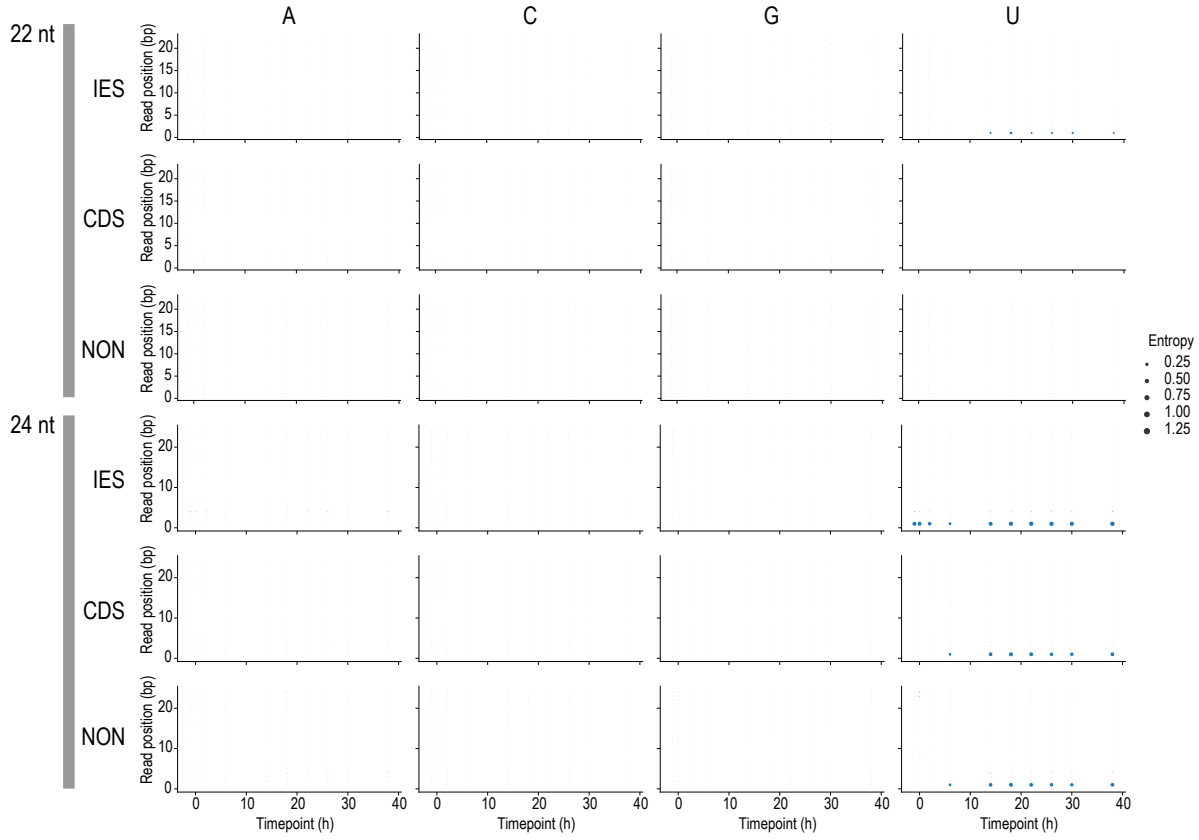


Figure S6.6. Per-position base entropy of 22 nt and 24 nt sRNAs from developmental time series showing conservation of 5'-U in 24 nt sRNAs. Each plot symbol represents positional sequence entropy (symbol size) for a given nucleotide base (columns) and position in the sRNA sequence (vertical axis) and time point (horizontal axis), in sRNAs mapping to different feature types (rows).



Supplementary tables

Table S6.1.

IES size classes, defined by peak calling on the length distribution of TA-bound IESs. Lower and upper lengths per size class are inclusive. Only TA-bound IESs on main assembly contigs are included in the counts and total lengths.

Peak center (bp)	Lower bound (bp)	Upper bound (bp)	No. IESs	Total IES length (bp)
65	64	66	207	13415
72	70	74	2717	195724
82	80	84	1035	85017
92	90	94	819	75458
101	99	103	688	69638
110	108	112	1592	175060
153	151	155	336	51478
174	173	175	377	65575
228	225	231	769	175422
389	385	393	876	340798

Table S6.2.

Summary of RepeatMasker annotations in *B. stoltei* MAC+IES assembly for each repeat class, as classified by RepeatClassifier. The most abundant repeat family (rnd-1_family-0) is also listed separately, despite being unclassified. Only one family, rnd-1_family-1, is classified as DNA/TcMar-Tc2. Total annotated length does not account for overlapping annotations.

Class	Number of annotated elements	Total sequence length annotated (bp)
Unknown (excluding rnd-1_family-0)	41836	11279760
rnd-1_family-0 (Unknown)	8369	2692873
Simple_repeat	6878	613736
Low_complexity	2511	123672
rnd-1_family-1 (DNA/TcMar-Tc2)	539	104263
LINE/RTE-X	94	51679
LTR/Pao	39	10475
LINE	28	38630
DNA/TcMar-Tc1	24	38070
Unknown/Helitron-2	23	11025

Table S6.3.

Top five most abundant repeat families in specific IES size classes (defined in Table S1). Repeats comprising > 20% of the total IES length of particular size classes are highlighted in bold font.

Repeat family	Number	Fraction of total IES length	IES size class (peak center bp)
rnd-1_family-397	712	0.044433	65
A-rich	134	0.008362	65
rnd-1_family-157	67	0.004181	65
rnd-1_family-151	65	0.004056	65
rnd-4_family-596	64	0.003994	65
A-rich	2735	0.012229	72
rnd-1_family-438	1898	0.008487	72
rnd-1_family-397	1511	0.006756	72
rnd-1_family-398	508	0.002271	72
(T)n	450	0.002012	72
A-rich	741	0.007556	82
rnd-1_family-397	441	0.004497	82
rnd-2_family-94	182	0.001856	82
rnd-1_family-0	171	0.001744	82
(AT)n	153	0.001560	82
A-rich	570	0.006505	92
rnd-1_family-205	400	0.004565	92
rnd-1_family-0	270	0.003081	92
rnd-2_family-11	209	0.002385	92
rnd-3_family-853	160	0.001826	92
A-rich	801	0.009991	101
rnd-3_family-853	679	0.008469	101
rnd-4_family-1308	277	0.003455	101
rnd-1_family-0	174	0.002170	101

(TATAA)n	128	0.001596	101
rnd-3_family-853	2275	0.011457	110
A-rich	1134	0.005711	110
rnd-2_family-11	336	0.001692	110
rnd-2_family-94	247	0.001244	110
rnd-1_family-210	239	0.001204	110
rnd-1_family-87	14889	0.236551	153
rnd-1_family-203	5865	0.093181	153
rnd-1_family-181	1331	0.021146	153
rnd-1_family-93	1059	0.016825	153
rnd-4_family-669	621	0.009866	153
rnd-1_family-82	19793	0.268358	174
rnd-1_family-80	6065	0.082231	174
rnd-1_family-93	4335	0.058775	174
rnd-1_family-65	3970	0.053826	174
rnd-1_family-224	2951	0.040010	174
rnd-1_family-160	18889	0.093361	228
rnd-1_family-10	10109	0.049965	228
rnd-1_family-16	9541	0.047158	228
rnd-1_family-14	9344	0.046184	228
rnd-4_family-669	9054	0.044750	228
rnd-1_family-0	294091	0.684765	389
rnd-4_family-95	38821	0.090391	389
rnd-3_family-190	9012	0.020984	389
rnd-4_family-1308	5945	0.013842	389
rnd-1_family-25	4430	0.010315	389

Table S6.4.

Numbers of transposase-related Pfam domains in MAC vs. MIC-limited sequences (IESs) for different ciliate species, based on hmmscan search of six-frame translations (6ft), six-frame translations split on stop codons (6ft split, shown in Figure 4E), or predicted coding sequences only (cds).

	Blepharisma stoltei				*Paramecium tetraurelia*				*Tetrahymena thermophila*				*Oxytricha trifallax*			
	MAC			MIC	MAC			MIC	MAC			MIC	MAC			MIC
Domain	6ft split	6ft	cds	ies 6ft	6ft split	6ft	cds	ies 6ft	6ft split	6ft	cds	ies 6ft	6ft split	6ft	cds	ies 6ft
DDE_1	1	0	3	14	0	0	0	0	0	0	1	83	0	0	0	0
DDE_3	2	2	6	15	1	3	0	7	0	0	3	868	0	0	2	451
DDE_Tnp_1_7	7	0	9	5	1	0	9	0	3	0	3	42	0	0	0	0
DDE_Tnp_IS1595	2	3	5	3	0	0	0	0	0	0	1	138	1	7	7	28
Exo_endo_phos_2	5	0	12	1	0	0	1	0	0	0	1	5	0	0	0	1
HTH_Tnp_Tc5	1	1	4	22	5	9	12	3	0	1	1	56	0	1	2	1
MULE	3	2	6	7	0	0	0	0	0	0	0	2	2	8	6	0
RVT_1	10	4	27	8	0	0	3	4	0	0	3	38	0	0	0	45
Transposase_mut	0	0	0	1	0	0	0	0	0	0	0	0	2	0	1	0
Dimer_Tnp_hAT	0	0	0	0	0	0	0	0	0	3	1	136	0	0	0	0
HTH_Tnp_1	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0
HTH_Tnp_Tc3_2	0	0	0	0	0	0	0	0	0	0	0	124	0	0	0	0
Transposase_1	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0
Helitron_like_N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Tnp_zf-ribbon_2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
DDE_Tnp_1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	0
DDE_Tnp_1_2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
DDE_Tnp_1_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
DDE_5	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0

Table S6.5.

Summary of RepeatMasker annotations for individual repeat families that were classified by RepeatClassifier. Repeats identified predominantly in IESs are highlighted in bold. RepeatClassifier classifications that appear to be errors or spurious annotations are surrounded in parentheses: family rnd-1_family-283 mostly comprises ubiquitin sequences, whereas rnd-4_family-1389 contains abundant WD40 repeats.

			All copies				Full length copies only			
Repeat family	Class (RepeatClassifier)	Cons len. (bp)	No.	Median copy len. (bp)	Total len. (bp)	No. on IESs	No.	Total len. (bp)	No. on IESs	% div. vs. cons
rnd-1_family-1	TcMar/Tc2	1833	539	91	104802	505	30	54844	30	0.5
rnd-1_family-73	DNA/TcMar-Tc1	1949	28	1640	38098	27	22	36273	22	0.6
rnd-1_family-273	LINE	3618	23	1319	38653	2	6	21708	0	16.9
rnd-1_family-276	LINE/RTE-X	3270	15	723	16197	4	2	6451	1	2.95
rnd-1_family-283	(LTR/Pao)	358	39	339	10514	3	24	8438	0	16.25
rnd-4_family-193	LINE/RTE-X	4628	79	279	35576	36	1	4628	1	9.5
rnd-4_family-1389	(Unknown/Helitron-2)	2108	24	268	11049	0	1	2108	0	5.8

Table S6.6.

Counts of intra- vs. intergenic localization for IESs in different size classes (defined in Table S6.1).

IES size class (peak center bp)	Intergenic	Intragenic	IES size class type	Ratio intra:inter-genic	Fraction pseudo-replicates with higher ratio
65	78	178	periodic	2.282051	0.434
72	851	2300	periodic	2.702703	1.000
82	352	883	periodic	2.508523	0.895
92	331	652	periodic	1.969789	0.013
101	264	549	periodic	2.079545	0.069
110	512	1324	periodic	2.585938	0.995
153	83	199	nonper	2.397590	0.615
174	143	295	nonper	2.062937	0.137
228	257	652	nonper	2.536965	0.913
389	373	767	nonper	2.056300	0.040

6.4. Bibliography

Arnaiz, Olivier, Nathalie Mathy, Céline Baudry, Sophie Malinsky, Jean-Marc Aury, Cyril Denby Wilkes, Olivier Garnier, et al. 2012. "The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences." *PLoS Genetics* 8 (10): e1002984. https://doi.org/10.1371/journal.pgen.1002984.

Baudry, Céline, Sophie Malinsky, Matthieu Restituito, Aurélie Kapusta, Sarah Rosa, Eric Meyer, and Mireille Bétermier. 2009. "PiggyMac, a Domesticated PiggyBac Transposase Involved in Programmed Genome Rearrangements in the Ciliate *Paramecium* Tetraurelia." *Genes & Development* 23 (21): 2478–83. https://doi.org/10.1101/gad.547309.

Bestor, T H. 1990. "DNA Methylation: Evolution of a Bacterial Immune Function into a Regulator of Gene Expression and Genome Structure in Higher Eukaryotes." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 326 (1235): 179–87. https://doi.org/10.1098/rstb.1990.0002.

Bischerour, Julien, Simran Bhullar, Cyril Denby Wilkes, Vinciane Régnier, Nathalie Mathy, Emeline Dubois, Aditi Singh, et al. 2018. "Six Domesticated PiggyBac Transposases Together Carry out Programmed DNA Elimination in *Paramecium*." *ELife* 7 (September). https://doi.org/10.7554/eLife.37927.

Catania, Francesco, Rebecca Rothering, and Valerio Vitali. 2021. "One Cell, Two Gears: Extensive Somatic Genome Plasticity Accompanies High Germline Genome Stability in *Paramecium*." *Genome Biology and Evolution* 13 (12). https://doi.org/10.1093/gbe/evab263.

Chalker, Douglas L, Eric Meyer, and Kazufumi Mochizuki. 2013. "Epigenetics of Ciliates." *Cold Spring Harbor Perspectives in Biology* 5 (12): a017764. https://doi.org/10.1101/cshperspect.a017764.

Cheng, Chao-Yin, Alexander Vogt, Kazufumi Mochizuki, and Meng-Chao Yao. 2010. "A Domesticated PiggyBac Transposase Plays Key Roles in Heterochromatin Dynamics and DNA Cleavage during Programmed DNA Deletion in *Tetrahymena* Thermophila." *Molecular Biology of the Cell* 21 (10): 1753–62. https://doi.org/10.1091/mbc.e09-12-1079.

Cheng, Chao-Yin, Janet M Young, Chih-Yi Gabriela Lin, Ju-Lan Chao, Harmit S Malik, and Meng-Chao Yao. 2016. "The PiggyBac Transposon-Derived Genes TPB1 and TPB6 Mediate Essential Transposon-like Excision during the Developmental Rearrangement of Key Genes in *Tetrahymena Thermophila*." *Genes & Development* 30 (24): 2724–36. https://doi.org/10.1101/gad.290460.116.

Chen, Qiujia, Wentian Luo, Ruth Ann Veach, Alison B Hickman, Matthew H Wilson, and Fred Dyda. 2020. "Structural Basis of Seamless Excision and Specific Targeting by PiggyBac Transposase." *Nature Communications* 11 (1): 3446. https://doi.org/10.1038/s41467-020-17128-1.

Chen, Xiao, John R Bracht, Aaron David Goldman, Egor Dolzhenko, Derek M Clay, Estienne C Swart, David H Perlman, et al. 2014. "The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development." *Cell* 158 (5): 1187–98. https://doi.org/10.1016/j.cell.2014.07.034.

Coyne, Robert S, Maoussi Lhuillier-Akakpo, and Sandra Duharcourt. 2012. "RNA-Guided DNA Rearrangements in Ciliates: Is the Best Genome Defence a Good Offence?" *Biology of the Cell* 104 (6): 309–25. https://doi.org/10.1111/boc.201100057.

Drews, Franziska, Abdulrahman Salhab, Sivarajan Karunanithi, Miriam Cheaib, Martin Jung, Marcel H. Schulz, and Martin Simon. 2021. "Broad Domains of Histone Marks in the Highly Compact *Paramecium* Macronuclear Genome." *BioRxiv*, August. https://doi.org/10.1101/2021.08.05.454756.

Dupeyron, Mathilde, Tobias Baril, Chris Bass, and Alexander Hayward. 2020. "Phylogenetic Analysis of the Tc1/Mariner Superfamily Reveals the Unexplored Diversity of Pogo-like Elements." *Mobile DNA* 11 (June): 21. https://doi.org/10.1186/s13100-020-00212-0.

Fang, Wenwen, Xing Wang, John R Bracht, Mariusz Nowacki, and Laura F Landweber. 2012. "Piwi-Interacting RNAs Protect DNA against Loss during *Oxytricha* Genome Rearrangement." *Cell* 151 (6): 1243–55. https://doi.org/10.1016/j.cell.2012.10.045.

Feng, Lifang, Guangying Wang, Eileen P Hamilton, Jie Xiong, Guanxiong Yan, Kai Chen, Xiao Chen, et al. 2017. "A Germline-Limited PiggyBac Transposase Gene Is Required for Precise Excision in *Tetrahymena* Genome Rearrangement." *Nucleic Acids Research* 45 (16): 9481–9502. https://doi.org/10.1093/nar/gkx652.

Feng, Yi, and Laura F Landweber. 2021. "Transposon Debris in Ciliate Genomes." *PLoS Biology* 19 (8): e3001354. https://doi.org/10.1371/journal.pbio.3001354.

Feschotte, C, and C Mouchès. 2000. "Evidence That a Family of Miniature Inverted-Repeat Transposable Elements (MITEs) from the *Arabidopsis Thaliana* Genome Has Arisen from a Pogo-like DNA Transposon." *Molecular Biology and Evolution* 17 (5): 730–37. https://doi.org/10.1093/oxfordjournals.molbev.a026351.

Feschotte, Cédric, Mark T Osterlund, Ryan Peeler, and Susan R Wessler. 2005. "DNA-Binding Specificity of Rice Mariner-like Transposases and Interactions with Stowaway MITEs." *Nucleic Acids Research* 33 (7): 2153–65. https://doi.org/10.1093/nar/gki509.

Feschotte, Cedric, Xiaoyu Zhang, and Susan R. Wessler. 2002. "Miniature Inverted-Repeat Trasnposable Elements and Their Relationship to Established DNA Transposons." In *Mobile DNA II*, edited by Nancy L. Craig, Robert Craigie, Martin Gellert, and Alan M. Lambowitz, 1147–58. Washington, D.C.: ASM Press.

