March 2018

IMPACTS OF GENOME AND NUCLEAR ARCHITECTURE ON MOLECULAR EVOLUTION IN EUKARYOTES

Xyrus Maurer-Alcalá

IMPACTS OF GENOME AND NUCLEAR ARCHITECTURE ON MOLECULAR EVOLUTION IN EUKARYOTES

A Dissertation Presented

by

XYRUS X. MAURER-ALCALÁ

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2018

Program in Organismic and Evolutionary Biology

IMPACTS OF GENOME AND NUCLEAR ARCHITECTURE ON MOLECULAR EVOLUTION IN EUKARYOTES

A Dissertation Presented

by

XYRUS X. MAURER-ALCALÁ

Approved as to style and content by:

_____
Laura A. Katz, Chair

_____
Courtney Babbitt, Member

_____
Robert L. Dorit, Member

_____
Michael Hood, Member

_____
Elizabeth R. Dumont, Program Director
Organismic and Evolutionary Biology

DEDICATION

To Kelsie for her continued support and nearly infinite patience with my ever-growing curiosity and excitement for microbes.

ACKNOWLEDGMENTS

I need to thank Laura Katz for inviting me to her lab and providing me with the freedom and guidance to develop as a young scientist. The excitement and energy she brought to research has been incredibly inspiring and I am deeply indebted to her for imparting that enthusiasm on my work and life. I would like to thank all the members in the lab that have provided insightful discussions, technical assistance and great times we shared. A special thank you to Jean-David Grattepanche for all the fun 'arguments' and for specifically providing guidance and insights into the post-doctoral stage of life. I also want to thank the numerous undergraduates with whom I have worked with, especially those that were kind enough to take an exhausted graduate student to lunch for food and insights into their lives, especially Olivia Pilling, Anna Rogers and Monica Wilson. I also acknowledge Judith Wopereis, not only for the help she gave me with the microscope facilities but her never-ending enthusiasm for microscopy.

I would like to thank the members of my dissertation committee, Courtney Babbitt, Rob Dorit, and Michael Hood for their insightful comments on the manuscripts, their support and discussions about science and life beyond graduate school. Many in the OEB program made this part of my life far more exciting and fun (and commiserating when we needed to), especially Daniel Peterson, Laura Doubleday and Chi-Yun Kuo. A massive thank you to Penny Jaques for somehow managing to keep a tabs on me for everything related to UMass as I found myself more engrained at Smith College.

Finally, thanks to the members of my family for staying awake through my explanations of my work and the never-ending enthusiasm and support for my research.

ABSTRACT

IMPACTS OF GENOME AND NUCLEAR ARCHITECTURE ON MOLECULAR EVOLUTION IN EUKARYOTES

FEBRUARY 2018

XYRUS X. MAURER-ALCALÁ

B.A., UNIVERSITY OF COLORADO BOULDER

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Laura A. Katz

The traditional view of genomes suggests that they are static entities changing slowly in sequence and structure through time (e.g. evolving over geological time-scales). This outdated view has been challenged as our understanding of the dynamic nature of genomes has increased. Changes in DNA content (i.e. polyploidy) are common to specific life-cycle stages in a variety of eukaryotes, as are changes in genome content itself. These dramatic genomic changes include chromosomal deletions (i.e. paternal chromosome deletion in insects; Goday and Esteban 2001; Ross, et al. 2010), developmentally regulated genome rearrangements (e.g. the V(D)J system in adaptive immunity in mammals; Schatz and Swanson 2011) and the specialization of a distinct somatic genome through epigenetically regulate DNA elimination during development (found in protists and some animals; Coyne, et al. 2012; Prescott 1994; Wang and Davis 2014; Wyngaard, et al. 2011).

What likely allows genomes to be highly flexible is the separation of germline (i.e. ‘heritable’) and somatic (i.e. ‘functional’) material, even in the context of a single nucleus. Germline-soma distinctions have been best described (and most easily seen) in lineages of multicellular eukaryotes (e.g. plants, animals and fungi) due to obvious sexual structures. Germline genomes of these taxa are restricted to specialized cells (e.g.

gametes; for example, pollen grains, eggs and spores) and remain undifferentiated (and often transcriptionally inactive), whereas the somatic cells (e.g. skin, leaves, hyphae) provide the basis for ensuring organismal survival to reproductive life-stages. Sequestered germline and somatic genomes are not restricted to these well-known multi-cellular lineages but are also well-described among ciliates (the focus of this dissertation) and some foraminifera. However, in these protists, germline and somatic genomes are not isolated into distinct cells and tissues but rather are isolated into distinct nuclei that share a common cytoplasm.