Fillingham, Jeffrey S, Trine A Thing, Nama Vythilingum, Alex Keuroghlian, Deanna Bruno, G Brian Golding, and Ronald E Pearlman. 2004. "A Non-Long Terminal Repeat Retrotransposon Family Is Restricted to the Germ Line Micronucleus of the Ciliated Protozoan *Tetrahymena Thermophila*." *Eukaryotic Cell* 3 (1): 157–69. https://doi.org/10.1128/EC.3.1.157-169.2004.

Fritz, G. 2000. "Human APE/Ref-1 Protein." *The International Journal of Biochemistry & Cell Biology* 32 (9): 925–29. https://doi.org/10.1016/s1357-2725(00)00045-5.

Gao, Bo, Yali Wang, Mohamed Diaby, Wencheng Zong, Dan Shen, Saisai Wang, Cai Chen, Xiaoyan Wang, and Chengyi Song. 2020. "Evolution of Pogo, a Separate Superfamily of IS630-Tc1-Mariner Transposons, Revealing Recurrent Domestication Events in Vertebrates." *Mobile DNA* 11 (July): 25. https://doi.org/10.1186/s13100-020-00220-0.

Gao, Feng, and Laura A Katz. 2014. "Phylogenomic Analyses Support the Bifurcation of Ciliates into Two Major Clades That Differ in Properties of Nuclear Division." *Molecular Phylogenetics and Evolution* 70 (January): 240–43. https://doi.org/10.1016/j.ympev.2013.10.001.

Giese, Arthur Charles. 1973. *Blepharisma: The Biology of a Light-Sensitive Protozoan*. illustrated ed. Stanford University Press.

Grewal, Shiv I S, and Songtao Jia. 2007. "Heterochromatin Revisited." *Nature Reviews. Genetics* 8 (1): 35–46. https://doi.org/10.1038/nrg2008.

Guérin, Frédéric, Olivier Arnaiz, Nicole Boggetto, Cyril Denby Wilkes, Eric Meyer, Linda Sperling, and Sandra Duharcourt. 2017. "Flow Cytometry Sorting of Nuclei Enables the First Global Characterization of *Paramecium* Germline DNA and Transposable Elements." *BMC Genomics* 18 (1): 327. https://doi.org/10.1186/s12864-017-3713-7.

Guermonprez, Hélène, Céline Loot, and Josep M Casacuberta. 2008. "Different Strategies to Persist: The Pogo-like Lemi1 Transposon Produces Miniature Inverted-Repeat Transposable Elements or Typical Defective Elements in Different Plant Genomes." *Genetics* 180 (1): 83–92. https://doi.org/10.1534/genetics.108.089615.

Hamilton, Eileen P, Aurélie Kapusta, Piroska E Huvos, Shelby L Bidwell, Nikhat Zafar, Haibao Tang, Michalis Hadjithomas, et al. 2016. "Structure of the Germline Genome of *Tetrahymena Thermophila* and Relationship to the Massively Rearranged Somatic Genome." *ELife* 5 (November). https://doi.org/10.7554/eLife.19090.

Han, Jeffrey S. 2010. "Non-Long Terminal Repeat (Non-LTR) Retrotransposons: Mechanisms, Recent Developments, and Unanswered Questions." *Mobile DNA* 1 (1): 15. https://doi.org/10.1186/1759-8753-1-15.

Hartl, D L, A R Lohe, and E R Lozovskaya. 1997. "Modern Thoughts on an Ancyent Marinere: Function, Evolution, Regulation." *Annual Review of Genetics* 31: 337–58. https://doi.org/10.1146/annurev.genet.31.1.337.

Herrick, G, S Cartinhour, D Dawson, D Ang, R Sheets, A Lee, and K Williams. 1985. "Mobile Elements Bounded by C4A4 Telomeric Repeats in *Oxytricha Fallax*." *Cell* 43 (3 Pt 2): 759–68. https://doi.org/10.1016/0092-8674(85)90249-1.

Huff, Jason T, Daniel Zilberman, and Scott W Roy. 2016. "Mechanism for DNA Transposons to Generate Introns on Genomic Scales." *Nature* 538 (7626): 533–36. https://doi.org/10.1038/nature20110.

Jahn, C L, S Z Doktor, J S Frels, J W Jaraczewski, and M F Krikau. 1993. "Structures of the *Euplotes Crassus* Tec1 and Tec2 Elements: Identification of Putative Transposase Coding Regions." *Gene* 133 (1): 71–78. https://doi.org/10.1016/0378-1119(93)90226-s.

Klobutcher, L A, and G Herrick. 1995. "Consensus Inverted Terminal Repeat Sequence of *Paramecium* IESs: Resemblance to Termini of Tc1-Related and *Euplotes* Tec Transposons." *Nucleic Acids Research* 23 (11): 2006–13. https://doi.org/10.1093/nar/23.11.2006.

———. 1997. "Developmental Genome Reorganization in Ciliated Protozoa: The Transposon Link." *Progress in Nucleic Acid Research and Molecular Biology* 56: 1–62. https://doi.org/10.1016/S0079-6603(08)61001-6.

Kubota, T, T Tokoroyama, Y Tsukuda, H Koyama, and A Miyake. 1973. "Isolation and Structure Determination of Blepharismin, a Conjugation Initiating Gamone in the Ciliate *Blepharisma*." *Science* 179 (4071): 400–402. https://doi.org/10.1126/science.179.4071.400.

Le Mouël, Anne, Alain Butler, François Caron, and Eric Meyer. 2003. "Developmentally Regulated Chromosome Fragmentation Linked to Imprecise Elimination of Repeated Sequences in Paramecia." *Eukaryotic Cell* 2 (5): 1076–90. https://doi.org/10.1128/EC.2.5.1076-1090.2003.

Lepère, Gersende, Mariusz Nowacki, Vincent Serrano, Jean-François Gout, Gérard Guglielmi, Sandra Duharcourt, and Eric Meyer. 2009. "Silencing-Associated and Meiosis-Specific Small RNA Pathways in *Paramecium Tetraurelia*." *Nucleic Acids Research* 37 (3): 903–15. https://doi.org/10.1093/nar/gkn1018.

Lhuillier-Akakpo, Maoussi, Andrea Frapporti, Cyril Denby Wilkes, Mélody Matelot, Michel Vervoort, Linda Sperling, and Sandra Duharcourt. 2014. "Local Effect of Enhancer of Zeste-like Reveals Cooperation of Epigenetic and Cis-Acting Determinants for Zygotic Genome Rearrangements." *PLoS Genetics* 10 (9): e1004665. https://doi.org/10.1371/journal.pgen.1004665.

Liu, Yifan, Sean D Taverna, Tara L Muratore, Jeffrey Shabanowitz, Donald F Hunt, and C David Allis. 2007. "RNAi-Dependent H3K27 Methylation Is Required for Heterochromatin Formation and DNA Elimination in *Tetrahymena*." *Genes & Development* 21 (12): 1530–45. https://doi.org/10.1101/gad.1544207.

Lynn, Denis H. 2010. *The Ciliated Protozoa*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-8239-9.

Malone, Colin D, Alissa M Anderson, Jason A Motl, Charles H Rexer, and Douglas L Chalker. 2005. "Germ Line Transcripts Are Processed by a Dicer-like Protein That Is Essential for Developmentally Programmed Genome Rearrangements of *Tetrahymena Thermophila*." *Molecular and Cellular Biology* 25 (20): 9151–64. https://doi.org/10.1128/MCB.25.20.9151-9164.2005.

Mitra, Rupak, Xianghong Li, Aurélie Kapusta, David Mayhew, Robi D Mitra, Cédric Feschotte, and Nancy L Craig. 2013. "Functional Characterization of PiggyBat from the Bat *Myotis Lucifugus* Unveils an Active Mammalian DNA Transposon." *Proceedings of the National Academy of Sciences of the United States of America* 110 (1): 234–39. https://doi.org/10.1073/pnas.1217548110.

Miyake, A, and J Beyer. 1974. "Blepharmone: A Conjugation-Inducing Glycoprotein in the Ciliate *Blepharisma*." *Science* 185 (4151): 621–23. https://doi.org/10.1126/science.185.4151.621.

Miyake, A, V Rivola, and T Harumoto. 1991. "Double Paths of Macronucleus Differentiation at Conjugation in *Blepharisma Japonicum*." *European Journal of Protistology* 27 (2): 178–200. https://doi.org/10.1016/S0932-4739(11)80340-8.

Mochizuki, Kazufumi, Noah A Fine, Toshitaka Fujisawa, and Martin A Gorovsky. 2002. "Analysis of a Piwi-Related Gene Implicates Small RNAs in Genome Rearrangement in *Tetrahymena*." *Cell* 110 (6): 689–99. https://doi.org/10.1016/s0092-8674(02)00909-1.

Mochizuki, Kazufumi, and Martin A Gorovsky. 2005. "A Dicer-like Protein in *Tetrahymena* Has Distinct Functions in Genome Rearrangement, Chromosome Segregation, and Meiotic Prophase." *Genes & Development* 19 (1): 77–89. https://doi.org/10.1101/gad.1265105.

Mochizuki, Kazufumi, and Henriette M Kurth. 2013. "Loading and Pre-Loading Processes Generate a Distinct SiRNA Population in *Tetrahymena*." *Biochemical and Biophysical Research Communications* 436 (3): 497–502. https://doi.org/10.1016/j.bbrc.2013.05.133.

Nowacki, Mariusz, Brian P Higgins, Genevieve M Maquilan, Estienne C Swart, Thomas G Doak, and Laura F Landweber. 2009. "A Functional Role for Transposases in a Large Eukaryotic Genome." *Science* 324 (5929): 935–38. https://doi.org/10.1126/science.1170023.

Prescott, D M, and A F Greslin. 1992. "Scrambled Actin I Gene in the Micronucleus of *Oxytricha Nova*." *Developmental Genetics* 13 (1): 66–74. https://doi.org/10.1002/dvg.1020130111.

Rzeszutek, Iwona, Xyrus X Maurer-Alcalá, and Mariusz Nowacki. 2020. "Programmed Genome Rearrangements in Ciliates." *Cellular and Molecular Life Sciences* 77 (22): 4615–29. https://doi.org/10.1007/s00018-020-03555-2.

Sandoval, Pamela Y, Estienne C Swart, Miroslav Arambasic, and Mariusz Nowacki. 2014. "Functional Diversification of Dicer-like Proteins and Small RNAs Required for Genome Sculpting." *Developmental Cell* 28 (2): 174–88. https://doi.org/10.1016/j.devcel.2013.12.010.

Seah, Brandon K B, and Estienne C Swart. 2021. "BleTIES: Annotation of Natural Genome Editing in Ciliates Using Long Read Sequencing." *Bioinformatics* 37 (21): 3929–31. https://doi.org/10.1093/bioinformatics/btab613.

Sellis, Diamantis, Frédéric Guérin, Olivier Arnaiz, Walker Pett, Emmanuelle Lerat, Nicole Boggetto, Sascha Krenek, et al. 2021. "Massive Colonization of Protein-Coding Exons by Selfish Genetic Elements in *Paramecium* Germline Genomes." *PLoS Biology* 19 (7): e3001309. https://doi.org/10.1371/journal.pbio.3001309.

Singh, Minakshi, Brandon K. B. Seah, Christiane Emmerich, Aditi Singh, Christian Woehle, Bruno Huettel, Adam Byerly, et al. 2021. "The *Blepharisma Stoltei* Macronuclear Genome: Towards the Origins of Whole Genome Reorganization." *BioRxiv*, December. https://doi.org/10.1101/2021.12.14.471607.

Stern, Adi, Leeat Keren, Omri Wurtzel, Gil Amitai, and Rotem Sorek. 2010. "Self-Targeting by CRISPR: Gene Regulation or Autoimmunity?" *Trends in Genetics* 26 (8): 335–40. https://doi.org/10.1016/j.tig.2010.05.008.

Sugiura, M, and T Harumoto. 2001. "Identification, Characterization, and Complete Amino Acid Sequence of the Conjugation-Inducing Glycoprotein (Blepharmone) in the Ciliate *Blepharisma* Japonicum." *Proceedings of the National Academy of Sciences of the United States of America* 98 (25): 14446–51. https://doi.org/10.1073/pnas.221457698.

Swart, Estienne C, John R Bracht, Vincent Magrini, Patrick Minx, Xiao Chen, Yi Zhou, Jaspreet S Khurana, et al. 2013. "The *Oxytricha Trifallax* Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes." *PLoS Biology* 11 (1): e1001473. https://doi.org/10.1371/journal.pbio.1001473.

Swart, Estienne C, Cyril Denby Wilkes, Pamela Y Sandoval, Miroslav Arambasic, Linda Sperling, and Mariusz Nowacki. 2014. "Genome-Wide Analysis of Genetic and Epigenetic Control of Programmed DNA Deletion." *Nucleic Acids Research* 42 (14): 8970–83. https://doi.org/10.1093/nar/gku619.

Taverna, Sean D, Robert S Coyne, and C David Allis. 2002. "Methylation of Histone H3 at Lysine 9 Targets Programmed DNA Elimination in *Tetrahymena*." *Cell* 110 (6): 701–11. https://doi.org/10.1016/s0092-8674(02)00941-8.

Udomkit, A, S Forbes, G Dalgleish, and D J Finnegan. 1995. "BS a Novel LINE-like Element in *Drosophila Melanogaster*." *Nucleic Acids Research* 23 (8): 1354–58. https://doi.org/10.1093/nar/23.8.1354.

Vitali, Valerio, Rebecca Hagen, and Francesco Catania. 2019. "Environmentally Induced Plasticity of Programmed DNA Elimination Boosts Somatic Variability in *Paramecium Tetraurelia*." *Genome Research* 29 (10): 1693–1704. https://doi.org/10.1101/gr.245332.118.

Vogt, Alexander, and Kazufumi Mochizuki. 2013. "A Domesticated PiggyBac Transposase Interacts with Heterochromatin and Catalyzes Reproducible DNA Elimination in *Tetrahymena*." *PLoS Genetics* 9 (12): e1004032. https://doi.org/10.1371/journal.pgen.1004032.

Wang, Shi, Lingling Zhang, Eli Meyer, and Mikhail V Matz. 2010. "Characterization of a Group of MITEs with Unusual Features from Two Coral Genomes." *Plos One* 5 (5): e10700. https://doi.org/10.1371/journal.pone.0010700.

Weiss, Adam. 2015. "Lamarckian Illusions." *Trends in Ecology & Evolution* 30 (10): 566–68. https://doi.org/10.1016/j.tree.2015.08.003.

Wuitschick, Jeffrey D, Jill A Gershan, Andrew J Lochowicz, Shuqiang Li, and Kathleen M Karrer. 2002. "A Novel Family of Mobile Genetic Elements Is Limited to the Germline Genome in *Tetrahymena Thermophila*." *Nucleic Acids Research* 30 (11): 2524–37. https://doi.org/10.1093/nar/30.11.2524.

Yao, Meng-Chao, Patrick Fuller, and Xiaohui Xi. 2003. "Programmed DNA Deletion as an RNA-Guided System of Genome Defense." *Science* 300 (5625): 1581–84. https://doi.org/10.1126/science.1084737.

Yuan, Yao-Wu, and Susan R Wessler. 2011. "The Catalytic Domain of All Eukaryotic Cut-and-Paste Transposase Superfamilies." *Proceedings of the National Academy of Sciences of the United States of America* 108 (19): 7884–89. https://doi.org/10.1073/pnas.1104208108.

Zahler, Alan M, Zachary T Neeb, Athena Lin, and Sol Katzman. 2012. "Mating of the Stichotrichous Ciliate *Oxytricha Trifallax* Induces Production of a Class of 27 Nt Small RNAs Derived from the Parental Macronucleus." *Plos One* 7 (8): e42371. https://doi.org/10.1371/journal.pone.0042371.

Zhou, Wanding, Gangning Liang, Peter L Molloy, and Peter A Jones. 2020. "DNA Methylation Enables Transposable Element-Driven Genome Expansion." *Proceedings of the National Academy of Sciences of the United States of America* 117 (32): 19359–66. https://doi.org/10.1073/pnas.1921719117.

Chapter 7
General Discussion

This section contains excerpts from the BioRxiv preprints DOI: https://doi.org/10.1101/2021.12.14.471607 and https://doi.org/10.1101/2022.05.02.489906. Details of author contributions are listed in Appendix A.2

Blepharisma, a distinctive genus of single-celled ciliates known for the red, light-sensitive pigment, blepharismin, in their sub-pellicular membranes (Giese 1973), and unusual nuclear/developmental biology (Miyake, Rivola, and Harumoto 1991), belongs to the early-diverging Heterotrichea lineage of ciliates. *Blepharisma* has been the subject of studies on photosensitivity, gamone-mediated cell conjugation and nuclear development during sexual reproduction for the greater part of the last century and have continued into the 21st century.

(Friedl, Miyake, and Heckmann 1983; Giese 1973; Larsen 1983; Kubota et al. 1973; Miyake 1982; Kovaleva, Raikov, and Miyake 1997; Miyake and Bleyman 1976; Honda and Miyake 1976; Sugiura, Yamanaka, and Suzaki 2016; Sugiura et al. 2005; Terazima, Harumoto, and Tsunoda 2013; Miyake, Rivola, and Harumoto 1991; Sugiura and Harumoto 2001; Sugiura et al. 2010; Terazima and Harumoto 2004; Stolte 1924). Conjugation in *Blepharisma* can be synchronized by pre-treatment with gamones (Miyake, 1968). *Blepharisma* is one of only two ciliate genera, where gamone-mediated cell pairing has been observed (Miyake, 1978) and the only genus for which one of these gamones, a small-molecule derivative of tryptophan (Kubota et al., 1973), has been chemically synthesized (Tokoroyama et al., 1973).

The process of genome rearrangement in ciliates is unique among eukaryotes and has been studied in model ciliates such as the oligohymenophorean ciliates *Paramecium* and *Tetrahymena* and the spirotrichous ciliate *Oxytricha*, which are on evolutionarily divergent branches of the ciliate phylogenetic tree (Vogt et al. 2013). Investigations of the molecular mechanisms and participants of genome reorganization in *Blepharisma* therefore provide the opportunity to investigate what might be the closest approximation to the ancestral state of this process in ciliates.

This study has generated several resources for the scientific community which can inform and enable further studies of *Blepharisma,* among them: i. highly contiguous assemblies for the MAC and the MAC+IES genomes of *Blepharisma* (Chapter 5, Chapter 6) generated from high-molecular weight DNA obtained from enriched *Blepharisma* somatic and germline nuclei and sequenced using both short-read and long-read sequencing (Chapter 3), ii. structural and functional annotations of the MAC and the MAC+IES genomes of *Blepharisma* (Chapter 3), iii. RNA-seq data across multiple consecutive timepoints during sexual reproduction through conjugation, complementing the gene annotations to indicate levels of gene expression during the time course (Chapter 4), iv. small RNA-seq data, complementing the gene and IES

annotations to indicate levels of sRNA expression during the time course and their corresponding loci of origin (Chapter 6). Furthermore, the analysis of PiggyBac homologs encoded in the *Blepharisma* MAC genome (Chapter 5) reveals several similarities in the mode of IES through PiggyBac homologs between *Blepharisma* and the model ciliate *Paramecium*, where this mechanism has been studied over the past few decades in great genomic and molecular detail (Baudry et al. 2009; Dubois et al. 2017; Bischerour et al. 2018; Duharcourt and Betermier 2014; Aury et al. 2006; Arnaiz et al. 2012).

7.1. The macronuclear and micronuclear genomes of *Blepharisma* are structurally and functionally annotated

Reliable genome annotation was made possible by the high completeness and contiguity of the MAC assembly. *Ab initio* genome annotation was performed by generating gene prediction models specifically for *Blepharisma* using the eukaryotic gene prediction software AUGUSTUS. The gene prediction models were complemented by RNA-seq data for accurately annotating *Blepharisma*'s miniscule 15-16 bp spliceosomal introns (Chapter 3, Chapter 5). Accurate gene prediction allowed us to locate genes of interest, encoding key protein domains, such as the PiggyBac DDE_Tnp_1_7 domain. Though the MAC genome of *Stentor* was assembled several years ago (Slabodnick et al. 2017), it has been difficult to find PiggyBac domains in the genome by directly annotating the CDS regions, in a manner similar to *Blepharisma*. One homolog of PiggyBac in *Stentor* was detected in its reference MAC genome, which was assembled solely from Illumina reads, as a region split across two open reading frames (Section 4.2.3.). This illustrates both the importance of a contiguous assembly for inferring pertinent genetic information, as well as the pitfalls of automated gene prediction, when performed on fragmented assemblies.

7.2. Expression of genes in the genome reorganization toolkit is upregulated during development of the new MAC

The transcriptomic data obtained at the different stages of conjugation, coupled with the annotated genome of the *Blepharisma* MAC facilitated the gene expression analysis for determining the role of different genes unregulated during our timeframe of interest, namely the duration in which DNA elimination i.e., IES excision occurs in the new, developing somatic nuclei of the conjugating cells (Chapter 4). Homologs of proteins implicated in genome editing were present among the genes most highly differentially upregulated during new MAC

development, notably the Dicer-like and Piwi proteins which are candidate genes responsible for development-specific sRNA biogenesis (Chapter 4). In current models of IES excision, MIC-limited sequence demarcation by deposition of methylation marks on histones occurs in an sRNA-dependent process (Chalker, Meyer, and Mochizuki 2013). These sequences are recognized by domesticated transposases whose excision is supported by additional proteins that somehow recognize these marks (Chalker, Meyer, and Mochizuki 2013). Together with MIC sequencing we observed abundant, development-specific sRNA production in *Blepharisma* resembling other model ciliates (Chapter 6).

7.3. Multiple transposase families and a putative IES excisase in Blepharisma

The annotated somatic and germline genomes, together with transcriptomic data allowed us to probe one of the key aspects of *Blepharisma* genome reorganization: the nature of the machinery responsible for IES excision in *Blepharisma*. *Blepharisma* has multiple transposase domains encoded in its somatic and germline genomes (Chapter 5, Figure 5.5A; Chapter 6, Figure 6.4E), all of which are upregulated to various degrees during MAC development (Chapter 4, Figures 4.6,4.7,4.8 and 4.9). Several MAC-encoded transposases with DDE_Tnp_1_7, DDE_3, DDE_Tnp_IS1595 (PFAM PF12762) and MULE (PFAM PF10551) protein domains, with the exception of the ones with the DDE_1 (PF03184) domain, have complete catalytic triads, constituted by the DDD/E motif. Moreover, these transposases appear to be domesticated in the somatic genome, as they lack TIRs or other flanking terminal repeats, which might indicate the presence of a transposon sequence environment (Volff 2006). It is therefore, not immediately clear, which one of the MAC encoded transposases might be involved in IES excision, based on the presence of a catalytic triad or expression patterns alone.

In the model ciliates *Paramecium* and *Tetrahymena*, the main IES excisase is a domesticated transposase of the PiggyBac family encoded in the MAC genome and is expressed at high levels exclusively during development of the new MAC (Cheng et al. 2010; Arnaiz et al. 2010). The domesticated PiggyBac transposase in *Paramecium* is called PiggyMac (Pgm) (Baudry et al. 2009) and that of *Tetrahymena* is called Tpb2 (*Tetrahymena* PiggyBac 2)(Cheng et al. 2010). Silencing their respective PiggyBac homologs through RNAi leads to massive retention of IESs in the developing MAC and leads to cell mortality (Arnaiz et al. 2010; Cheng et al. 2010), demonstrating these PiggyBac excisases are the main effectors of IES elimination. In *Oxytricha*, IES elimination is proposed to be dependent on the self-excision of TBE transposons from the

germline genome, also exclusively expressed during MAC development. Silencing of TBEs leads to widespread IES retention and cell death (Nowacki et al. 2009), indicating their similarly critical role in IES elimination in *Oxytricha*. The oligohymenophorean PiggyBac transposases are characterized by the DDE_Tnp_1_7 (PF13843) protein domain, and the *Oxytricha* TBEs encode a transposase with the DDE_3 (PF13358) protein domain. While other transposase domains are found in the somatic and germline genomes of *Paramecium*, *Tetrahymena* and *Oxytricha* (Chapter 6, Figure 6.4E), investigations of their possible involvement or otherwise in IES excision have yet to be reported.

7.4. *Blepharisma* IESs share several characteristics with *Paramecium* IESs

The sequencing of the *Blepharisma* germline nuclei and assembly of the IES regions allowed characterization of the IES lengths and their boundaries (Chapter 6). Most IESs in *Blepharisma* possess a 5'-TA-3' dinucleotide boundary on both ends of the IES, only one copy of which remains in the MAC after the IES has been excised. This TA-dinucleotide boundary is also present in the majority of IESs in *Paramecium*, *Tetrahymena*, *Oxytricha* and *Euplotes* (Klobutcher and Herrick 1995; Steele et al. 1994; Arnaiz et al. 2012; Hamilton et al. 2016; Chen et al. 2014). In *Paramecium*, this dinucleotide repeat is part of the longer consensus sequence 5'-TAYAGYNR-3', which bears resemblance to the boundaries of the *Euplotes* Tec elements and the terminal inverted repeats (TIR) of Tc1/mariner transposons (Klobutcher and Herrick 1995) but is not observed in the TA-bound IES of *Tetrahymena* (Hamilton et al. 2016) or *Oxytricha* (Chen et al. 2014).

The similarity of the *Paramecium* and *Euplotes* IES boundary sequences to Tc1/mariner transposon TIRs was one of the first indications that IESs may be descendants of transposons (Mayer, Mikami, and Forney 1998; Mayer and Forney 1999; Klobutcher and Herrick 1997). The TA-delineated IES in *Paramecium* and *Tetrahymena* are excised by their principal excisases, Pgm and Tpb2 respectively. The majority of *Blepharisma* IESs have conspicuous 5'-TA-3' dinucleotide boundaries, which viewed in the light of the presence of the domesticated PiggyBac homologs in the somatic genome indicate that the *Blepharisma* PiggyMac (BPgm) may be involved in IES excision. The *Blepharisma* IESs also show a periodic distribution of IES lengths (Chapter 6, Figure 6.1A), very similar to the periodicity in IES lengths (Arnaiz et al. 2012) seen in *Paramecium* IESs. IESs in *Tetrahymena*, while also possessing weakly delineated TA-dinucleotide boundaries and being excised by a PiggyBac homolog, do not show such periodicity

in the length distribution. In contrast, they are distributed unimodally, with a peak at ~3 kbp (Hamilton et al. 2016). However, IESs in *Tetrahymena* are excised imprecisely and are found predominantly in intergenic regions (Hamilton et al. 2016), in contrast to those of *Blepharisma* (Chapter 6) and *Paramecium* (Arnaiz et al. 2012), which are intragenically located and precisely excised.

Intragenic IESs, excised precisely, with periodicity in length distribution, possessing TA-boundaries, complemented with the presence of a catalytically complete PiggyMac in the somatic genome, and therefore strongly support the role of the *Blepharisma* PiggyMac as the main IES excisase, just as the *Paramecium* Pgm has been observed to serve this purpose in the only other ciliate where this particular combination IES properties have been observed (Arnaiz et al. 2012; Baudry et al. 2009). In addition to this, *Blepharisma* also mirrors the presence of several catalytically inactive PiggyBac homologs in the somatic genome, which are co-expressed with the putative main *Blepharisma* PiggyMac. In *Paramecium,* six-domesticated PiggyBacs, of which only one is catalytically complete, coordinate IES excision (Bischerour et al. 2018). The catalytically inactive homologs of PiggyMac have been characterized as PiggyMac-likes (PgmLs), and their gene silencing has been demonstrated to cause abnormalities in IES excision in *Paramecium* (Bischerour et al. 2018).

A family of 24 nucleotide sRNA mapping to IES regions, which increases in abundance during development of the new MAC was also found in *Blepharisma* (Chapter 6). A similar enrichment of sRNAs complementary to IES- and TE-regions occurs in both *Tetrahymena* and *Paramecium* in the later stages of sexual reproduction (Sandoval et al. 2014; Schoeberl et al. 2012). This adds another parallel which can be drawn between the PiggyBac-mediated IES-excision machinery of *Paramecium* and *Tetrahymena* to that of *Blepharisma.* This also contrasts to the so-called macRNAs of *Oxytricha* which match to genomic regions between IESs rather than to IESs (Chen et al. 2014; Zahler et al. 2012). All of these observations lend further credence to the likelihood of the main IES excisase in *Blepharisma* being a PiggyBac transposase.

7.5. The last common ancestor of ciliates possessed a PiggyBac

In addition to the BPgm and putative *Blepharisma* PgmLs found in the somatic genome, there are five additional PiggyBac homologs encoded in the IES, called the *Blepharisma* PiggyMics. Two of the five PiggyMics are upregulated during MAC development, though not to the same levels as their MAC-encoded counterparts (Chapter 5, Figure 5.5B). None of the

PiggyMics have a complete catalytic triad, though PiggyMic1 comes close by having an Aspartate residue translocated one position downstream of the canonical site for a complete catalytic triad. The presence of germline-limited PiggyBac homologs is also seen in *Tetrahymena*, where Tpb6, a MIC-limited PiggyBac contributes to the precise excision of a class of intragenic IES during genome reorganization (Feng et al. 2017).

Paramecium and *Tetrahymena* are oligohymenophorean ciliates, and though they are some evolutionary distance apart within the clade of oligohymneophoreans (Gao et al. 2016), they are close enough to warrant the assumption, that if a number of *Paramecium* species and *Tetrahymena* have PiggyBac excisases (Bischerour et al. 2018; Cheng et al. 2010), the last common ancestor of oligohymenophorean ciliates might too. For another oligohymenophorean ciliate, a close relative of *Tetrahymena*, the marine parasitic ciliate *Ichthyopthirus multifilis* no PiggyBac homolog could be found (Chapter 4). That does not mean that such domains are not present in *Ichthyopthirus*. It implies only that if there are any, they may be undetectable due to the incompleteness or fragmented nature of the genome assembly (Chapter 5) or that they do not have sufficient homology to the known ciliate PiggyBacs to be detected.

A phylogenetic tree of PiggyBac homologs found in five different eukaryotic lineages, namely the opisthokonts, the archaeplastids, the atramenopiles and one amoebozoan, in addition to all the ciliate PiggyBacs known from multiple *Paramecium* species, *Tetrahymena*, *Condylostoma* and *Blepharisma* showed that not only do all the heterotrichous PiggyBac homologs (*Condylostoma* and *Paramecium*) share a common ancestor, but also that all the ciliate PiggyBac homologs share a common ancestor (Chapter 5, Figure 5.6). This indicates that the last common ancestor of the Heterotrichs and the Oligohymneophoreans, which would possibly be the last common ancestor of all ciliates may have possessed a PiggyBac. Since PiggyBac domains remain undetectable among the spirotrichous ciliates and other oligohymenophorean ciliates, the shared ancestry of the ciliate PiggyBacs raises the question: were PiggyBac homologs lost in other ciliate lineages or were the heterotrichous and oligohymenophorean PiggyBac homologs acquired independently from a common source? The former possibility is more parsimonious; however, the latter cannot be ruled out, due to the sparse sampling and limited availability of complete and annotated genome assemblies of non-model ciliates.

7.6. The *Blepharisma* germline genome indicates the early origin of IESs

The origin of IESs from transposons was initially postulated to take into account the IESs of *Oxytricha* and *Euplotes* and their possible relation to TBEs and Tec elements, respectively, which are autonomous transposons, encoding DDE_3 domain transposases (Hunter et al. 1989; Jahn et al. 1993). The terminal repeats flanking these IES-transposons also shared similarity with the IES boundaries of *Paramecium* (Klobutcher and Herrick 1995).

The *Blepharisma* germline contains two classes of transposon-derived IESs. One of the IES-limited repeat families, the rnd-1_family-73, bears resemblance to the Tc1/mariner transposons (BstTc1), and possesses TIRs which indicate that it may still be functional as an autonomous transposon. The BstTc1 elements encode a DDE_3 domain transposase. The other IES-limited repeat family, the rnd-1_family_1, possess flanking TIRs and TSDs, and encode DDE_1 and HTH_Tnp_Tc5 domains. This domain architecture is characteristic of Pogo transposases, and these instances in the *Blepharisma* germline have been named Bogo (*Blepharisma* Pogo) elements. In addition to the full-length copies of the Bogo elements, Bogo elements devoid of the transposase elements are also found in the *Blepharisma* germline. These elements represent a miniature inverted transposable element (MITE) form of the Bogo transposons (BogoMITE). Several DDE transposon families are known to generate MITES including PiggyBac (Wang et al. 2010), Tc1/Mariner, Pif/Harbinger, hAT and Mutator (Venkatesh and Nandini 2020; Fattash et al. 2013).

7.7. MITIES in *Blepharisma* represent an intermediate stage in IES generation and transposon domestication

The presence of both BogoMITEs and full-length Bogo elements in the *Blepharisma* germline present the first glimpse of an intermediate state of transposon domestication and IES generation according to the IBAF model (Klobutcher and Herrick 1997). This led to the formulation of the Invasion-Bloom-Abdication-Fade (IBAF) model of IES generation (Klobutcher and Herrick 1997), according to which an invading transposon ("Invasion") would eventually give rise to a class of IESs by first proliferating in the genome ("Bloom"), losing its transposase to capture by the host ("Abdication"), which would free it from purifying selection, and finally degenerating into sequences which bear resemblance to transposon, but are no longer

active TEs ("Fade"). Non-autonomous Tc1/mariner transposon-derived sequences are found in the *Paramecium* germline, namely the Thon, Sardine and Anchois elements (Arnaiz et al. 2012).

MITEs represent the product of the "Abdication" stage, where the transposase of the original TE has been lost, but transposase-less copies, MITEs, can still proliferate in the genome with the help of the full-length transposon, diluting the replication of the TE itself. MITEs have been identified in the germline genome of *Paramecium* as well, notably those of the Thon and Merou families, however these occur in relatively small numbers (Sellis et al. 2021). For additional families of mobile IESs which resemble MITES and have abundant copies (Sellis et al. 2021), the autonomous TEs corresponding were not reported. The *Blepharisma* germline presents the sole instance so far of abundant MITEs constituting a class of IESs in the germline genome, together with the autonomous TE which gave rise to them.

The abundance of repetitive and TE-rich sequences in the germline-limited regions of the genome has often been interpreted as the genome defense mechanism, whereby the soma is protected from the harmful effects of the TEs by sequestering them in the germline, by means of IES excision (Drotos et al. 2022). However, given the TE-based origins of the IESs and their excision mechanisms like domesticated transposases or self-excising elements (Klobutcher and Herrick 1997), an alternative explanation not beset by similar teleological implications can be considered. IESs and their effective excision during MAC development is arguably the reason why these TE-derived elements have been tolerated in the ciliate genome. If invading TEs were not sheltered from the selection in the somatic genome, by means of an efficient and precise excision machinery, they would not be tolerated in this genome. Similar non-adaptive considerations have recently also been applied to mechanisms of TE silencing such as DNA-methylation, which by shielding the TEs from selection, allow them to persist in the genome (Zhou et al. 2020).

7.8. Conclusion and outlook

The principal aim of this thesis was to leverage the basal position of *Blepharisma* on the ciliate tree to infer the degree of conservation of IES excision mechanisms found in ciliates and to learn about the state of these processes in the last ciliate common ancestor. The assemblies and transcriptomic data generated enroute to address these inquiries and made available to the ciliate community will continue to enrich the genomic study of *Blepharisma* and the origin of genome rearrangements in ciliates.

The somatic genome of *Blepharisma* shows that its PiggyBac homologs share common ancestry with those of the oligohymenophorean ciliates *Paramecium* and *Tetrahymena*. This implies that the common ancestor of the heterotrichs and oligohymenophoreans, the last ciliate common ancestor also possessed a PiggyBac. The discovery of IESs in *Blepharisma* leads to an important hypothesis that not only were IES present in the last common ciliate ancestor, but also that any individual ciliate lineages on later diverging branches of the ciliate phylogeny which appear to lack IESs, may have lost them independently.