Ciliates are a diverse and ancient clade of eukaryotes (~1-1.2 GYA old) and their study has led to the discovery of broad uniting eukaryotic features such as telomeres (Blackburn and Gall 1978) and self-splicing RNAs (Kruger, et al. 1982). As in the "macrobial" eukaryotes, the somatic genome (macronucleus; MAC) is transcriptionally active, transcribing all the genes necessary to maintain the cell, while the germline genome (micronucleus; MIC) remains transcriptionally inactive during the asexual portions of the life cycle. While the germline chromosomes in ciliates are physically similar to other 'traditional' eukaryotic chromosomes (e.g. being multi-Mbp with centromeres), the physical structure of the somatic chromosomes is highly variable. For example, in the model ciliate *Tetrahymena thermophila*, the somatic genome is composed of 225 unique chromosomes (most of them being ~200-400Kbp), with each at approximately 45 copies, whereas *Oxytricha trifallax*'s somatic genome is composed of ~16,000 gene-sized chromosomes (~2-3Kbp) with each chromosome at its own independent copy number (average copy number ~2,000).

Despite dramatic differences in somatic genome architecture in ciliates, the development of a new somatic genome involves. For all ciliates studied to date, this metamorphosis from 'traditional' germline chromosomal architecture to the incredibly

variable somatic genome architecture includes large-scale genome rearrangements and DNA elimination. This transformation involves the epigenetically-guided retention of somatically destined DNA from the background germline genome. While genomic rearrangements in most other eukaryotes are often fatal and are symptoms of well-known diseases (e.g. some cancers), this traditionally 'catastrophic' event is a fundamental part of ciliate life-cycles.

Although studies of ciliate germline genomes have largely been restricted to only a few genera, there appear to be broad similarities in gene organization that may be phylogenetically conserved. Ciliate germline genome architecture has been categorized as either non-scrambled or scrambled, where non-scrambled architectures are often defined as possessing macronuclear destined sequences (MDSs; soma) that are separated by germline-limited DNA and remain in consecutive order (e.g. 1-2-3-4; Figure 3.1A and Figure 4.4A). Scrambled germline architectures are highly variable, but are broadly defined as MDSs being maintained in non-consecutive order (e.g. 1-3-4-2) and/or on opposing strands of DNA (Figure 3.1 B-D and Figure 4.4B). The germline genomes of *Chilodonella uncinata* (the main focus of this dissertation) possess a combination of scrambled and non-scrambled architectures. Before my thesis work, only those ciliates with gene-sized chromosomes have been demonstrated to have scrambled germline loci. Interestingly, previous work has implicated somatic genome architecture impacting the observable accelerated rates of protein evolution in ciliates, where the proteins of those ciliates possessing 'gene-sized' chromosomes experience the greatest evolutionary rates. These observations highlight the need for further work exploring the evolutionary impacts of different germline genome architectures, as the germline structure itself has direct impact on the development of the somatic genome.

While this dissertation aims to elucidate some aspects of the evolution of germline-soma distinctions and the impact of genome and nuclear architecture (Chapters 2-4), there remain several fundamental questions that we can start addressing. For instance, in this work we observe that the most expanded gene families in *Chilodonella uncinata* are composed of genes that are disproportionately found at scrambled germline loci (Chapter 3). A major step future step will be to explore the functional implications of this increased paralog diversity through forward and reverse genetics techniques. Similarly, it will be incredibly valuable to better understand the nuclear architecture of the differing genomic contents of the three distinct nuclei present during ciliate development (i.e. the degrading parental MAC, the 'new' MIC, and the developing MAC). There may be observable compartmentalization that is exploitable or critical to the accurate rearrangement of the germline genome into a functional somatic genome. Finally, with the increasingly apparent utility of single-cell 'omics techniques (which we use in Chapters 3 and 4), there is opportunity to probe into taxonomic groups where physical germline-soma separations exist, which will provide a far more expansive understanding of the evolutionary and functional impacts of harboring multiple distinct genomes inside of a single cell/organism.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

AN EPIGENETIC TOOLKIT ALLOWS FOR DIVERSE GENOME ARCHITECTURES IN EUKARYOTES

1.1 Abstract

Genome architecture varies considerably among eukaryotes in terms of both size and structure (e.g. distribution of sequences within the genome, elimination of DNA during formation of somatic nuclei). The diversity in eukaryotic genome architectures and the dynamic processes that they undergo are only possible due to the well-developed nature of an epigenetic toolkit, which likely existed in the Last Eukaryotic Common Ancestor (LECA). This toolkit may have arisen as a means of navigating the genomic conflict that arose from the expansion of transposable elements within the ancestral eukaryotic genome. This toolkit has been coopted to support the dynamic nature of genomes in lineages across the eukaryotic tree of life. Here we highlight how the changes in genome architecture in diverse eukaryotes are regulated by epigenetic processes by focusing on DNA elimination, genome rearrangements, and adaptive changes to genome architecture. The ability to epigenetically modify and regulate genomes has contributed greatly to the diversity of eukaryotes observed today.