The future study of *Blepharisma* as a model organism provides other interesting opportunities as well. It possesses a secondary pathway of MAC development, where germline nuclei which have undergone neither meiosis nor recombination give rise to the new MAC, albeit only in cultures where selfing occurs frequently (Miyake, Rivola, and Harumoto 1991). Any disturbance in the processes preceding genome rearrangement will also disrupt DNA elimination. Genome rearrangement during sexual reproduction through the conventional pathway is preceded by processes such as meiosis of the germline nuclei, generation of gametic nuclei, fusion of gametic nuclei in karyogamy and generation of the zygotic nucleus, which can finally give rise to the new germline and somatic nuclei of the cell. Studying genome rearrangement occurring through "apomixis", would therefore provide an opportunity to observe these processes independent of the preceding meiotic and recombination stages. Additionally, the mechanism of UGA-translation in *Blepharisma* has not yet been deciphered. Experimental assays in *Blepharisma* suggest that its eRF1, the eukaryotic release factor involved in translation termination, is capable of recognizing all three standard stop (Eliseev et al. 2011). This indicates that potentially some form of readthrough, using near-cognate pairing with the existing tryptophan tRNA might be involved in UGA translation, or an as yet to be discovered cognate tRNA species.

7.9. Bibliography

Arnaiz, Olivier, Jean François Goût, Mireille Bétermier, Khaled Bouhouche, Jean Cohen, Laurent Duret, Aurélie Kapusta, Eric Meyer, and Linda Sperling. 2010. "Gene Expression in a Paleopolyploid: A Transcriptome Resource for the Ciliate *Paramecium* Tetraurelia." *BMC Genomics* 11 (1): 1–13. https://doi.org/10.1186/1471-2164-11-547/TABLES/5.

Arnaiz, Olivier, Nathalie Mathy, Céline Baudry, Sophie Malinsky, Jean-Marc Aury, Cyril Denby Wilkes, Olivier Garnier, et al. 2012. "The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences." Edited by Harmit S. Malik. *PLoS Genetics* 8 (10): e1002984. https://doi.org/10.1371/journal.pgen.1002984.

Aury, Jean-Marc, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M. Porcel, Béatrice Ségurens, et al. 2006. "Global Trends of Whole-Genome Duplications Revealed by the Ciliate *Paramecium* Tetraurelia." *Nature* 444 (7116): 171–78. https://doi.org/10.1038/nature05230.

Baudry, Céline, Sophie Malinsky, Matthieu Restituito, Aurélie Kapusta, Sarah Rosa, Eric Meyer, and Mireille Bétermier. 2009. "PiggyMac, a Domesticated PiggyBac Transposase Involved in Programmed Genome Rearrangements in the Ciliate *Paramecium* Tetraurelia." *Genes & Development* 23 (21): 2478–83. https://doi.org/10.1101/gad.547309.

Bischerour, Julien, Simran Bhullar, Cyril Denby Wilkes, Vinciane Régnier, Nathalie Mathy, Emeline Dubois, Aditi Singh, et al. 2018. "Six Domesticated PiggyBac Transposases Together Carry out Programmed DNA Elimination in *Paramecium*." *ELife* 7 (September): 1–24. https://doi.org/10.7554/eLife.37927.

Chalker, Douglas L., Eric Meyer, and Kazufumi Mochizuki. 2013. "Epigenetics of Ciliates." *Cold Spring Harbor Perspectives in Biology* 5 (12): a017764–a017764. https://doi.org/10.1101/cshperspect.a017764.

Chen, Xiao, John R. R. Bracht, Aaron David David Goldman, Egor Dolzhenko, Derek M. M. Clay, Estienne C. C. Swart, David H. H. Perlman, et al. 2014. "The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development." *Cell* 158 (5): 1187–98. https://www.sciencedirect.com/science/article/pii/S0092867414009842.

Cheng, Chao-Yin, Alexander Vogt, Kazufumi Mochizuki, and Meng-Chao Yao. 2010. "A Domesticated PiggyBac Transposase Plays Key Roles in Heterochromatin Dynamics and DNA Cleavage during Programmed DNA Deletion in *Tetrahymena* Thermophila." Edited by Kerry S. Bloom. *Molecular Biology of the Cell* 21 (10): 1753–62. https://doi.org/10.1091/mbc.e09-12-1079.

Drotos, Katherine H I, Maxim V Zagoskin, Tony Kess, T Ryan Gregory, and Grace A Wyngaard. 2022. "Throwing Away DNA : Programmed Downsizing in Somatic Nuclei," 1–18.

Dubois, Emeline, Nathalie Mathy, Vinciane Régnier, Julien Bischerour, Céline Baudry, Raphaëlle Trouslard, and Mireille Bétermier. 2017. "Multimerization Properties of PiggyMac, a Domesticated PiggyBac Transposase Involved in Programmed Genome Rearrangements."

Nucleic Acids Research 45 (6): 3204–16. https://doi.org/10.1093/nar/gkw1359.

Duharcourt, Sandra, and Mireille Betermier. 2014. "Programmed Rearrangement in Ciliates: *Paramecium*." *Microbiology Spectrum* 2 (6): 1–20. https://doi.org/10.1128/microbiolspec.MDNA3-0035-2014.

Eliseev, B. D., E. Z. Alkalaeva, P. N. Kryuchkova, S. A. Lekomtsev, Wei Wang, Ai Hua Liang, and L. Yu Frolova. 2011. "Translation Termination Factor ERF1 of the Ciliate *Blepharisma* Japonicum Recognizes All Three Stop Codons." *Molecular Biology* 45 (4): 614–18. https://doi.org/10.1134/S0026893311040030.

Fattash, Isam, Rebecca Rooke, Amy Wong, Caleb Hui, Tina Luu, Priyanka Bhardwaj, and Guojun Yang. 2013. "Miniature Inverted-Repeat Transposable Elements: Discovery, Distribution, and Activity1." *Genome* 56 (9): 475–86. https://doi.org/10.1139/GEN-2012-0174/ASSET/IMAGES/GEN-2012-0174TAB2.GIF.

Feng, Lifang, Guangying Wang, Eileen P. Hamilton, Jie Xiong, Guanxiong Yan, Kai Chen, Xiao Chen, et al. 2017. "A Germline-Limited PiggyBac Transposase Gene Is Required for Precise Excision in *Tetrahymena* Genome Rearrangement." *Nucleic Acids Research* 45 (16): 9481–9502. https://doi.org/10.1093/nar/gkx652.

Friedl, E., Akio Miyake, and K. Heckmann. 1983. "Requirement of Successive Protein Syntheses for the Progress of Meiosis in *Blepharisma*." *Experimental Cell Research* 145 (1): 105–13. https://doi.org/10.1016/S0014-4827(83)80013-5.

Gao, Feng, Alan Warren, Qianqian Zhang, Jun Gong, Miao Miao, Ping Sun, Dapeng Xu, Jie Huang, Zhenzhen Yi, and Weibo Song. 2016. "The All-Data-Based Evolutionary Hypothesis of Ciliated Protists with a Revised Classification of the Phylum Ciliophora (Eukaryota, Alveolata)." *Scientific Reports* 6 (1): 24874. https://doi.org/10.1038/srep24874.

Giese, Arthur C. 1973. *Blepharisma: The Biology of a Light-Sensitive Protozoan*. Stanford University Press. https://books.google.de/books?id=5S6sAAAAIAAJ.

Hamilton, Eileen P, Aurélie Kapusta, Piroska E Huvos, Shelby L Bidwell, Nikhat Zafar, Haibao Tang, Michalis Hadjithomas, et al. 2016. "Structure of the Germline Genome of *Tetrahymena* Thermophila and Relationship to the Massively Rearranged Somatic Genome." *ELife* 5 (November). https://doi.org/10.7554/elife.19090.

Honda, Hisao, and Akio Miyake. 1976. "Cell-to-Cell Contact by Locally Differentiated Surfaces in Conjugation of *Blepharisma*." *Developmental Biology* 52 (2): 221–30. https://doi.org/10.1016/0012-1606(76)90242-6.

Hunter, D. J., K. Williams, S. Cartinhour, and G. Herrick. 1989. "Precise Excision of Telomere-Bearing Transposons during *Oxytricha* Fallax Macronuclear Development." *Genes & Development* 3 (12b): 2101–12. https://doi.org/10.1101/gad.3.12b.2101.

Jahn, Carolyn L., Stella Z. Doktor, John S. Frels, John W. Jaraczewski, and Mark F. Krikau. 1993. "Structures of the Euplotes CrassusTec1 and Tec2 Elements: Identification of Putative Transposase Coding Regions." *Gene* 133 (1): 71–78. https://doi.org/10.1016/0378-1119(93)90226-S.

Klobutcher, Lawrence A., and Glenn Herrick. 1997. "Developmental Genome Reorganization in

Ciliated Protozoa: The Transposon Link." *Progress in Nucleic Acid Research and Molecular Biology* 56 (April): 1–62. https://doi.org/10.1016/s0079-6603(08)61001-6.

Klobutcher, Lawrence A, and Glenn Herrick. 1995. "Consensus Inverted Terminal Repeat Sequence of Pammecium LESs: Resemblance to Termini of Tc1-Related and Euplotes Tec Transposons." *Nucleic Acids Research* 23 (11). https://academic.oup.com/nar/article/23/11/2006/2400577.

Kovaleva, Valentina G., Igor B. Raikov, and Akio Miyake. 1997. "Fine Structure of Conjugation of the Ciliate *Blepharisma* Japonicum I. Changes of the Old Macronucleus." *Archiv Für Protistenkunde* 148 (4): 343–50. https://doi.org/10.1016/S0003-9365(97)80014-0.

Kubota, T., T. Tokoroyama, Y. Tsukuda, H. Koyama, and Akio Miyake. 1973. "Isolation and Structure Determination of Blepharismin, a Conjugation Initiating Gamone in the Ciliate *Blepharisma*." *Science* 179 (4071): 400–402. https://doi.org/10.1126/science.179.4071.400.

Larsen, Hans Find. 1983. "Observations on the Morphology and Ecology of *Blepharisma* Lateritium (Ehrenberg, 1831) Kahl, 1932." *Archiv Fur Protistenkunde* 127 (1): 65–80. https://doi.org/10.1016/S0003-9365(83)80006-2.

Mayer, Kimberly M., and James D. Forney. 1999. "A Mutation in the Flanking 5'-TA-3' Dinucleotide Prevents Excision of an Internal Eliminated Sequence From the *Paramecium* Tetraurelia Genome." *Genetics* 151 (2): 597–604. https://doi.org/10.1093/GENETICS/151.2.597.

Mayer, Kimberly M., Kazuyuki Mikami, and James D. Forney. 1998. "A Mutation in *Paramecium* Tetraurelia Reveals Functional and Structural Features of Developmentally Excised DNA Elements." *Genetics* 148 (1): 139–49. https://doi.org/10.1093/GENETICS/148.1.139.

Miyake, Akio. 1982. "Conjugation of Ciliates in Biochemistry of Multicellular Morphogenesis." In *Biochemistry of Differentiation and Morphogenesis*, 211–30. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-68833-1_21.

Miyake, Akio, and Lea K. Bleyman. 1976. "Gamones and Mating Types in the Genus *Blepharisma* and Their Possible Taxonomic Application." *Genetical Research* 27 (2): 267–75. https://doi.org/10.1017/S0016672300016451.

Miyake, Akio, Valeria Rivola, and Terue Harumoto. 1991. "Double Paths of Macronucleus Differentiation at Conjugation in *Blepharisma* Japonicum." *European Journal of Protistology* 27 (2): 178–200. https://doi.org/10.1016/S0932-4739(11)80340-8.

Nowacki, Mariusz, Brian P. Higgins, Genevieve M. Maquilan, Estienne C. Swart, Thomas G. Doak, and Laura F. Landweber. 2009. "A Functional Role for Transposases in a Large Eukaryotic Genome." *Science* 324 (5929): 935–38. https://doi.org/10.1126/science.1170023.

Sandoval, Pamela Y., Estienne C. Swart, Miroslav Arambasic, and Mariusz Nowacki. 2014. "Functional Diversification of Dicer-like Proteins and Small RNAs Required for Genome Sculpting." *Developmental Cell* 28 (2): 174–88. https://doi.org/10.1016/j.devcel.2013.12.010.

Schoeberl, Ursula E., Henriette M. Kurth, Tomoko Noto, and Kazufumi Mochizuki. 2012. "Biased Transcription and Selective Degradation of Small RNAs Shape the Pattern of

DNA Elimination in *Tetrahymena*." *Genes & Development* 26 (15): 1729–42. https://doi.org/10.1101/GAD.196493.112.

Sellis, Diamantis, Frédéric Guérin, Olivier Arnaiz, Walker Pett, Emmanuelle Lerat, Nicole Boggetto, Sascha Krenek, et al. 2021. *Massive Colonization of Protein-Coding Exons by Selfish Genetic Elements in Paramecium Germline Genomes*. *PLoS Biology*. Vol. 19. https://doi.org/10.1371/journal.pbio.3001309.

Slabodnick, Mark M., J. Graham Ruby, Sarah B. Reiff, Estienne C. Swart, Sager Gosai, Sudhakaran Prabakaran, Ewa Witkowska, et al. 2017. "The Macronuclear Genome of *Stentor Coeruleus* Reveals Tiny Introns in a Giant Cell." *Current Biology* 27 (4): 569–75. https://doi.org/10.1016/j.cub.2016.12.057.

Steele, Charlotte Joy, Genevieve A. Barkocy-Gallagher, L B Preer, and John R. Preer. 1994. "Developmentally Excised Sequences in Micronuclear DNA of *Paramecium*." *Proceedings of the National Academy of Sciences* 91 (6): 2255–59. https://doi.org/10.1073/pnas.91.6.2255.

Stolte, H. A. 1924. "Morphologische Und Physiologische Untersuchungen an *Blepharisma* Undulans. Stein." *Archiv Für Protistenkunde* 48: 245–301.

Sugiura, Mayumi, and Terue Harumoto. 2001. "Identification, Characterization, and Complete Amino Acid Sequence of the Conjugation-Inducing Glycoprotein (Blepharmone) in the Ciliate *Blepharisma* Japonicum." *Proceedings of the National Academy of Sciences* 98 (25): 14446–51. https://doi.org/10.1073/pnas.221457698.

Sugiura, Mayumi, Seiko Kawahara, Hideo Iio, and Terue Harumoto. 2005. "Developmentally and Environmentally Regulated Expression of Gamone 1: The Trigger Molecule for Sexual Reproduction in *Blepharisma* Japonicum." *Journal of Cell Science* 118 (12): 2735–41. https://doi.org/10.1242/jcs.02359.

Sugiura, Mayumi, Hiromi Shiotani, Toshinobu Suzaki, and Terue Harumoto. 2010. "Behavioural Changes Induced by the Conjugation-Inducing Pheromones, Gamone 1 and 2, in the Ciliate *Blepharisma* Japonicum." *European Journal of Protistology* 46 (2): 143–49. https://doi.org/10.1016/j.ejop.2010.01.002.

Sugiura, Mayumi, Mika Yamanaka, and Toshinobu Suzaki. 2016. "Rapid Response to Nutrient Depletion on the Expression of Mating Pheromone , Gamone 1 , in *Blepharisma* Japonicum" 49 (1): 27–36.

Terazima, Masayo Noda, and Terue Harumoto. 2004. "Defense Function of Pigment Granules in the Ciliate *Blepharisma* Japonicum against Two Predatory Protists, Amoeba Proteus (Rhizopodea) and Climacostomum Virens (Ciliata)." *Zoological Science* 21 (8): 823–28. https://doi.org/10.2108/zsj.21.823.

Terazima, Masayo Noda, Terue Harumoto, and Kasumi Tsunoda. 2013. "Mitochondria Toxicity of Blepharismin, a Defense Toxin Produced by Ciliated Protozoan *Blepharisma* Japonicum against Predatory Protists." *Japanese Journal of Protozoology*.

Venkatesh, and B. Nandini. 2020. "Miniature Inverted-Repeat Transposable Elements (MITEs), Derived Insertional Polymorphism as a Tool of Marker Systems for Molecular Plant Breeding." *Molecular Biology Reports 2020 47:4* 47 (4): 3155–67. https://doi.org/10.1007/S11033-020-05365-Y.

Vogt, Alexander, Aaron David Goldman, Kazufumi Mochizuki, and Laura F. Landweber. 2013. "Transposon Domestication versus Mutualism in Ciliate Genome Rearrangements." *PLoS Genetics*. Public Library of Science. https://doi.org/10.1371/journal.pgen.1003659.

Volff, Jean Nicolas. 2006. "Turning Junk into Gold: Domestication of Transposable Elements and the Creation of New Genes in Eukaryotes." *BioEssays* 28 (9): 913–22. https://doi.org/10.1002/bies.20452.

Wang, Shi, Lingling Zhang, Eli Meyer, and Mikhail V. Matz. 2010. "Characterization of a Group of MITEs with Unusual Features from Two Coral Genomes." *PLOS ONE* 5 (5): e10700. https://doi.org/10.1371/JOURNAL.PONE.0010700.

Zahler, Alan M, Zachary T Neeb, Athena Lin, and Sol Katzman. 2012. "Mating of the Stichotrichous Ciliate *Oxytricha* Trifallax Induces Production of a Class of 27 Nt Small RNAs Derived from the Parental Macronucleus." https://doi.org/10.1371/journal.pone.0042371.

Zhou, Wanding, Gangning Liang, Peter L. Molloy, and Peter A. Jones. 2020. "DNA Methylation Enables Transposable Element-Driven Genome Expansion." *Proceedings of the National Academy of Sciences of the United States of America* 117 (32): 19359–66. https://doi.org/10.1073/PNAS.1921719117/-/DCSUPPLEMENTAL.

Chapter 8

Materials and methods

This section contains excerpts from the BioRxiv preprints DOI: https://doi.org/10.1101/2021.12.14.471607 and https://doi.org/10.1101/2022.05.02.489906. Details of author contributions are listed in Appendix A.2

General reagents were analytical grade and purchased from Sigma-Aldrich or Merck unless otherwise indicated.

8.1. Strains and localities

The strains used and their original isolation localities were: *Blepharisma stoltei* ATCC 30299, Lake Federsee, Germany (Repak 1968); *Blepharisma stoltei* HT-IV, Aichi prefecture, Japan; *Blepharisma japonicum* R1072, from an isolate from Bangalore, India (Harumoto et al. 1998).