1.2 Introduction

Epigenetic mechanisms regulate gene expression, modify genome structures, silence mobile genetic elements, and are widespread among eukaryotes, suggesting that at least some were present in the last eukaryotic common ancestor (LECA; Cerutti and Casas-Mollano 2006; Parfrey, et al. 2008; Shabalina and Koonin 2008). For example, the

RNAi pathway that is involved in the post-transcriptional regulation of transposable elements (TEs) also plays a role in guiding large-scale chromatin remodeling processes such as *de novo* DNA methylation in plants (Matzke, et al. 2007; Wassenegger, et al. 1994) and diatoms (Veluchamy, et al. 2013), as well as in modifying histones (Kloc, et al. 2008; Volpe, et al. 2002). Evidence for transgenerational epigenetic inheritance, a concept the emerged from Barbara McClintock's discovery of the impact of transposable elements (TEs) on phenotypes in corn, is now well established in plants and animals where it often involves chromatin modifications (Heard and Martienssen 2014). While less is known about microeukaryotic lineages, there is a growing body of literature suggesting that epigenetic processes underlie the structure and function of genomes in diverse lineages.

One hypothesis for the proliferation of epigenetic mechanisms in eukaryotes is that they evolved first to manage genome conflict that resulted from the expansion of TEs and then became coopted for other uses (Fedoroff 2012). Silencing of TEs can be done post-transcriptionally or through heterochromatin formation targeting mobile elements (Aravin, et al. 2001; Klenov, et al. 2007), and both require epigenetic mechanisms that are now deployed more generally throughout the genome. As described below, several eukaryotic lineages have managed to reduce the negative impact of TEs through developmentally regulated genome rearrangements, which include the loss of 'germline-specific' genome sequences during the generation of somatic nuclei (Wang and Davis 2014). Other lineages have coopted epigenetic mechanisms to regulate gene expression and nuclear architecture (Espada and Esteller 2007; Landeira and Navarro 2007).

Here we describe the links between epigenetic mechanisms and the diversity of genome architectures in lineages from across the eukaryotic tree of life. Available data are most abundant for plants, animals and fungi, and we discuss only select data from these multicellular lineages as reviews exist to cover many topics within these clades (Diez, et al. 2014; Feng, et al. 2010; Slotkin and Martienssen 2007). Data from the rest of the eukaryotic tree of life are patchy, and come largely from model lineages (e.g. ciliates), and parasites and pathogens (e.g. *Entamoeba, Plasmodium, Phytophthora).* We are confident that examples of the roles of epigenetic processes in shaping genomes will only expand as poorly-sampled lineages receive greater scrutiny. We also believe that the value of this review includes highlighting the exceptions to biological principles (e.g. the concept of a static genome within species) that emerge from studies of diverse eukaryotic lineages.

1.2.1 Diversity of Eukaryotic Genome Contents

Understanding the impact of epigenetic processes in eukaryotes requires an appreciation of the tremendous variation in size and content of eukaryotic genomes (Fedoroff 2012). This is perhaps best exemplified by the C-value paradox whereby genome size is highly variable and does not obviously correlate with any measure of complexity, particularly in eukaryotes (Cavaliersmith 1978; Fedoroff 2012; Gregory 2001). Among eukaryotes, size variation can be extreme with genomes ranging from only 2.3 Mbp in the microsporidian fungus *Encephalitozoon intestinalis* (Opisthokonta; Fungi; Corradi, et al. 2010), 3 Gbp in *Homo sapiens* (Opisthokonta; Metazoa; Morton 1991), to over 20 Gbp in the gymnosperm *Pinus taeda* (Loblolly pine; Plantae; Wegrzyn, et al. 2014) and an estimated 670 Gbp in the *Amoeba dubia* (Amoebozoa; Friz 1968).

Variation in the number of TEs is one factor that contributes to variation in genome sizes, with the proportion of transposable elements comprising more than 50% of the genome content in some lineages (Fedoroff 2012). Transposable elements are rare in other lineages including the ancient-asexual Bdelloid rotifers (Opisthokonta; Metazoa; Arkhipova and Meselson 2005) and the somatic macronuclei of ciliates (SAR; Coyne, et al. 2012) where they comprise less than 10% of the genome.

1.3 DNA Elimination in Establishing Somatic Genomes

One example of epigenetic control of eukaryotic genome structure can be seen in the purging of portions of the genome during the development of somatic nuclei. This distinction between germline and somatic nuclei defines both animals and ciliates, and is also found in a subset of foraminifera (Figure 1.1; Katz 2001).

1.3.1 Distinct Germline and Somatic Genomes in Animals

Beyond simply differing between haploid and diploid, multiple non-sister animal lineages generate somatic genomes with distinct contents that often includes reduced levels of TEs and other repetitive elements (Figure 1.1; Wang and Davis 2014). During early animal development, the germline genome is physically sequestered into specialized tissues where it often remains heavily heterochromatinized for much of the life cycle (Maatouk, et al. 2006; Robert, et al. 2005). The loss of germline-specific DNA, also described as chromatin diminution, has been documented in a diversity of non-monophyletic animal lineages (Wang and Davis 2014) and molecular details have been worked out in ascarid worms (Bachmann-Waldmann, et al. 2004), copepods (Drouin 2006), and in early-diverging vertebrates (i.e. hagfish and lampreys) (i.e. hagfish and

lampreys; Kohno, et al. 1998; Nakai, et al. 1991; Smith, et al. 2009). In copepods, for example, the zygotic genome expands through successive rounds of endoreplication and/or TE proliferation (Drouin 2006; Sun, et al. 2014; Wyngaard, et al. 2011), which is then followed by large-scale elimination of germline-limited sequences (Wyngaard, et al. 2011). In *Cyclops kolensis* (Opisthokonta), the genome is amplified from ~ 1 Gbp up to ~75 Gbp (Wyngaard, et al. 2011). Recently, Sun et al. (2014) sequenced portions of both the somatic and germline genomes of *Mesocyclops edax* (Opisthokotna) revealing that TEs are rare in the somatic genome, and younger (i.e. less degenerate) TEs appear to be more effectively eliminated (absent) from the somatic genome (Sun, et al. 2014). Given the broad distribution of examples of DNA elimination during the formation of somatic nuclei in lineages across the animal tree of life (Wang and Davis 2014), we suspect that this process may be even more widespread and may have evolved as a means of managing the genome conflict introduced by the invasion of TEs.

1.3.2 Distinct Germline and Somatic Genomes in Ciliates

Ciliates are marked by the presence of distinct germline and somatic genomes within a shared cytoplasm. Because of mechanistic similarities in some elements of chromosome processing, Klobutcher and Herrick (1997) argued that nuclear dualism in ciliates arose as a means of eliminating TEs from the somatic genome (Figure 1.1; SAR). The somatic macronucleus harbors gene-rich chromosomes that are the result from developmentally regulated genome processing following conjugation (i.e. sex). These processes include DNA elimination, genome rearrangements and genome amplification (Jahn and Klobutcher 2002; Prescott 1994). In contrast, the germline micronucleus is enriched in repetitive regions that interrupt gene-coding regions (Coyne, et al. 2012;

Prescott 1994). Many of these repetitive regions harbor signatures of TEs, suggesting that an ancient proliferation of TEs was counterbalanced by the evolution/cooption of mechanisms for DNA elimination of germline-limited sequences during somatic development (Klobutcher and Herrick 1997). For example, a domesticated PiggyBac transposase (i.e. PiggyMAC) is responsible for excision of germline-limited DNA, effectively deleting TEs from the somatic genome.

The molecular mechanisms behind genome reduction have been worked out in some ciliate lineages and involve a suite of epigenetic players (Chalker, et al. 2013; Liu, et al. 2007; Mochizuki, et al. 2002). In the model ciliate *Tetrahymena thermophila* (SAR), which only eliminates ~30% of its germline genome, small RNAs are enriched in germline specific sequences and are believed to serve as scan RNAs during the development of the somatic nucleus (Mochizuki and Gorovsky 2004). In contrast, the ciliate *Stylonychia lemnae* (SAR), which eliminates >90% of its germline genome, small RNAs appear to target somatic sequences to be kept (Chalker and Yao 2011). These same small RNAs also contribute to heterochromatin formation, by guiding repressive histone modifications (Liu, et al. 2007) and DNA methylation (Bracht, et al. 2012) in regions to be eliminated.

1.4 Transposable Elements, Epigenetics, and the Potential for Adaptation

The idea that epigenetic mechanisms evolved at least in part as a means of silencing transposable elements is well-established and has been reviewed elsewhere (Lisch 2009; Slotkin and Martienssen 2007; Yoder, et al. 1997). Some well documented examples of epigenetic silencing of transposable elements include: RNA-directed *de novo*

DNA methylation in plants and diatoms (Rogato, et al. 2014; Saze, et al. 2012), repeat-induced point mutations in fungi (Galagan and Selker 2004), and small RNA guided transposon silencing in animals (Halic and Moazed 2009). Despite the ability of diverse eukaryotes to effectively 'purge' or silence TEs throughout development, TEs and their associated processing/silencing in genomes can also play an adaptive role (Fedoroff 2012; Heard and Martienssen 2014; Lai, et al. 2005) and perhaps even influence patterns of speciation (Belyayev 2014). For example, cell-to-cell heterogeneity and life stage specific control of gene expression – both of which are categorized as stochastic developmental variation – are underlain by epigenetic modifications to chromatin and have been argued to be adaptive in lineages as diverse as bacteria, yeast, animals, plants, apicomplexa, ciliates, green algae, slime molds, and choanoflagellates (Cortes, et al. 2012; Levy, et al. 2012; Rouxel, et al. 2011; Vogt 2015). The broad distribution of stochastic developmental variation among lineages of bacteria, archaea and eukaryotes suggests that this phenomenon may have been present in the last universal common ancestor (LUCA; Vogt 2015).