8.2. Cell cultivation, harvesting and cleanup

For genomic DNA isolation *B. stoltei* ATCC 30299 and HT-IV cells were cultured in Synthetic Medium for *Blepharisma* (SMB) (A Miyake and Beyer 1973) at 27˚C. Belpharismas were fed *Chlorogonium elongatum* grown in Tris-acetate phosphate (TAP) medium (Andersen 2004) at room temperature. *Chlorogonium* cells were pelleted at 1500 g at room temperature for 3 minutes to remove most of the TAP medium, and resuspended in 50 mL SMB. 50 ml of dense *Chlorogonium* was used to feed 1 litre of *Blepharisma* culture once every three days.

Blepharisma stoltei ATCC 30299 and HT-IV cells used for RNA extraction were cultured in Lettuce medium inoculated with *Enterbacter aerogenes* and maintained at 25˚C (A Miyake et al. 1990).

Blepharisma cultures were concentrated by centrifugation in pear-shaped flasks at 100 g for 2 minutes using a Hettich Rotanta 460 centrifuge with swing out buckets. Pelleted cells were washed with SMB and centrifuged again at 100 g for 2 minutes. The washed pellet was then transferred to a cylindrical tube capped with a 100 μm-pore nylon membrane at the base and immersed in SMB to filter residual algal debris from the washed cells. The cells were allowed to diffuse through the membrane overnight into the surrounding medium. The next day, the cylinder with the membrane was carefully removed while attempting to minimize dislodging any debris collected on the membrane. Cell density after harvesting was determined by cell counting under the microscope.

8.3. DNA isolation from whole cells and macronuclei, library preparation and sequencing

B. stoltei macronuclei were isolated by sucrose gradient centrifugation (Lauth et al. 1976). DNA was isolated with a Qiagen 20/G genomic-tip kit according to the manufacturer's instructions. Purified DNA from the isolated MACs was fragmented, size selected and used to prepare libraries according to standard PacBio HiFi SMRTbell protocols. The libraries were sequenced in circular consensus mode to generate HiFi reads.

Total genomic DNA from *B. stoltei* ATCC 30299, *B. stoltei* HT-IV and *B. undulans* was isolated with the SigmaAldrich GenElute Mammalian genomic DNA kit. A sequencing library was prepared with a NEBnext FS DNA Library Prep Kit for Illumina and sequenced on an Illumina HiSeq 3000 sequencer, generating 150 bp paired-end reads. Total genomic DNA form *B. japonicum* was isolated with the SigmaAldrich GenElute Mammalian genomic DNA kit and and sequencing library was prepared with the TruSeq Nano DNA Library Prep Kit (Illumina) and sequenced on an Illumina NovaSeq6000 to generate 150 bp paired-end reads.

Total genomic DNA from *B. stoltei* ATCC 30299 was also isolated by the phenol-chloroform method (Sambrook and Russell 2006). A sequencing library was prepared with a NEBnext FS DNA Library Prep Kit for Illumina and sequenced on an Illumina HiSeq 3000 sequencer, generating 150 bp paired-end reads.

8.4. *Enrichment of micronuclei, isolation and sequencing of MIC genomic DNA*

B. stoltei ATCC 30299 cells were harvested and cleaned to yield 400 mL of cell suspension (1600 cells/mL). This suspension was twice concentrated by centrifugation (100 g; 2 min; room temperature) in pear-shaped flasks and in 50 mL tubes to ~8 mL. 10 mL chilled Qiagen Buffer C1 (from the Qiagen Genomic DNA Buffer Set, Qiagen no. 19060) and 30 mL chilled, autoclaved deionized water were added. The suspension was mixed by gently inverting the tube until no clumps of cells were visible, and then centrifuged (1300 g; 15 min; 4˚C). The pellet was washed with chilled 2 mL Buffer C1 and 6 mL water, mixed by pipetting gently with a wide-bore pipette tip, centrifuged (1300 g; 15 min; 4˚C), and resuspended with chilled 2 mL Buffer C1 and 6 mL water by pipetting gently with a wide-bore pipette tip.

The nuclei suspension was layered over a discrete sucrose gradient of 20 mL 10% (w/v) sucrose in TSC medium (0.1% (v/v) Triton X-100, 0.01% (w/v) spermidine trihydrochloride

and 5mM $CaCl_2$) on top of 40% (w/v) sucrose in TSC medium (Lauth et al. 1976). Gradients were centrifuged (250 g; 10 min; 4˚C). 10 to 12 mL fractions were collected by careful pipetting from above, and the nuclei were pelleted by centrifugation (3000 g; 10 min; 4˚C). DNA was extracted from pelleted nuclei with the Qiagen Genomic tips 20/G and HMW DNA extraction buffer set (Qiagen no. 19060) according to the manufacturer's instructions. DNA concentration was measured by the Qubit dsDNA High-Sensitivity assay kit. Fragment size distribution in each sample was assessed by a Femto Pulse analyzer.

B. stoltei ATCC 30299 DNA isolated from the MIC-enriched fraction on two separate occasions was used to prepare two sets of DNA sequencing libraries. A low-input PacBio SMRTbell library was prepared without shearing the DNA and was sequenced in the CLR- (continuous long read) sequencing mode on a PacBio Sequel II instrument. Paired-end short-read libraries were prepared for four sucrose gradient fractions (top (T), middle (M), middle lower (ML), bottom (B)) and sequenced with 100 bp BGI-Seq paired-end reads on a BGI-Seq instrument.

8.5. Genome assembly

8.5.1. Chapter 3

Short-read assemblies were assemblies were generated for *B. stoltei* ATCC 30299, *B. stoltei* HT-IV, *B. undulans* and *B. japonicum*, using SPAdes genome assembler (v3.14.0) (sapdes.py –12 <file with interleaved forward and reverse paired-end reads>) –threads 16 -o <output_folder>. All assemblies were inspected with the quality assessment tool QUAST (Gurevich et al. 2013).

Long-read assemblies were generated for *B. stoltei* long-reads using Ra (v.0.9) (ra -x pb -t 16 <input reads >) and Raven (v.1.1.5)(raven -t 16 <input sequences>).

8.5.2. Chapter 5

Two MAC genome assemblies for *B. stoltei* ATCC 30299 (70× and 76× coverage) were produced with Flye (version 2.7-b1585) (Kolmogorov et al. 2019) for the two separate PacBio Sequel II libraries (independent replicates) using default parameters and the switches: --pacbio-hifi -g 45m. The approximate genome assembly size was chosen based on preliminary Illumina genome assemblies of approximately 40 Mb. Additional assemblies using the combined coverage (145×) of the two libraries were produced using either Flye version 2.7-b1585 or 2.8.1-b1676,

and the same parameters. Two rounds of extension and merging were then used, first comparing the 70× and 76× assemblies to each other, then comparing the 145× assembly to the former merged assembly. Assembly graphs were all relatively simple, with few tangles to be resolved (Figure S1). Minimap2 (H. Li 2018) was used for pairwise comparison of the assemblies using the parameters: -x asm5 --frag=yes --secondary=no, and the resultant aligned sequences were visually inspected and manually merged or extended where possible using Geneious (version 2020.1.2) (Kearse et al. 2012).

Visual inspection of read mapping to the combined assembly was then used to trim off contig ends where there was little correspondence between the assembly consensus and the mapped reads - which we classify as "cruft". Read mapping to cruft regions was often lower or uneven, suggestive of repeats. Alternatively, these features could be due to trace MIC sequences, or sites of alternative chromosome breakage during development which lead to sequences that are neither purely MAC nor MIC. A few contigs with similar dubious mapping of reads at internal locations, which were also clear sites of chromosome fragmentation (evident by abundant telomere-bearing reads in the vicinity) were split apart and trimmed back as for the contig ends. Telomere-bearing reads mapped to the non-trimmed region nearest to the trimmed site were then used to define contig ends, adding representative telomeric repeats from one of the underlying sequences mapped to each of the ends. The main genome assembly with gene predictions can be obtained from the European Nucleotide Archive (ENA) (PRJEB40285; accession GCA_905310155). “Cruft” sequences are also available from the same accession.

Two separate assemblies were generated for *Blepharisma japonicum.* A genome assembly for *Blepharisma japonicum* strain R1072 was generated from Illumina reads, using SPAdes genome assembler (v3.14.0) (Prjibelski et al. 2020). An assembly with PacBio Sequel long reads was produced with Ra (v0.2.1) (Vaser and Sikic 2019), which uses the Overlap-Layout-Consensus paradigm. The assembly produced with Ra was more contiguous, with 268 contigs, in comparison to 1510 contigs in the SPAdes assembly, and was chosen as the reference assembly for *Blepharisma japonicum* (ENA accession: ERR6474383).

Condylostoma magnum genomic reads (study accession PRJEB9019) from a previous study (Swart et al. 2016) were reassembled to improve contiguity and remove bacterial contamination. Reads were trimmed with bbduk.sh from the BBmap package v38.22 (https://sourceforge.net/projects/bbmap/), using minimum PHRED quality score 2 (both ends)

and k-mer trimming for Illumina adapters and Phi-X phage sequence (right end), retaining only reads ≥25 bp. Trimmed reads were error-corrected and reassembled with SPAdes v3.13.0 (Prjibelski et al. 2020) using k-mer values 21, 33, 55, 77, 99. To identify potential contaminants, the unassembled reads were screened with phyloFlash v3.3b1 (Gruber-Vodicka, Seah, and Pruesse 2020) against SILVA v132 (Quast et al. 2013); the coding density under the standard genetic code and prokaryotic gene model were also estimated using Prodigal v2.6.3 (Hyatt et al. 2010). Plotting the coverage vs. GC% of the initial assembly showed that most of the likely bacterial contigs (high prokaryotic coding density, lower coverage, presence of bacterial SSU rRNA sequences) had >=40% GC, so we retained only contigs with <40% GC as the final *C. magnum* genome bin. The final assembly is available from the ENA bioproject PRJEB48875 (accession GCA_920105805).

All assemblies were inspected with the quality assessment tool QUAST (Gurevich et al. 2013).

8.6. Gene prediction

We created a wrapper program, Intronarrator, to predict genes in *Blepharisma* and other heterotrichs, accommodating their tiny introns. Intronarrator can be downloaded and installed together with dependencies via Conda from GitHub (https://github.com/Swart-lab/Intronarrator). Intronarrator directly infers introns from spliced RNA-seq reads mapped by HISAT2 from the entire developmental time course we generated. RNA-seq reads densely cover almost the entire *Blepharisma* MAC genome, aside from intergenic regions, and most potential protein-coding genes (Figure 4B). After predicting the introns and removing them to create an intron-minus genome, Intronarrator runs AUGUSTUS (version 3.3.3) using its intronless model. It then adds back the introns to the intronless gene predictions to produce the final gene predictions.

Introns are inferred from "CIGAR" string annotations in mapped RNA-seq BAM files, using the regular expression "[0-9]+M([0-9][0-9])N[0-9]+M" to select spliced reads. For intron inference we only used primary alignments with: MAPQ >= 10; just a single "N", indicating one potential intron, per read; and at least 6 mapped bases flanking both the 5' and 3' intron boundaries (to limit spurious chance matches of a few bases that might otherwise lead to incorrect intron prediction). The most important parameters for Intronarrator are a cut-off of 0.2 for the fraction of spliced reads covering a potential intron, and a minimum of 10 or more

spliced reads to call an intron. The splicing fraction cut-off was chosen based on the overall distribution of splicing (Figure 2A-C). From our visual examination of mapped RNA-seq reads and gene predictions, values less than this were typically "cryptic" excision events (Saudemont et al. 2017) which remove potentially essential protein-coding sequences, rather than genuine introns. Intronarrator classifies an intron as sense (7389 in total, excluding alternative splicing), when the majority of reads (irrespective of splicing) mapping to the intron are the same strand, and antisense (554 in total) when they are not. The most frequently spliced intron was chosen in rare cases of overlapping alternative intron splicing.

To eliminate spurious prediction of protein-coding genes overlapping ncRNA genes, we also incorporated ncRNA prediction in Intronarrator. Infernal (Nawrocki, Kolbe, and Eddy 2009) (default parameters; e-value < 1e-6) was used to predict a restricted set of conserved ncRNAs models (i.e., tRNAs, rRNAs, SRP, and spliceosomal RNAs) from RFAM 14.0 (Kalvari et al. 2018). These ncRNAs were hard-masked (with "N" characters) before AUGUSTUS gene prediction. Both Infernal ncRNA predictions (excluding tRNAs) and tRNA-scan SE 2.0 (Chan et al. 2019) (default parameters) tRNA predictions are annotated in the *B. stoltei* ATCC 30299 assembly deposited in the European Nucleotide Archive.

Since we found that *Blepharisma stoltei*, like *Blepharisma japonicum* (Swart et al. 2016), uses a non-standard genetic code, with UGA codon translated as tryptophan, gene predictions use the "The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (transl_table=4)" from the NCBI genetic codes. The default AUGUSTUS gene prediction parameters override alternative (mitochondrial) start codons permitted by NCBI genetic code 4, other than ATG. So, all predicted *B. stoltei* gene coding sequences begin with ATG.

RNA-seq read mapping relative to gene predictions of Contig_1 of *B. stoltei* ATCC30299 was visualized with PyGenomeTracks (Lopez-Delisle et al. 2021).

8.7. Functional gene annotation

Pannzer2 (Törönen, Medlar, and Holm 2018) (default parameters) and EggNog (version 2.0.1) (Huerta-Cepas et al. 2019) were used for gene annotation. Annotations were combined and are available from the Max Planck Society's Open Research Repository, Edmond (https://dx.doi.org/10.17617/3.8c). Protein domain annotations were performed using hmmscan

from HMMER3 (version 3.3, Nov 2019) (Eddy 2011) vs. the PFAM database (Pfam-A.full, 33.0, retrieved on June 23, 2020) with default parameters.

8.8. Gamone 1/ Cell-Free Fluid (CFF) isolation and conjugation activity assay

B. stoltei ATCC 30299 cells were cultured and harvested and concentrated to a density of 2000 cells/mL according to the procedure described in "Cell cultivation, Harvesting and Cleanup". This concentrated cell culture was incubated overnight at 27˚C. The next day, the cells were harvested, and the supernatant collected and preserved at 4˚C at all times after extraction. The supernatant was then filtered through a 0.22 μm-pore filter. BSA (10 mg/mL) was added to produce the final CFF at a final BSA concentration of 0.01%.

To assess the activity of the CFF, serial dilutions of the CFF were made to obtain the gamone activity in terms of units (U) (AKIO Miyake 1981).The activity of the isolated CFF was 2^{10} U.

8.9. Conjugation time course and RNA isolation for high-throughput sequencing

B. stoltei cells for the complementary strains, ATCC 30299 and HT-IV, were cultivated and harvested by gentle centrifugation to achieve a final cell concentration of 2000 cells/ml for each strain. Non-gamone treated ATCC 30299 (A1) and HT-IV cells (H1) were collected (time point: -3 hours). Strain ATCC 30299 cells were then treated with synthetic gamone 2 (final concentration 1.5 μg/mL) and strain HT-IV cells were treated with cell-free fluid with a gamone 1 activity of ~2^{10} U/ml for three hours (Figure S6).

Homotypic pair formation in both cultures was checked after three hours. More than 75% of the cells in both cultures formed homotypic pairs. At this point the samples A2 (ATCC 30299) and H2 (HT-IV) were independently isolated for RNA extraction as gamone-treated control cells just before mixing. For the rest of the culture, homotypic pairs in both cultures were separated by pipetting them gently with a wide-bore pipette tip. Once all pairs had been separated, the two cultures were mixed together. This constitutes the experiment's 0-h time point. The conjugating culture was observed and samples collected for RNA isolation or cell fixation at 2 h, 6 h, 14 h, 18 h, 22 h, 26 h, 30 h and 38 h (Figure S6). Further details of the sample staging approach are described in (A Miyake, Rivola, and Harumoto 1991) and (Sugiura et al. 2012). At each time point including samples A1, H1, A2 and H2, 7 mL of culture was harvested for RNA-extraction using Trizol. The total RNA obtained was then separated into a

small RNA fraction < 200 nt and a fraction with RNA fragments > 200 nt using the Zymo RNA Clean and Concentrator-5 kit according to the manufacturer's instructions. RNA-seq libraries were prepared by BGI according to their standard protocols and sequenced on a BGISeq 500 instrument.

Separate 2 mL aliquots of cells at each time point for which RNA was extracted were concentrated by centrifuging gently at 100 rcf. 50 μL of the concentrated cells were fixed with Carnoy's fixative (ethanol:acetic acid, 6:1), stained with DAPI and imaged to determine the state of nuclear development (A Miyake, Rivola, and Harumoto 1991).

8.10. RNA-seq read mapping

To permit correct mapping of tiny introns RNA-seq data was mapped to the *B. stoltei* ATCC 30299 MAC genome using a version of HISAT2 (Kim et al. 2019) with modified source code, with the static variable minIntronLen in hisat2.cpp lowered to 9 from 20 (change available in the HISAT2 github fork: https://github.com/Swart-lab/hisat2/; commit hash 86527b9). HISAT2 was run with default parameters and parameters --min-intronlen 9 --max-intronlen 500. It should be noted that RNA-seq from timepoints in which *B. stoltei* ATCC 30299 and *B. stoltei* HT-IV cells were mixed together were only mapped to the former genome assembly, and so reads for up to three alleles may map to each of the genes in this assembly.

8.11. Gene expression analysis

Features from RNA-seq reads mapped to the *B. stoltei* ATCC 30299 MAC and MAC+IES genomes over the developmental time-course were extracted using featureCounts from the Subread package (Liao, Smyth, and Shi 2014). Further analysis was performed using the R software environment. Genes with a total read count of less than 50, across all timepoints, were filtered out of the dataset. The remaining genes were passed as a DGElist object to edgeR (Robinson, McCarthy, and Smyth 2010). Each time point, representing one library, was normalized for library size using the edgeR function calcNormFactors. The normalized read counts were transformed into TPM (transcripts per million) values (B. Li et al. 2010; Wagner, Kin, and Lynch 2012). The TPM-values for different genes were compared across timepoints to examine changes in gene expression. Heatmaps showing log2(TPM) changes across timepoints were plotted using the tidyverse collection of R packages (https://www.tidyverse.org/) and RColorBrewer (https://rdrr.io/cran/RColorBrewer/). Tabulated gene expression estimates

together with protein annotations are available from Edmond (https://dx.doi.org/10.17617/3.8c).