1.4.1 Epigenetic Mechanisms and Expansive TE Burden in Plants

The prevalence of TEs in plants led to the concept that a diverse epigenetic toolkit evolved for genome defense from TEs and viruses (Matzke and Mosher 2014), and that this toolkit has become part of an adaptive, TE-mediated response to stress (Matzke and Mosher 2014b; Molinier, et al. 2006). The diverse suite of epigenetic mechanisms in plants can been attributed to the large portion of genomes comprised of both functional TEs and repetitive elements (i.e. degraded TEs; >80% in some plants such as Zea mays; Plantae; Tenaillon, et al. 2011). Silencing of TEs in plants occurs through RNA-directed

DNA methylation, where transcribed TEs are processed into the small RNAs that guide their own *de novo* methylation (Law, et al. 2011; Matzke and Mosher 2014b). During non-stressed growth, epigenetic proteins ensure the maintenance of heterochromatin and genomic stability in the vast TE rich chromosomal regions (Stancheva 2005; Zilberman and Henikoff 2004).

Evidence for the adaptive impact of TEs in adaptive responses in plants has emerged in recent decades. Upon abiotic stress in *Arabidopsis* (Plantae), TE activity increases measurably, leading to distinct changes in genome organization through both homologous recombination and copy number variation of TEs and protein coding genes (DeBolt 2010; Molinier, et al. 2006). Interestingly, these effects are heritable through multiple generations of progeny, suggesting the possibility that this response is adaptive (DeBolt 2010; Molinier, et al. 2006; Tricker 2015). For example, increased rates of homologous recombination are heritable in *Nicotiana tabacum* (tobacco; Plantae), where stress induces global changes in hypermethylation of DNA and loci-specific hypomethylation that allows for recombination (Kathiria, et al. 2010). It is possible that the impacts of genome rearrangement are adaptive to some individuals due to beneficial changes in gene regulation or even gene copy number (Figure 1.1).

1.4.2 Epigenetic Modifications of Genome Structures in Eukaryotic Parasites

We focus on the role of epigenetics in parasites to exemplify processes in eukaryotic microbes, largely due to the lack of data in non-parasitic lineages; we do recognize that data are beginning to emerge from lineages such as dinoflagellates, stramenopiles and other marine algae (Lin 2011; Lopez-Gomollon, et al. 2014; Maumus, et al. 2011). Epigenetic mechanisms play a role in phenotypic plasticity and in the ability

of parasites to modify host physiology and behavior (Croken, et al. 2012; Gomez-Diaz, et al. 2012; Hari Dass and Vyas 2014; Marr, et al. 2014). Moreover, mechanisms like pathogen-induced chromatin modifications also play a role in bacterial disease (Gomez-Diaz, et al. 2012), suggesting that they may be very ancient.

The apicomplexan parasite *Plasmodium falciparum* (Figure 1.1; SAR), the causative agent of malaria, relies on epigenetic mechanisms to regulate the transcription of genes necessary for its varying life cycle stages (Ay, et al. 2015; Cortes, et al. 2012; Deshmukh, et al. 2013; Gomez-Diaz, et al. 2012; Salcedo-Amaya, et al. 2010). Transitions between life cycle stages in *Plasmodium* is in part driven by post-translational modifications of histones (Cortes, et al. 2012) and in part by large scale reorganization of nuclear architecture (Ay, et al. 2015). *Plasmodium falciparum* also differentially modifies the expression of the *var* genes that underlie antigenic variation through epigenetic modification of histones in small chromatin domains; the *var* genes are located in subtelomeric regions and their expression is regulated both by localized modification of chromatin and position within the nucleus (Cortes, et al. 2012). Epigenetic mechanisms in the apicomplexan *Toxoplasma gondii* (Figure 1.1; SAR) have evolved to alter the behavior of one of their hosts, the rat, to make it less fearful of cats, which are the final hosts for the parasite (Hari Dass and Vyas 2014).

Life cycle variation is also epigenetically regulated in the parasite *Giardia intestinalis* (Figure 1.1; Excavata; Sonda, et al. 2010). Changes in histone acetylation correspond to transition from free-living to encysted states (Sonda, et al. 2010). Another interesting feature about the structure of the *G. intestinalis* genome is the restriction of active retrotransposons to subtelomeric regions (Arkhipova and Morrison 2001). The

variation in the number of retrotransposons (and their recombination) may contribute to the variable karyotypes observed among strains of *Giardia* (Arkhipova and Morrison 2001; Le Blancq and Adam 1998; Poxleitner, et al. 2008). These homologous regions could allow for recombination in the absence of traditional meiosis, providing *Giardia* with an alternative means to generate genomic diversity after the fusion of its two nuclei (Poxleitner, et al. 2008; Ramesh, et al. 2005).

Another disease-causing group of Excavata, the kinetoplastids (e.g. *Leishmania* and *Trypanosoma;* Figure 1.1; Excavata), also deploy epigenetic mechanisms in causing disease (e.g. Leishmaniasis, African sleeping sickness) and evading host immune systems. The genus *Trypanosoma* relies on epigenetic modification of VSG (variable surface glycoprotein) genes to evade host immune systems (Croken, et al. 2012), including inducing homologous recombination of VSG genes nestled in subtelomeric regions. Similar to the *var* genes in *Plasmodium*, changes in nuclear position of the active VSG gene initiate changes in chromatin structure (e.g. chromatin condensation) that lead to differential and mono-allelic VSG expression (Landeira and Navarro 2007). Beyond altering their own genome, the parasite *Leishmania donovani* (the causative agent of leishmaniasis) is able to induce epigenetic modifications in host macrophages that allow for the successful invasion by the parasite (Marr, et al. 2014).

Epigenetics may also underlie karyotype variation in the genus *Entamoeba* (Figure 1.1; Amoebozoa), which includes *Entamoeba histolytica*, the causative agent of dysentery (Weedall and Hall 2011). As in *Giardia*, karyotype variation may be generated by recombination between transposable elements within the genome, and may contribute to the ability of *Entamoeba* to escape host immune systems (Andersson, et al. 2007).

Adding a further layer of complexity, differential methylation of TEs in *Entamoeba* has been linked to varying levels of virulence (Croken, et al. 2012; Kumari, et al. 2011). Together, these data indicate the role the epigenetic toolkit plays in virulence of this human pathogen.

Genome architecture also drives patterns of substitutions in the genomes of some eukaryotic lineages. Oomycetes and some filamentous fungi (Figure 1.1; SAR; Stramenopiles and Opisthokonta; Fungi respectively) have managed to physically partition their genomes into core regions with greater conservation that are interrupted by gene-poor plastic regions (Gijzen 2009; Haas, et al. 2009; Raffaele and Kamoun 2012). This is most apparent in *Phytophthora infestans*, the causative agent in the Irish potato famine, whose 240 Mbp genome is divided unevenly as the regions of conserved 'house-keeping' genes that comprise about 25% of the total genome size. The gene-poor regions that comprise the bulk of the *P. infestans* genome are rich in mobile and repetitive elements and are associated with pathogenicity and epigenetic silencing (Haas, et al. 2009). This division of function within the *P. infestans* genome behaves almost as two functionally and spatially distinct genomes, and is determined by epigenetic mechanisms. RNAi-mediated heterochromatin formation not only controls the activity of mobile elements but also has major impacts on the transcription of nearby effector genes (more than half of all effector genes in *P. infestans* are within <2kb of a TE) where increasing proximity can alter an effector gene's transcription due to the spreading of heterochromatin from targeted loci (van West, et al. 2008; Vetukuri, et al. 2013). The combination of complex epigenetic silencing and the evolutionary impacts of the

repetitive genome on gene evolution (e.g. copy number variation, and recombination) contribute to the incredible virulence of the pathogenic oomycetes.

1.5 Perspective

Epigenetic mechanisms that regulate transposable elements as part of genome defense have been coopted and contribute to the development of diversity across the eukaryotic tree of life. Eukaryotes share a core epigenetic toolkit (though individual components vary among lineages) comprised of proteins and RNAs that regulate histone and DNA modifications, and that enable RNA scanning mechanisms. These epigenetic processes have expanded among eukaryotic lineages and have enabled eukaryotes to explore diverse genomic landscapes. The resulting epigenetic toolkit provides the basis for the dynamic processes that have contributed to the overall diversity and success of eukaryotic lineages.

1.6 Glossary

Endoreplication: Replication of the genome without any following cell division that leads to changes in ploidy.

Heterochromatin: Tightly packed chromatin that blocks transcription from occurring and is associated with histone modifications.

Histone modification: Post-transcriptional modifications of the histone proteins at varying amino acid residues. The most well-known are histone methylation and acetylation, which are often generalized to be repressive and activating modifications, respectively.

Macronucleus: Somatic and transcriptionally active nucleus in ciliates. Contains streamlined chromosomes that leack centromeric sequences and are often gene-rich. In some ciliate lineages, processing of germline chromosomes leads to macronuclei with chromosomes coding for single-genes and that can be highly amplified.

Micronucleus: The germline nucleus in ciliates that is heterochromatinized and has a more traditional genome architecture (e.g. long chromosomes with centromeric sequences). Micronuclear genomes also contain transposable element sequences that sometimes interrupt protein-coding genes.

Stochastic developmental variation: Seemingly random changes in phenotype such as heterogeneity in gene expression among cells. Stochastic developmental variation provides populations with genetic diversity that may allow exploration of adaptive landscapes.

Transposable elements: Regions of DNA that are capable of changing their position in the genome.