8.12. Repeat annotation

Interspersed repeat element families were predicted with RepeatModeler v2.0.1 (default settings, random number seed 12345) with the following dependencies: rmblast v2.9.0+ (http://www.repeatmasker.org/RMBlast.html), TRF 4.09 (Benson, 1999), RECON (Bao and Eddy, 2002), RepeatScout 1.0.6 (Price et al., 2005), RepeatMasker v4.1.1 (http://www.repeatmasker.org/RMDownload.html). Repeat families were also classified in the pipeline by RepeatClassifier v2.0.1 through comparison against RepeatMasker's repeat protein database and the Dfam database. Consensus sequences of the predicted repeat families, produced by RepeatModeler, were then used to annotate repeats with RepeatMasker, using rmblast as the search engine.

Terminal inverted repeats (TIRs) of selected repeat element families were identified by aligning the consensus sequence from RepeatModeler, and/or selected full-length elements, with their respective reverse complements using MAFFT (Katoh and Standley, 2013) (plugin version distributed with Geneious). TIRs from the Dfam DNA transposon termini signatures database (v1.1, https://www.dfam.org/releases/dna_termini_1.1/dna_termini_1.1.hmm.gz) (Storer et al., 2021) were searched with hmmsearch (HMMer v3.2.1) against the IES sequences, to identify matches to TIR signatures of major transposon subfamilies.

8.13. Cell fixation and imaging

B. stoltei cells were harvested as above ("Cell cultivation"), and fixed with an equal volume of "ZFAE" fixative, containing zinc sulfate (0.25 M, Sigma Aldrich), formalin, glacial acetic acid and ethanol (Carl Roth), freshly prepared by mixing in a ratio of 10:2:2:5. Fixed cells were pelleted (1000 g; 1 min), resuspended in 1% TritonX-100 in PHEM buffer to permeabilize (5 min; room temperature), pelleted and resuspended in 2% (w/v) formaldehyde in PHEM buffer to fix further (10 min; room temp.), then pelleted and washed twice with 3% (w/v) BSA in TBSTEM buffer (~10 min; room temp.). For indirect immunofluorescence, washed cells were incubated with primary antibody rat anti-alpha tubulin (Abcam, ab6161; 1:100 dilution in 3% w/v BSA/TBSTEM; 60 min; room temp.) then secondary antibody goat anti-rat IgG H&L labeled with AlexaFluor 488 (Abcam, ab150157, 1:500 dilution in 3% w/v BSA/TBSTEM; 20

min; room temp.). Nuclei were counterstained with DAPI (1 μg/mL) in 3% (w/v) BSA/TBSTEM. A z-stack of images was acquired using a confocal laser scanning microscope (Leica TCS SP8), equipped with a HC PL APO 40× 1.30 Oil CS2 objective and a 1 photomultiplier tube and 3 HyD detectors, for DAPI (405 nm excitation, 420-470 nm emission) and Alexa Fluor 488 (488 nm excitation, 510-530 nm emission). Scanning was performed in sequential exposure mode. Spatial sampling was achieved according to Nyquist criteria. ImageJ (Fiji) (Schindelin et al. 2012) was used to adjust image contrast and brightness and overlay the DAPI and AlexaFluor 488 channels. The z-stack was temporally color-coded.

8.14. Variant calling

Illumina total genomic DNA-seq libraries for *B. stoltei* strains ATCC 30299 (ENA accession: ERR6061285) and HT-IV (ERR6064674) were mapped to the ATCC 30299 reference assembly with bowtie2 v2.4.2 (Langmead and Salzberg 2012). Alignments were tagged with the MC tag (CIGAR string for mate/next segment) using samtools (Danecek et al. 2021) fixmate. The BAM file was sorted and indexed, read groups were added with bamaddrg (commit 9baba65, https://github.com/ekg/bamaddrg), and duplicate reads were removed with Picard MarkDuplicates v2.25.1 (http://broadinstitute.github.io/picard/). Variants were called from the combined BAM file with freebayes v1.3.2 (Garrison and Marth 2012) in diploid mode, with maximum coverage 1000 (option -g). The resultant VCF file was combined and indexed with bcftools v1.12 (Danecek et al. 2021), then filtered to retain only SNPs with quality score > 20, and at least one alternate allele.

8.15. Annotation of alternative telomere addition sites

Alternative telomere addition sites (ATASs) were annotated by mapping PacBio HiFi reads to the curated reference MAC assembly described above, using minimap2 and the following flags: -x asm20 --secondary=no --MD. We expect reads representing alternative telomere additions to have one portion mapping to the assembly (excluding telomeric regions), with the other portion containing telomeric repeats being soft-clipped in the BAM record. For each mapped read with a soft-clipped segment, we extracted the clipped sequence, and the coordinates and orientation of the clip relative to the reference. We searched for ≥ 24 bp tandem direct repeats of the telomere unit (i.e., ≥3 repeats of the 8 bp unit) in the clipped segment with NCRF v1.01.02 (Harris, Cechova, and Makova 2019), which can detect tandem repeats in the presence of noise, e.g., from sequencing error. The orientation of the telomere sequence, the

distance from the end of the telomeric repeat to the clip junction ('gap'), and the number of telomere-bearing reads vs. total mapped reads at each junction were also recorded. Junctions with zero gap between telomere repeat and clip junction were annotated as ATASs. The above procedure was implemented in the MILTEL module of the software package BleTIES v0.1.3 (Seah and Swart 2021).

MILTEL output was processed with Python scripts depending on Biopython (Cock et al. 2009), pybedtools (Dale, Pedersen, and Quinlan 2011), Bedtools (Quinlan and Hall 2010), and Matplotlib (Hunter 2007), to summarize statistics of junction sequences and telomere permutations at ATAS junctions, and to extract genomic sequences flanking ATASs for sequence logos. Logos were drawn with Weblogo v3.7.5 (Crooks et al. 2004), with sequences oriented such that the telomere would be added on the 5' end of the ATAS junctions.

To calculate the expected minichromosome length, we assumed that ATASs were independent and identically distributed in the genome following a Poisson distribution. About 47×10^3 ATASs were annotated, supported on average by a single read. Given a genome of 42 Mbp at 145× coverage, the expected rate of encountering an ATAS is 47×10^3 / (145 × 42 Mbp), so the distance between ATASs (i.e., the minichromosome length) is exponentially distributed with expectation (145 × 42 Mbp) / 47×10^3 = 130 kbp.

8.16. Genetic code prediction

We used the program PORC (Prediction Of Reassigned Codons; available from https://github.com/Swart-lab/PORC) previously written to predict genetic codes in protist transcriptomes (Swart et al. 2016) to predict the *B. stoltei* genetic code. This program was used to translate the draft *B. stoltei* ATCC 30299 genome assembly in all six frames (with the standard genetic code). Like the program FACIL (Dutilh et al. 2011) that inspired PORC, the frequencies of amino acids in PFAM (version 34.0) protein domain profiles aligned to the six frame translation by HMMER 3.1b2 (Eddy 2011) (default search parameters; domains used for prediction with conditional E-values < 1e-20), and correspondingly also to the underlying codon, are used to infer the most likely amino acid encoded by each codon (Figure 1B).

8.17. Assessment of genome completeness

A BUSCO (version 4.0.2) (Waterhouse et al. 2018) analysis of the assembled MAC genomes of *B. stoltei* and *B. japonicum* was performed on the set of predicted proteins (BUSCO

mode -prot) using the BUSCO Alveolata database. The completeness of the *Blepharisma* genomes was compared to the protein-level BUSCO analysis of the published genome assemblies of ciliates *T. thermophila*, *P. tetraurelia*, *S. coeruleus* and *I. multifiliis* (Figure 11).

8.18. Gene expression analysis

Features from RNA-seq reads mapped to the *B. stoltei* ATCC 30299 MAC and MAC+IES genomes over the developmental time-course were extracted using featureCounts from the Subread package (Liao, Smyth, and Shi 2014). Further analysis was performed using the R software environment. Genes with a total read count of less than 50, across all timepoints, were filtered out of the dataset. The remaining genes were passed as a DGElist object to edgeR (Robinson, McCarthy, and Smyth 2010). Each time point, representing one library, was normalized for library size using the edgeR function calcNormFactors. The normalized read counts were transformed into TPM (transcripts per million) values (B. Li et al. 2010; Wagner, Kin, and Lynch 2012). The TPM-values for different genes were compared across timepoints to examine changes in gene expression. Heatmaps showing log2(TPM) changes across timepoints were plotted using the tidyverse collection of R packages (https://www.tidyverse.org/) and RColorBrewer (https://rdrr.io/cran/RColorBrewer/). Tabulated gene expression estimates together with protein annotations are available from Edmond (https://dx.doi.org/10.17617/3.8c).

8.19. Sequence visualization and analysis

Nucleotide and amino acid sequences were visualized using Geneious Prime (Biomatters Ltd.) (Kearse et al. 2012). Multiple sequence alignments were performed with MAFFT version 7.450 (Katoh et al. 2002; Katoh and Standley 2013). Phylogenetic trees were constructed with PhyML version 3.3.20180621 (Guindon et al. 2010).

8.20. Identification and correction of MIC-encoded PiggyBac homologs

We sought coding regions present within *Blepharisma* IESs to gauge the expression and type of MIC-limited genes (IES assembly and gene prediction described in Seah et al. 2022). After gene prediction within IESs with Intronarrator, predicted protein domains were annotated by HMMER (v3.3) (Eddy 2011). Several transposase families were represented in protein domains identified with coding regions of IESs. However, gene prediction within IESs was hampered by the presence of intermittent A-residues in the consensus sequence which occur due

to the inaccuracy inherent in long-reads, from which the IES regions were assembled. These errors cause IES gene-prediction to falter by generating inaccurate ORFs. To circumvent this, a six-frame translation of the MIC-limited genome regions was performed using a custom script, which was then used to detect PFAM domains, using HMMER and the Pfam-A database 32.0 (release 9) (Mistry et al. 2021). Domain annotations for diagrams were generated with the InterproScan 5.44-79.0 pipeline (Jones et al. 2014)

Four instances of the Pfam domain DDE_Tnp_1_7, characteristic of PiggyBac transposases, were detected in an initial gene prediction within *Blepharisma* IESs. The four genes corresponding to the DDE_Tnp_1_7 domain had high RNA-seq coverage of combined reads from all timepoints across development. The IESs with the PiggyBac domains on Contig 17 and Contig 39 each had two ORFs with a partial DDE_1_7 domain, separated by a few hundred bp. Alignment of short-read MIC-enriched DNA reads mapped to the IES regions containing the putative PiggyBac homologs indicated that several A-nucleotides in the assembled IESs were insertion errors in the IES assembly, which were corrected with the short-read alignment. Open reading frames of predicted genes in these corrected regions were adjusted accordingly. The prefix "cORF" (corrected ORFs) was used to indicate the short-read corrected sequences of the PiggyMics.

Short-read MIC-enriched DNA sequences were aligned to the IES regions containing putative PiggyBac homologs with Hisat2 (2.0.0-beta) with modified source code (described above). Indel errors in the IES assembly were corrected manually, then used to predict coding regions. Pfam domains were annotated on MIC PiggyBac homologs with corrected ORFs using the InterproScan (v. 1.1.4) (Quevillon et al. 2005) plugin in Geneious v11.1.5 (Biomatter Ltd.). DDE_Tnp_1_7 domains were detected in the corrected ORFs, which in some cases spanned IES regions lacking predicted genic regions before correction. A multiple sequence alignment of the correct MIC PiggyBac homologs with other ciliate PiggyBac-derived proteins (PGBDs) and eukaryotic PiggyBac-like elements (PBLEs) that contain the PiggyBac transposase domain DDE_Tnp_1_7 (PF13843) was performed with MAFFT (v4.1) via the Geneious plugin (algorithm L-INS-i, BLOSUM62 scoring matrix, gap open penalty 1.53, offset value 0.123). A phylogenetic tree was constructed using the FastTree (v 2.1.11) plugin for Geneious (Whelan-Goldman model).

8.21. d_N/d_S estimation

We generated pairwise coding sequence alignments of PiggyMac paralog nucleotide sequences from *P. tetraurelia* and *P. octaurelia* using MAFFT version 7.450 (Katoh and Standley 2013) (Katoh et al. 2002) (algorithm: "auto", scoring matrix: 200PAM/k=2, gap open penalty 1.53, offset value 0.123) using the "translation align" panel of Geneious Prime (version 2020.1.2) (Kearse et al. 2012). PAML version 4.9 (Yang 2007) was used to estimate d_N/d_S values in pairwise mode (runmode = -2, seqtype = 1, CodonFreq = 2). For *Blepharisma stoltei*, we generated pairwise coding sequence alignments of the *Blepharisma* PiggyMac homolog, BPgm (Contig_49.g1063; BSTOLATCC_MAC17466), with the *Blepharisma* Pgm-likes (BPgmLs) using Translation Align panel of Geneious v11.1.5 (Genetic code: *Blepharisma*, Protein alignment options: MAFFT alignment (v7.450) (Katoh and Standley 2013), scoring matrix: BLOSUM62, Gap open penaly: 1.53, offset value: 0.1). PAML version 4.9 was used to estimate dN/dS values in pairwise mode (runmode = -2, seqtype = 1, CodonFreq = 2).

8.22. Phylogenetic analysis of eukaryotic PiggyBac-like elements

Protein sequences of PBLEs were obtained from Bouallègue et al (Bouallègue et al. 2017). Protein sequences of *Paramecium* and *Tetrahymena* Pgms and PgmLs were obtained from *Paramecium*DB (Arnaiz, Meyer, and Sperling 2020) (PGM, PGMLs1-5) and ciliate.org (Stover et al. 2012) (Tpb1, Tpb2, Tpb7, LIA5), respectively. *Condylostoma* and *Blepharisma* Pgms and PgmLs were obtained from genome assemblies (accessions GCA_920105805 and GCA_905310155, respectively). Sequence manipulation was done using Geneious (Biomatters Ltd.). The Geneious plug-in for InterProScan (Jones et al. 2014) was used to identify DDE_Tnp_1_7 domains using the PFAM-A database (Mistry et al. 2021). The DDE_Tnp_1_7 domain and regions adjacent to it were extracted and aligned using the MAFFT plug-in (v7.450) for Geneious (Katoh and Standley 2013) (Algorithm: L-INS-i, Scoring matrix: BLOSUM62, Gap open penalty: 1.53, Offset value: 0.123). Phylogenetic trees using this alignment were generated with the FastTree2 (v2.2.11) Geneious plug-in using the Whelan-Goldman model. The phylogenetic trees were visualized with FigTree (v1.4.4) (Andrew Rambaut, http://tree.bio.ed.ac.uk/).

8.23. Repeat annotation

Interspersed repeat element families were predicted with RepeatModeler v2.0.1 (default settings, random number seed 12345) with the following dependencies: rmblast v2.9.0+

(http://www.repeatmasker.org/RMBlast.html), TRF 4.09 (Benson, 1999), RECON (Bao and Eddy, 2002), RepeatScout 1.0.6 (Price et al., 2005), RepeatMasker v4.1.1 (http://www.repeatmasker.org/RMDownload.html). Repeat families were also classified in the pipeline by RepeatClassifier v2.0.1 through comparison against RepeatMasker's repeat protein database and the Dfam database. Consensus sequences of the predicted repeat families, produced by RepeatModeler, were then used to annotate repeats with RepeatMasker, using rmblast as the search engine.

Terminal inverted repeats (TIRs) of selected repeat element families were identified by aligning the consensus sequence from RepeatModeler, and/or selected full-length elements, with their respective reverse complements using MAFFT (Katoh and Standley, 2013) (plugin version distributed with Geneious). TIRs from the Dfam DNA transposon termini signatures database (v1.1, https://www.dfam.org/releases/dna_termini_1.1/dna_termini_1.1.hmm.gz) (Storer et al., 2021) were searched with hmmsearch (HMMer v3.2.1) against the IES sequences, to identify matches to TIR signatures of major transposon subfamilies.

8.24. IES prediction from PacBio subreads

PacBio subreads (CLR reads) from a MIC-enriched sample (ENA accession ERR6548140) were aligned to the somatic genome reference assembly (accession PRJEB40285) (Singh et al. 2021) with minimap2 v2.17-r941 (H. Li 2018), with options: -ax map-pb --secondary=no --MD. Mapped reads were sorted and indexed with samtools v1.10 (H. Li et al. 2009), and then used for predicting IESs with BleTIES MILRAA v0.1.9, with options: --type subreads --junction_flank 5 --min_ies_length 15 --min_break_coverage 10 --subreads_pos_max_cluster_dist 5. The BleTIES pipeline has been previously described (Seah and Swart 2021) and uses spoa v4.0.3 (Vaser et al. 2017) for assembly. After inspecting the initial IES predictions, we removed IES predictions with length <50 bp and retention score <0.075, which we judged to be more likely to be spurious or to have insufficient coverage for an accurate assembly.

Terminal direct repeats (TDRs) at the boundary of a given IES were defined as a sequence of any length that was exactly repeated on both ends of the IES, such that one copy lies within the IES, and the other in the MAC-destined sequence. Because the sequence is identical, it is not possible to determine from sequencing data alone where the physical excision of the IES would occur; such ambiguous excision junctions have been termed "floating IESs" (Sellis et al.

2021). Therefore, TDRs were always reported starting from the left-most coordinate. If the TDR sequence contained 5'-TA-3', the corresponding IES was also considered to be "TA-bound", even if the TDR was longer than the 2 bp 5'-TA-3' sequence.

Reconstructed IES sequences were computationally inserted into the MAC assembly with BleTIES Insert, to produce a hybrid MAC+IES assembly, which approximates the part of the MIC genome that is collinear with the MAC.

8.25. Identification and comparison of IES length classes

Visual inspection of the length distribution of BleTIES-predicted IESs showed sharp peaks every ~10 bp between ~65 and 115 bp. Peak calling on the graph of number of IESs (TA-bound only) vs. length (bp) was performed with the function find_peaks from the Python package scipy.signal v1.3.1 (Virtanen et al. 2020), with height cutoff 100. IES size classes were defined with the width at half peak height. In *Paramecium tetraurelia*, where most IESs are TA-bound, the IES termini have a short, weakly conserved inverted repeat (Klobutcher and Herrick 1995; Arnaiz et al. 2012). To search for similar motifs in *B. stoltei*, sequences flanking TA-bound IES junctions were extracted, with one from each pair reverse-complemented so that the sequences were always in the orientation 5'-(MDS segment)-TA-(IES segment)-3'. Sequence logos of the junctions (10 bp MDS, 14 bp within IES, not including the TA itself) were drawn for each IES length class with Weblogo (Crooks et al. 2004). Only TA-bound IESs were used for the sequence logos because they could be aligned relative to the 5'-TA-3' repeat, whereas for IESs bound by other types of junctions there is no common reference point to align the boundaries of the IES.

8.26. Probability of terminal direct repeat-bound IESs

Under a null model where all bases in a sequence are independently and identically distributed, the probability P_n of having any possible terminal direct repeat (TDR) of length n bounding a given sequence feature is the sum if probabilities of all possible sequences k of length n, squared: Pn=kKpk2, which can be simplified to Pn =(bBpb2)n, where B is the alphabet of bases and p_b is the individual probability of each base. The number of possible sequences k of length n is simply $|K| = |B|^n$.

The probability of having a TDR of length at least 2 is equal to the probability of having a TDR of length 2, because all cases of TDR length > 2 implicitly have a TDR of length = 2.

Therefore the probability of having a TDR of length exactly *n*, i.e. match in bases 1 to *n*, and mismatch on base *n*+1 is Pn (mismatch) = Pn(1-bBpb2). The expected number of TDRs in *Blepharisma* were calculated by using the empirical base frequencies of the MAC+IES genome assembly for p_b, and multiplying this probability by the number of IESs.

8.27. Identification of terminal inverted repeats (TIRs) and palindromes in IESs

The BleTIES-assembled IES sequences for *Blepharisma* were used to identify exact, ungapped terminal inverted repeats (TIRs). Starting from the ends of the IES sequence immediately within the flanking TDRs, each base was compared to the reverse complement of the corresponding base on the opposite end for a match, extending the TIR until a mismatch was encountered, up to a maximum length of 25 bp. The same procedure was used for *Paramecium tetraurelia* using IESs sequences downloaded from *Paramecium*DB (https://*paramecium*.i2bc.paris-saclay.fr/files/*Paramecium*/*tetraurelia*/51/annotations/p*tetraurelia*_mac_51_with_ies, accessed 14 October 2021), except that the coordinates of TDRs were first renumbered and extended beyond the "TA" motif if possible, following the BleTIES coordinate numbering convention, in case there are potential TDRs that are longer than a simple TA. The expected number of TIRs of given lengths under a null model was computed in the same way as the expectation for TDRs (see "Probability of terminal direct repeat-bound IESs").

Long TIRs (≥10 bp) were clustered by sequence identity to look for IESs of potentially related origin, using the cluster_fast algorithm (Edgar 2010) implemented in Vsearch v2.13.6 (Rognes et al. 2016) at 80% identity and the CD-HIT definition of sequence identity (-iddef 0). For each resulting cluster of similar TIRs, the cluster centroid was used as the representative sequence shown in Figure TIRS. TDRs associated with each cluster's IESs were grouped by length, and for each TDR length a degenerate consensus was reported with the degenerate_consensus function of the Bio.motifs module in Biopython v1.74.

Palindromic IESs were defined as IESs that align to their own reverse complement with a sequence identity ≥90% (matching columns over sequence length); this definition was less strict and permitted inexact matches unlike the TIR search, to allow for sequence divergence and assembly errors. IES sequences were aligned with the PairwiseAligner function from Bio.Align in BioPython v1.74, using global mode and parameter match_score = 1.0, with all other scores set to zero.

Palindromic IESs were clustered with Vsearch cluster_fast as described above, except that one sequence (BSTOLATCC_IES35757) was manually removed after inspection of results because it appears to contain two different nested palindromic sequences. Cluster centroids were aligned pairwise as above and used to calculate a matrix of edit distances (matching columns / alignment length). The distance matrix was clustered with average linkage clustering to produce a sequence distance dendrogram with the functions average and dendrogram from scipy.cluster.hierarchy v1.3.1 (Virtanen et al. 2020).

8.28. Comparison of intragenic:intergenic IES ratios

Intragenic vs. intergenic IESs were defined by overlap of predicted IES annotations with "gene" feature annotations on the MAC reference (ENA accession GCA_905310155), using Bedtools v2.30.0 (Quinlan and Hall 2010) and pybedtools v0.8.1 (Dale, Pedersen, and Quinlan 2011).

To test whether the underrepresentation of IESs within gene features was statistically significant, compared to the null hypothesis of IESs and gene feature locations being independently distributed, we assumed that the number of intragenic IESs would follow a binomial distribution with individual probability equal to the fraction of the genome that is covered by gene features. The p-value of the observed number of intragenic IESs would then be equal to the cumulative probability density up to and including the observed value.

8.29. Developmental time series small RNA-seq

Complementary mating strains *B. stoltei* ATCC 30299 and HT-IV were pre-treated with Gamone 2 and Gamone 1 respectively, and then mixed to initiate conjugation as described previously; sRNA and mRNA were isolated from total RNA at the same time points ("Conjugation time course", (Singh et al. 2021)). sRNA libraries were prepared with the BGISeq-500 Small RNA Library protocol, which selects 18 to 30nt sRNAs by polyacrylamide gel electrophoresis, and sequenced on a BGISeq 500 instrument.

8.30. Small RNA libraries mapping and comparison

Small RNA libraries were mapped to the MAC+IES assembly with bowtie2 v2.4.2 (Langmead and Salzberg 2012) using default parameters. Total reads mapping to CDS vs. IES features were counted with featureCounts v2.0.1 (Liao, Smyth, and Shi 2014). To account for different total sequence lengths represented by CDSs, IESs, and intergenic regions, the read

counts were converted to relative expression values (reads per kbp transcript per million reads mapped, RPKM (Mortazavi et al. 2008)) using the total lengths of each feature type in place of transcript length in the original definition of RPKM, with the following formula:

10^9 × (reads mapped to feature type) / (total reads mapped × total length of feature type).

Reads mapping to CDSs, IESs, or neither (but excluding tRNA and rRNA features) were extracted with samtools view, with 22 and 24 nt reads extracted to separate files. Read length distributions for each sequence length and feature type were summarized with samtools stats.

8.31. Gene prediction and domain annotation in IES regions

To predict protein-coding genes in IESs, non-IES nucleotides in the MAC+IES assembly were first masked with 'N's. The Intronarrator pipeline (https://github.com/Swart-lab/Intronarrator), a wrapper around Augustus (Stanke and Waack 2003), was run with the same parameters as for the *B. stoltei* MAC genome, i.e. a cut-off of 0.2 for the fraction of spliced reads covering a potential intron, and ≥10 reads to call an intron (Singh et al. 2021). Without masking, gene predictions around IESs were poor, with genuine MDS-limited genes (with high RNAseq coverage) incorrectly extended into IES regions. The possibility of genes spanning IES boundaries was not catered for. Domain annotations for diagrams were generated with the InterproScan 5.44-79.0 pipeline (Jones et al. 2014) incorporating HMMER (v3.3, Nov 2019, hmmscan) (Eddy 2011).

For comparison of transposase-related domain content in MAC vs. MIC, reference sequences were obtained from public databases for *Paramecium tetraurelia* (https://*paramecium*.i2bc.paris-saclay.fr/files/*Paramecium*/*tetraurelia*/51/annotations/p*tetraurelia*_mac_51_with_ies/), *Tetrahymena thermophila* (http://www.ciliate.org/system/downloads/3-upd-cds-fasta-2021.fasta), and *Oxytricha trifallax* (https://oxy.ciliate.org/common/downloads/oxy/Oxy2020_CDS.fasta, https://knot.math.usf.edu/mds_ies_db/data/gff/oxytri_mic_non_mds.gff). IES gene prediction in *Blepharisma* was hampered by intermittent polynucleotide tract length errors, due to the assembly of IESs from PacBio CLR reads. To mitigate this, a six-frame translation of the MIC-limited genome regions was performed using a custom script, then scanned against the Pfam-A database 32.0 (release 9) (Mistry et al. 2021) with hmmscan (HMMER), with i-E-value cutoff $\leq 10^{-6}$.

8.32. Repeat annotation and clustering

To evaluate the repetitive sequence content in IESs, we applied a repeat prediction and annotation to the combined MAC+IES assembly, instead of clustering whole IESs by sequence similarity. This was so that: (i) Repeats shared between the MDS and IES could be identified. (ii) Complex structures such as nested repeats could be detected. (iii) Repeat families were predicted de novo, permitting discovery of novel elements. (iv) Repeats did not have to be strictly identical to be grouped into a family.

Interspersed repeat element families were predicted from the MAC+IES genome assembly with RepeatModeler v2.0.1 (default settings, random number seed 12345) with the following dependencies: rmblast v2.9.0+ (http://www.repeatmasker.org/RMBlast.html), TRF 4.09 (Benson 1999), RECON (Bao and Eddy 2002), RepeatScout 1.0.6 (A. L. Price, Jones, and Pevzner 2005), RepeatMasker v4.1.1 (http://www.repeatmasker.org/RMDownload.html). Repeat families were also classified in the pipeline by RepeatClassifier v2.0.1 through comparison against RepeatMasker's repeat protein database and the Dfam database. Consensus sequences of the predicted repeat families, produced by RepeatModeler, were then used to annotate repeats in the MAC+IES assembly with RepeatMasker, using rmblast as the search engine.

The consensus sequences for rnd-1_family-0 and rnd-1_family-73 were manually curated for downstream analyses. For rnd-1_family-0 (BogoMITE) the original consensus predicted by RepeatModeler for rnd-1_family-0 was 784 bp long, but this was a spurious inverted duplication of the basic ~390 bp unit; the duplication had been favored in the construction of the consensus because RepeatModeler attempts to find the longest possible match to represent each family. For family rnd-1_family-73 (containing BstTc1 transposon), the actual repeat unit was longer than the boundaries predicted by RepeatModeler. In most IESs that contain this repeat (19 of 22), it was flanked by and partially overlapping with short repeat elements from families rnd-4_family-1308 and rnd-1_family-117, which are spurious predictions. Repeat unit boundaries were manually defined by alignment of full length repeats and their flanking regions.

Terminal inverted repeats of selected repeat element families were identified by aligning the consensus sequence from RepeatModeler, and/or selected full-length elements, with their respective reverse complements using MAFFT (Katoh and Standley 2013) (plugin version distributed with Geneious).

TIRs from the Dfam DNA transposon termini signatures database (v1.1, https://www.dfam.org/releases/dna_termini_1.1/dna_termini_1.1.hmm.gz) (Storer et al. 2021) were searched with hmmsearch (HMMer v3.2.1) against the IES sequences, to identify matches to TIR signatures of major transposon subfamilies.

8.33. Phylogenetic analysis of Tc1/Mariner-superfamily transposases

Repeat family rnd-1_family-1 was initially classified as a "TcMar/Tc2" family transposable element by RepeatClassifier. 30 full length copies (>95% of the consensus length) were annotated by RepeatMasker, all of which fell within IESs and contained CDS predictions. However, CDSs were of varying lengths because of frameshifts caused by indels, which may be biological or due to assembly error; nonetheless, the nucleotide sequences had high pairwise identity (about 98%, except for one outlier). We chose Contig_12.g100018 as the representative CDS sequence for phylogenetic analysis because it was one of the longest predicted and both predicted Pfam domains (*HTH_Tnp_Tc5* and *DDE_1*) appeared to be intact.

For repeat family rnd-1_family-73, the initial classification was "DNA/TcMar-Tc1". As described above, CDS predictions were of variable lengths, and the longest CDSs were not necessarily the best versions of the sequence because of potential frameshift errors. For phylogenetic analysis, we chose Contig_51.g100045 as the representative copy, because a complete *DDE_3* Pfam domain was predicted by HMMER that could align with other DDE/D domains from reference alignments described below.

The representative CDSs of the rnd-1_family-1 and rnd-1_family-73 transposases were aligned with MAFFT (E-INS-i mode) against a published DDE/D domain reference alignment (Supporting Information Dataset_S01 of (Yuan and Wessler 2011)) to identify the residues at the conserved catalytic triad and the amino acid distance between the conserved residues.

For the phylogenetic analysis of the DDE/D domains in the Tc1/Mariner superfamily, both MAC- and MIC-limited genes containing DDE_1 and DDE_3 domains were separately aligned for each Pfam domain with MAFFT v7.450 (algorithm: E-INS-i, scoring matrix: BLOSUM62, Gap open penalty: 1.53) and trimmed to the DDE/D domain with Geneious and incomplete domains were removed. As reference, 204 sequences from a published alignment (Additional File 4 of (Dupeyron et al. 2020)) were selected to represent the 53 groups defined in that study, choosing only complete domains (with all three conserved catalytic residues) and all

Oxytricha trifallax TBE and *Euplotes crassus* Tec transposase sequences. Thirteen *Paramecium* Tc1/Mariner DDE/D domain consensus sequences were added (Additional File 4 of (Guérin et al. 2017)). Sequences were aligned with MAFFT (E-INS-i mode) and trimmed to only the DDE/D domain boundaries with Geneious. Phylogeny was inferred with FastTree2 v2.1.11 (M. N. Price, Dehal, and Arkin 2010) using the WAG substitution model. The tree was visualized with Dendroscope v3.5.10 (Huson and Scornavacca 2012), rooted with bacterial IS630 sequences as outgroup

8.34. Phylogenetic analysis of retrotransposon-derived sequences

All the nucleotide sequences ≥500 bp for the repeat families identified by RepeatClassifier as LINE or LINE/RTE-x: rnd-1_family-273, rnd-1_family-276 and rnd-4_family-193 were aligned to one another with MAFFT v7.450 (automatic algorithm) (Katoh and Standley 2013), with the option to automatically determine sequence direction (via the MAFFT plugin for Geneious Prime (Kearse et al. 2012)). Since the alignment appeared to be poor between the rnd-4-family-193 sequences and the rest, we generated separate alignments for this family from the other two, also with MAFFT (E-INS-i mode). Maximum likelihood phylogenies were generated by PhyML (Guindon et al. 2010) version 3.3.20180621 with the HKY85 substitution model.

8.35. Sequence visualization and analysis

Nucleotide and amino acid sequences were visualized using Geneious Prime (Biomatters Ltd.) (Kearse et al. 2012).

8.36. Data availability

The draft *Blepharisma stoltei* ATCC 30299 MAC genome assembly is accessible from bleph.ciliate.org and from the European Nucleotide Archive (ENA) bioproject PRJEB40285 under the accession GCA_905310155. PacBio CCS reads (ERR5873783 and ERR5873334) and subreads (ERR5962314) used to assemble the genome are also available from ENA. Illumina DNA-seq data for the *B. stoltei* ATCC 30299 and HT-IV strains is available from accessions ERR6061285 and ERR6064674, respectively. The RNA-seq developmental time course is available from the bioproject PRJEB45374 (accessions ERR6049461-ERR6049485).

Illumina and PacBio Sequel sequencing data for *Blepharisma japonicum* strain R1702 is available from the ENA bioproject PRJEB46921 (Illumina accessions: ERR6473251, ERR6474356; PacBio accession: ERR6474383).

Annotated draft MAC+IES genome for *Blepharisma stoltei* strain ATCC 30299 (European Nucleotide Archive (ENA) Bioproject PRJEB46944 under accession GCA_914767885). IES sequences and annotations, MAC gene predictions with intervening IESs, and gene predictions within IESs (EDMOND, doi: 10.17617/3.83; genome browser, https://bleph.ciliate.org. Sequencing data for the MIC-enriched nuclear fractions (PacBio CLR reads: ENA accession ERR6510520 and ERR6548140; BGI-seq reads: ENA accessions ERR6474675, ERR6496962, ERR6497067, ERR6501836). Small RNA libraries from developmental time series (ENA Bioproject PRJEB47200 under accessions ERR6565537-ERR6565561). Repeat family predictions and annotations by RepeatModeler and RepeatMasker (EDMOND, doi: 10.17617/3.82). Alignment and phylogeny of Tc1/Mariner superfamily transposase domains (EDMOND, doi:10.17617/3.JLWBFM)

8.37. Bibliography

Andersen, Robert A. 2004. *Algal Culturing Techniques*. 1st Edition.

Arnaiz, Olivier, Nathalie Mathy, Céline Baudry, Sophie Malinsky, Jean-Marc Aury, Cyril Denby Wilkes, Olivier Garnier, et al. 2012. "The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences." *PLoS Genetics* 8 (10): e1002984. https://doi.org/10.1371/journal.pgen.1002984.

Arnaiz, Olivier, Eric Meyer, and Linda Sperling. 2020. "*Paramecium*DB 2019: Integrating Genomic Data across the Genus for Functional and Evolutionary Biology." *Nucleic Acids Research* 48 (D1): D599–605. https://doi.org/10.1093/nar/gkz948.

Bao, Zhirong, and Sean R Eddy. 2002. "Automated de Novo Identification of Repeat Sequence Families in Sequenced Genomes." *Genome Research* 12 (8): 1269–76. https://doi.org/10.1101/gr.88502.

Benson, G. 1999. "Tandem Repeats Finder: A Program to Analyze DNA Sequences." *Nucleic Acids Research* 27 (2): 573–80. https://doi.org/10.1093/nar/27.2.573.

Bouallègue, Maryem, Jacques-Deric Rouault, Aurélie Hua-Van, Mohamed Makni, and Pierre Capy. 2017. "Molecular Evolution of PiggyBac Superfamily: From Selfishness to Domestication." *Genome Biology and Evolution* 9 (2): 323–39. https://doi.org/10.1093/gbe/evw292.