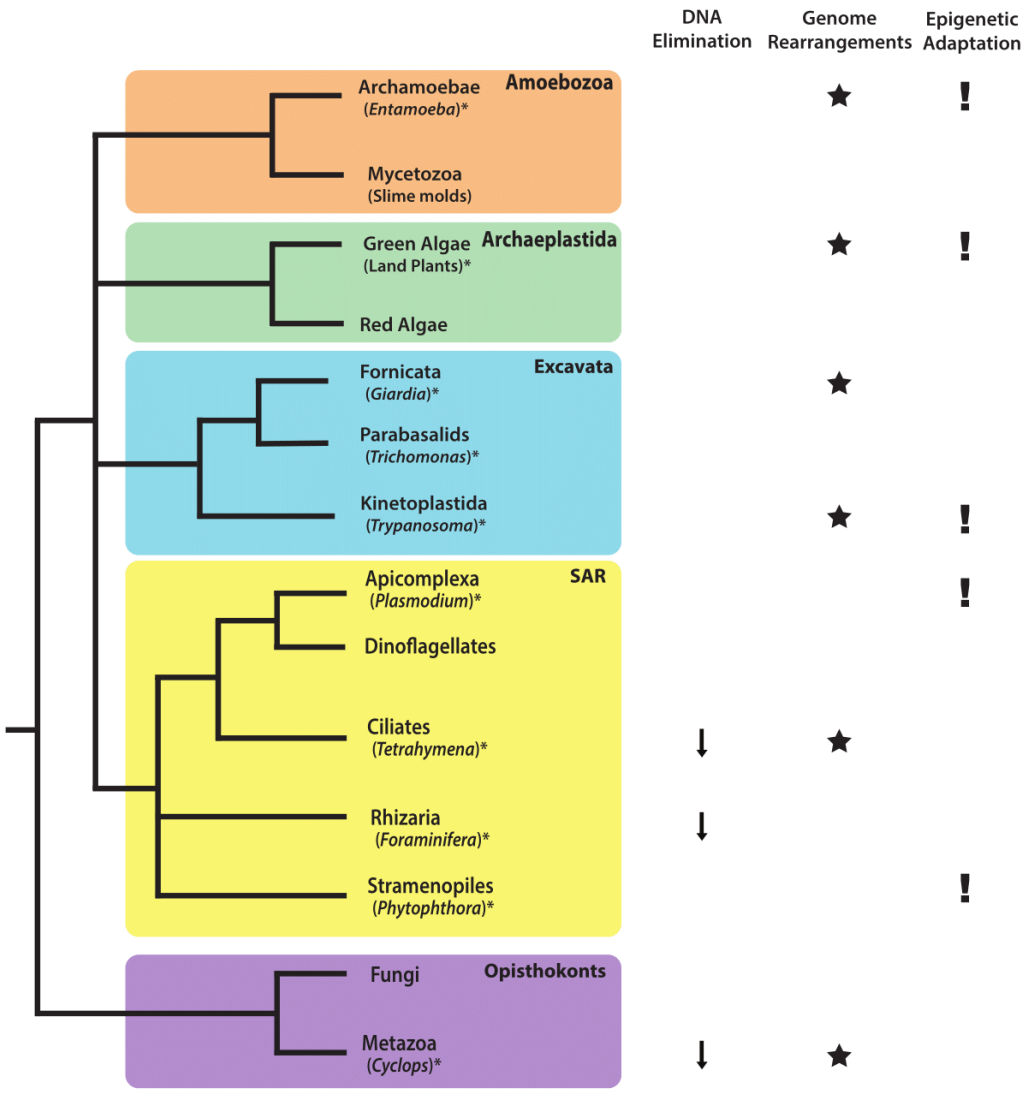


Figure 1.1 Distribution of epigenetic processes across the eukaryotic tree of life. These exemplar epigenetically regulated processes are widespread across eukaryotes. Organisms denoted with '*' are discussed in this review.

CHAPTER 2

NUCLEAR ARCHITECTURE AND PATTERNS OF MOLECULAR EVOLUTION ARE CORRELATED IN THE CILIATE *CHILODONELLA UNCINATA*

2.1 Abstract

The relationship between nuclear architecture and patterns of molecular evolution in lineages across the eukaryotic tree of life is not well understood, partly because molecular evolution is traditionally explored as changes in base pairs along a linear sequence without considering the context of nuclear position of chromosomes. The ciliate *Chilodonella uncinata* is an ideal system to address this relationship between nuclear architecture and patterns of molecular evolution as the somatic macronucleus of this ciliate is comprised of a peripheral DNA rich area (orthomere) and a DNA poor central paramere (i.e. a heteromeric macronucleus). Moreover, because the somatic chromosomes of *C. uncinata* are highly processed into "gene-sized" chromosomes (i.e. nanochromosomes), we can assess fine-scale relationships between location and sequence evolution. By combining fluorescence microscopy and analyses of transcriptome data from *C. uncinata*, we find that highly expressed genes have the greatest codon usage bias and are enriched in DNA poor regions. In contrast, genes with less biased sequences tend to be concentrated in DNA abundant areas, at least during vegetative growth. Our analyses are consistent with recent work in better-studied systems (e.g. plants and animals) where nuclear architecture plays a role in gene expression. At the same time, the unusual localization of nanochromosomes suggests that the highly structured nucleus in *C. uncinata* may create a 'gene bank' that facilitates rapid changes in expression of genes required only in specific life history stages. By using "non-model" organisms like as *C.*

uncinata, we can also explore the universality of eukaryotic features while also providing examples of novel properties (i.e. the presence of a 'gene bank') that build from these features.