Chan, Patricia P., Brian Y. Lin, Allysia J. Mak, and Todd M. Lowe. 2019. "TRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes." *BioRxiv*, April. https://doi.org/10.1101/614032.

Cock, Peter J A, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. "Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics." *Bioinformatics* 25 (11): 1422–23. https://doi.org/10.1093/bioinformatics/btp163.

Crooks, G E, G Hon, J M Chandonia, and S E Brenner. 2004. "WebLogo: A Sequence Logo Generator." *Genome Research* 14 (6): 1188–90. https://doi.org/10.1101/gr.849004.

Dale, Ryan K, Brent S Pedersen, and Aaron R Quinlan. 2011. "Pybedtools: A Flexible Python Library for Manipulating Genomic Datasets and Annotations." *Bioinformatics* 27 (24): 3423–24. https://doi.org/10.1093/bioinformatics/btr539.

Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008.

Dupeyron, Mathilde, Tobias Baril, Chris Bass, and Alexander Hayward. 2020. "Phylogenetic Analysis of the Tc1/Mariner Superfamily Reveals the Unexplored Diversity of Pogo-like Elements." *Mobile DNA* 11 (June): 21. https://doi.org/10.1186/s13100-020-00212-0.

Dutilh, Bas E, Rasa Jurgelenaite, Radek Szklarczyk, Sacha A F T van Hijum, Harry R Harhangi, Markus Schmid, Bart de Wild, et al. 2011. "FACIL: Fast and Accurate Genetic Code Inference and Logo." *Bioinformatics* 27 (14): 1929–33. https://doi.org/10.1093/bioinformatics/btr316.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7 (10): e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster than BLAST." *Bioinformatics* 26 (19): 2460–61. https://doi.org/10.1093/bioinformatics/btq461.

Garrison, Erik, and Gabor Marth. 2012. "Haplotype-Based Variant Detection from Short-Read Sequencing." v2 ed. arXiv. https://arxiv.org/abs/1207.3907.

Gruber-Vodicka, Harald R, Brandon K B Seah, and Elmar Pruesse. 2020. "PhyloFlash: Rapid Small-Subunit RRNA Profiling and Targeted Assembly from Metagenomes." *MSystems* 5 (5). https://doi.org/10.1128/mSystems.00920-20.

Guérin, Frédéric, Olivier Arnaiz, Nicole Boggetto, Cyril Denby Wilkes, Eric Meyer, Linda Sperling, and Sandra Duharcourt. 2017. "Flow Cytometry Sorting of Nuclei Enables the First Global Characterization of *Paramecium* Germline DNA and Transposable Elements." *BMC Genomics* 18 (1): 327. https://doi.org/10.1186/s12864-017-3713-7.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." *Systematic Biology* 59 (3): 307–21. https://doi.org/10.1093/sysbio/syq010.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics* 29 (8): 1072–75. https://doi.org/10.1093/bioinformatics/btt086.

Harris, Robert S, Monika Cechova, and Kateryna D Makova. 2019. "Noise-Cancelling Repeat Finder: Uncovering Tandem Repeats in Error-Prone Long-Read Sequencing Data." *Bioinformatics* 35 (22): 4809–11. https://doi.org/10.1093/bioinformatics/btz484.

Harumoto, Terue, Akio Miyake, Naoko Ishikawa, Rika Sugibayashi, Kazutaka Zenfuku, and Hideo lio. 1998. "Chemical Defense by Means of Pigmented Extrusomes in the Ciliate *Blepharisma Japonicum*." *European Journal of Protistology* 34 (4): 458–70. https://doi.org/10.1016/S0932-4739(98)80014-X.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, et al. 2019. "EggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses." *Nucleic Acids Research* 47 (D1): D309–14. https://doi.org/10.1093/nar/gky1085.

Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Huson, Daniel H, and Celine Scornavacca. 2012. "Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks." *Systematic Biology* 61 (6): 1061–67. https://doi.org/10.1093/sysbio/sys062.

Hyatt, Doug, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March): 119. https://doi.org/10.1186/1471-2105-11-119.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30 (9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

Kalvari, Ioanna, Joanna Argasinska, Natalia Quinones-Olvera, Eric P Nawrocki, Elena Rivas, Sean R Eddy, Alex Bateman, Robert D Finn, and Anton I Petrov. 2018. "Rfam 13.0: Shifting to a Genome-Centric Resource for Non-Coding RNA Families." *Nucleic Acids Research* 46 (D1): D335–42. https://doi.org/10.1093/nar/gkx1038.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform." *Nucleic Acids Research* 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.

Katoh, Kazutaka, and Daron M Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kearse, Matthew, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, et al. 2012. "Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data." *Bioinformatics* 28 (12): 1647–49. https://doi.org/10.1093/bioinformatics/bts199.

Kim, Daehwan, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. 2019. "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37 (8): 907–15. https://doi.org/10.1038/s41587-019-0201-4.

Klobutcher, L A, and G Herrick. 1995. "Consensus Inverted Terminal Repeat Sequence of *Paramecium* IESs: Resemblance to Termini of Tc1-Related and *Euplotes* Tec Transposons." *Nucleic Acids Research* 23 (11): 2006–13. https://doi.org/10.1093/nar/23.11.2006.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A Pevzner. 2019. "Assembly of Long, Error-Prone Reads Using Repeat Graphs." *Nature Biotechnology* 37 (5): 540–46. https://doi.org/10.1038/s41587-019-0072-8.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lauth, M R, B B Spear, J Heumann, and D M Prescott. 1976. "DNA of Ciliated Protozoa: DNA Sequence Diminution during Macronuclear Development of *Oxytricha*." *Cell* 7 (1): 67–74. https://doi.org/10.1016/0092-8674(76)90256-7.

Liao, Yang, Gordon K Smyth, and Wei Shi. 2014. "FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Li, Bo, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey. 2010. "RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty." *Bioinformatics* 26 (4): 493–500. https://doi.org/10.1093/bioinformatics/btp692.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Lopez-Delisle, Lucille, Leily Rabbani, Joachim Wolff, Vivek Bhardwaj, Rolf Backofen, Björn Grüning, Fidel Ramírez, and Thomas Manke. 2021. "PyGenomeTracks: Reproducible Plots for Multivariate Genomic Datasets." *Bioinformatics* 37 (3): 422–23. https://doi.org/10.1093/bioinformatics/btaa692.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, et al. 2021. "Pfam: The Protein Families Database in 2021." *Nucleic Acids Research* 49 (D1): D412–19. https://doi.org/10.1093/nar/gkaa913.

Miyake, A, and J Beyer. 1973. "Cell Interaction by Means of Soluble Factors (Gamones) in Conjugation of *Blepharisma* Intermedium." *Experimental Cell Research* 76 (1): 15–24. https://doi.org/10.1016/0014-4827(73)90413-8.

Miyake, A, T Harumoto, B Salvi, and V Rivola. 1990. "Defensive Function of Pigment Granules in *Blepharisma* Japonicum." *European Journal of Protistology* 25 (4): 310–15. https://doi.org/10.1016/S0932-4739(11)80122-7.

Miyake, A, V Rivola, and T Harumoto. 1991. "Double Paths of Macronucleus Differentiation at Conjugation in *Blepharisma Japonicum*." *European Journal of Protistology* 27 (2): 178–200. https://doi.org/10.1016/S0932-4739(11)80340-8.

Miyake, AKIO. 1981. "Cell Interaction by Gamones in *Blepharisma*." In *Sexual Interactions in Eukaryotic Microbes*, 95–129. Elsevier. https://doi.org/10.1016/B978-0-12-524160-1.50010-2.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Nawrocki, Eric P, Diana L Kolbe, and Sean R Eddy. 2009. "Infernal 1.0: Inference of RNA Alignments." *Bioinformatics* 25 (10): 1335–37. https://doi.org/10.1093/bioinformatics/btp157.

Price, Alkes L, Neil C Jones, and Pavel A Pevzner. 2005. "De Novo Identification of Repeat Families in Large Genomes." *Bioinformatics* 21 Suppl 1 (June): i351-8. https://doi.org/10.1093/bioinformatics/bti1018.

Price, Morgan N, Paramvir S Dehal, and Adam P Arkin. 2010. "FastTree 2 — Approximately Maximum-Likelihood Trees for Large Alignments." *Plos One* 5 (3): e9490. https://doi.org/10.1371/journal.pone.0009490.

Prjibelski, Andrey, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, and Anton Korobeynikov. 2020. "Using SPAdes de Novo Assembler." *Current Protocols in Bioinformatics* 70 (1): e102. https://doi.org/10.1002/cpbi.102.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (Database issue): D590-6. https://doi.org/10.1093/nar/gks1219.

Quevillon, E, V Silventoinen, S Pillai, N Harte, N Mulder, R Apweiler, and R Lopez. 2005. "InterProScan: Protein Domains Identifier." *Nucleic Acids Research* 33 (Web Server issue): W116-20. https://doi.org/10.1093/nar/gki442.

Quinlan, Aaron R, and Ira M Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Repak, Arthur J. 1968. "Encystment and Excystment of the Heterotrichous Ciliate *Blepharisma Stoltei* Isquith." *Journal of Protozoology* 5: 407–12.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. "VSEARCH: A Versatile Open Source Tool for Metagenomics." *PeerJ* 4 (October): e2584. https://doi.org/10.7717/peerj.2584.

Sambrook, Joseph, and David W Russell. 2006. "Purification of Nucleic Acids by Extraction with Phenol:Chloroform." *CSH Protocols* 2006 (1): pii: pdb.prot4455. https://doi.org/10.1101/pdb.prot4455.

Saudemont, Baptiste, Alexandra Popa, Joanna L Parmley, Vincent Rocher, Corinne Blugeon, Anamaria Necsulea, Eric Meyer, and Laurent Duret. 2017. "The Fitness Cost of Mis-Splicing Is the Main Determinant of Alternative Splicing Patterns." *Genome Biology* 18 (1): 208. https://doi.org/10.1186/s13059-017-1344-6.

Schindelin, Johannes, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, et al. 2012. "Fiji: An Open-Source Platform for Biological-Image Analysis." *Nature Methods* 9 (7): 676–82. https://doi.org/10.1038/nmeth.2019.

Seah, Brandon K B, and Estienne C Swart. 2021. "BleTIES: Annotation of Natural Genome Editing in Ciliates Using Long Read Sequencing." *Bioinformatics* 37 (21): 3929–31. https://doi.org/10.1093/bioinformatics/btab613.

Sellis, Diamantis, Frédéric Guérin, Olivier Arnaiz, Walker Pett, Emmanuelle Lerat, Nicole Boggetto, Sascha Krenek, et al. 2021. "Massive Colonization of Protein-Coding Exons by Selfish Genetic Elements in *Paramecium* Germline Genomes." *PLoS Biology* 19 (7): e3001309. https://doi.org/10.1371/journal.pbio.3001309.

Singh, Minakshi, Brandon K. B. Seah, Christiane Emmerich, Aditi Singh, Christian Woehle, Bruno Huettel, Adam Byerly, et al. 2021. "The *Blepharisma Stoltei* Macronuclear Genome: Towards the Origins of Whole Genome Reorganization." *BioRxiv*, December. https://doi.org/10.1101/2021.12.14.471607.

Stanke, Mario, and Stephan Waack. 2003. "Gene Prediction with a Hidden Markov Model and a New Intron Submodel." *Bioinformatics* 19 Suppl 2 (October): ii215-25. https://doi.org/10.1093/bioinformatics/btg1080.

Storer, Jessica, Robert Hubley, Jeb Rosen, Travis J Wheeler, and Arian F Smit. 2021. "The Dfam Community Resource of Transposable Element Families, Sequence Models, and Genome Annotations." *Mobile DNA* 12 (1): 2. https://doi.org/10.1186/s13100-020-00230-y.

Stover, Nicholas A, Ravinder S Punia, Michael S Bowen, Steven B Dolins, and Theodore G Clark. 2012. "*Tetrahymena* Genome Database Wiki: A Community-Maintained Model Organism Database." *Database: The Journal of Biological Databases and Curation* 2012 (March): bas007. https://doi.org/10.1093/database/bas007.

Sugiura, Mayumi, Yuri Tanaka, Toshinobu Suzaki, and Terue Harumoto. 2012. "Alternative Gene Expression in Type I and Type II Cells May Enable Further Nuclear Changes during Conjugation of *Blepharisma* Japonicum." *Protist* 163 (2): 204–16. https://doi.org/10.1016/j.protis.2011.07.007.

Swart, Estienne Carl, Valentina Serra, Giulio Petroni, and Mariusz Nowacki. 2016. "Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination." *Cell* 166 (3): 691–702. https://doi.org/10.1016/j.cell.2016.06.020.

Törönen, Petri, Alan Medlar, and Liisa Holm. 2018. "PANNZER2: A Rapid Functional Annotation Web Server." *Nucleic Acids Research* 46 (W1): W84–88. https://doi.org/10.1093/nar/gky350.

Vaser, Robert, and Mile Sikic. 2019. "Yet Another de Novo Genome Assembler." *BioRxiv*, May. https://doi.org/10.1101/656306.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads." *Genome Research* 27 (5): 737–46. https://doi.org/10.1101/gr.214270.116.

Virtanen, Pauli, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17 (3): 261–72. https://doi.org/10.1038/s41592-019-0686-2.

Wagner, Günter P, Koryu Kin, and Vincent J Lynch. 2012. "Measurement of MRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent among Samples." *Theory in Biosciences = Theorie in Den Biowissenschaften* 131 (4): 281–85. https://doi.org/10.1007/s12064-012-0162-3.

Waterhouse, Robert M, Mathieu Seppey, Felipe A Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V Kriventseva, and Evgeny M Zdobnov. 2018. "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics." *Molecular Biology and Evolution* 35 (3): 543–48. https://doi.org/10.1093/molbev/msx319.

Yang, Ziheng. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8): 1586–91. https://doi.org/10.1093/molbev/msm088.

Yuan, Yao-Wu, and Susan R Wessler. 2011. "The Catalytic Domain of All Eukaryotic Cut-and-Paste Transposase Superfamilies." *Proceedings of the National Academy of Sciences of the United States*

of America 108 (19): 7884–89. https://doi.org/10.1073/pnas.1104208108.

Appendix

Author contributions according to CRediT- Contributor Roles Taxonomy (Brand, Allen, Altman, Hlava, & Scott, 2015).

Author abbreviations

M.S.1	Minakshi Singh
K.B.B.S.	Kwee Boon Brandon Seah
C.E.	Christiane Emmerich
A.S.	Aditi Singh
C.W.	Christian Woehle
B.H.	Bruno Huettel
A.B.	Adam Byerly
N.S.	Naomi Alexandra Stover
M.S.2	Mayumi Sugiura
T.H.	Terue Harumoto
E.C.S.	Estienne Carl Swart

A.1. Author contributions for Chapters 4, 5 and 6

Chapter 4

Conceptualization, **M.S.1**., K.B.B.S., E.C.S.

Methodology, **M.S.1**., K.B.B.S., C.E., A.S., C.W., B.H., E.C.S.

Software, **M.S.1**., E.C.S.
Investigation, **M.S.1**., E.C.S.

Writing – Original Draft, **M.S.1**., K.B.B.S., A.S., C.W., B.H., E.C.S.

Writing – Review & Editing, **M.S.1**, K.B.B.S., A.S., E.C.S.

Funding Acquisition, E.C.S.

Resources, A.B. and N.A.S.

Supervision, M.S.2., T.H., E.C.S.

Chapter 5

Conceptualization, **M.S.1**., K.B.B.S., E.C.S.

Methodology, **M.S.1**., K.B.B.S., C.E., A.S., C.W., B.H., E.C.S.

Software, **M.S.1**., K.B.B.S., E.C.S.
Investigation, **M.S.1**., K.B.B.S., C.W., B.H., E.C.S.

Writing – Original Draft, **M.S.1**., K.B.B.S., A.S., C.W., B.H., E.C.S.

Writing – Review & Editing, **M.S.1**, K.B.B.S., A.S., E.C.S.

Funding Acquisition, E.C.S.

Resources, A.B. and N.A.S.

Supervision, M.S.2., T.H., E.C.S.

Chapter 6

Data curation: K.B.B.S, E.C.S., **M.S.1**

Formal analysis: K.B.B.S, **M.S.1**, E.C.S., C.W.

Funding acquisition: N.S., E.C.S.

Investigation: K.B.B.S, **M.S.1**., E.C.S.

Methodology: K.B.B.S, **M.S.1**., E.C.S., M.S.2, T.H., A.S., C.E.

Resources: M.S.2., T.H., C.W., B.H., A.B., N.S.

Software: K.B.B.S, E.C.S., **M.S.1**, A.B., N.S.

Supervision: M.S.2, T.H., E.C.S.

Visualization: K.B.B.S, **M.S.1**., E.C.S.

Writing – original draft: K.B.B.S, **M.S.1**., E.C.S.

Writing – review & editing: K.B.B.S, **M.S.1**, E.C.S., M.S.2, T.H., A.S., C.W.

A.2. Author contributions for Chapters 7 and 8

Chapter 7

Data curation: **M.S.1**, K.B.B.S, E.C.S.

Formal analysis: **M.S.1**, K.B.B.S, E.C.S.

Investigation: **M.S.1**, K.B.B.S, E.C.S.

Methodology: **M.S.1**., K.B.B.S, E.C.S., M.S.2, T.H., A.S., C.E.

Software: K.B.B.S, E.C.S., **M.S.1**

Writing – original draft: **M.S.1**., K.B.B.S, E.C.S.

Writing – review & editing: **M.S.1**, K.B.B.S, E.C.S.

Chapter 8

Methodology: **M.S.1**., K.B.B.S, E.C.S., M.S.2, T.H., A.S., C.E.

Data curation: **M.S.1**, K.B.B.S, E.C.S.

Formal analysis: **M.S.1**, K.B.B.S, E.C.S.

Investigation: **M.S.1**, K.B.B.S, E.C.S.

Methodology: **M.S.1**., K.B.B.S, E.C.S., M.S.2, T.H., A.S., C.E.

Software: K.B.B.S, E.C.S., **M.S.1**

Writing – original draft: **M.S.1**., K.B.B.S, E.C.S.

Writing – review & editing: **M.S.1**, K.B.B.S, E.C.S.