2.2 Introduction

Our understanding of the spatial organization of DNA in the interphase nucleus has changed dramatically over the past two decades, largely due to the myriad studies performed on mammalian cell lines (e.g. Cremer, et al. 2001; Kupper, et al. 2007; Tai, et al. 2014). From this work, a model of the interphase nucleus has emerged where decondensed chromosomes are allocated to distinct nuclear regions (i.e. chromosome territories) that are delineated by chromatin poor (i.e. interchromatin) compartments. This chromosome territory-interchromatin compartment model is now accepted as a major organizing principle of the interphase nucleus, due to the widespread conservation of this architecture among animals (Cremer, et al. 2001; Tanabe, et al. 2002) as well as plants, though studies here are more limited (e.g. Fransz, et al. 2002).

Studies of mammalian cells have shown that variations in the radial distribution of individual chromosomes are linked to the morphology of the nucleus itself (Cremer, et al. 2001; Sun, et al. 2000). For example, analyses of "flat" nuclei of fibroblasts reveal chromosomes that are radially arranged by their size such that large chromosomes are found surrounding shorter ones (Cremer, et al. 2001; Sun, et al. 2000). In animal tissues with more spherical nuclei, chromosome distribution correlates best with gene density per chromosome: gene poor chromosomes, often rich in repetitive elements, are typically inactive as heterochromatin and are situated close to the nuclear envelope (Akhtar and

Gasser 2007). Gene dense chromosomes remain euchromatic, occupying the nucleus' center (Kupper, et al. 2007) and are closer to transcriptional foci than expected by chance, supporting the non-random distribution of chromosomes in the nucleus (Meister, et al. 2010). Together, gene density and transcriptional activity likely regulate the position of entire chromosomes (Mahy, et al. 2002). Although based predominantly on a single lineage of eukaryotes, animals, this organization of heterochromatin surrounding a euchromatin core, coupled with the CT-IC model, has become the standard view of the eukaryotic nucleus.

There are few studies examining the nuclear architecture in lineages other than animals and plants, though examples of atypical chromosomes are known. Such examples include the variant surface glycoprotein (VSG) genes found on mini-chromosomes in the parasitic trypanosome *Trypanosoma brucei* (Navarro, et al. 2007), the crystalline chromosomes of dinoflagellates (Bachvaroff, et al. 2014; de la Espina, et al. 2005) and the fragmented and amplified chromosomes found in some ciliates (e.g. Postberg, et al. 2005; Prescott 1994). Despite the presence of unusual chromosomes, Postberg et al. (2005) have suggested that aspects of the CT-IC model also exist in the ciliate *Stylonychia lemnae* and may be a common eukaryotic nuclear feature. The "gene-sized" nanochromosomes in *S. lemnae* form chromatin dense regions, resembling chromosome territories, surrounded by a diffuse chromatin poor network throughout the somatic macronucleus (Postberg, et al. 2005).

Analyses of interactions between nuclear architecture and patterns of molecular evolution (i.e. changes in DNA sequences) are limited and also largely restricted to animal lineages. There is a well-documented relationship between high codon bias (i.e.

strong selection on silent sites) and high levels of gene expression (e.g. Duret 2002; Duret and Mouchiroud 1999; Ma, et al. 2014) but these studies generally do not assess the relationship to nuclear architecture. In *Drosophila*, gene family members residing in euchromatic regions are significantly more biased in codon usage than orthologous members in heterochromatic portions of the same chromosome (Diaz-Castillo and Golic 2007). Such euchromatic regions of chromosomes are typically found in closer proximity to areas of active transcription (Simonis, et al. 2006), suggesting that nuclear architecture may reflect molecular evolution, at least in some animal lineages.

Taking advantage of the presence of nanochromosomes in the somatic macronuclei of *Chilodonella uncinata*, we address the relationship between nuclear architecture and genome evolution. Like other ciliates with extensively-processed somatic chromosomes (e.g. the classes Spirotrichea and Armophorea), *C. uncinata* has a heterochromatin rich germline micronucleus and a spherical macronucleus containing nanochromosomes that are highly and unevenly amplified (Bellec and Katz 2012; Huang and Katz 2014; Radzikowski and Steinbruck 1990; Riley and Katz 2001). Unlike other ciliates whose chromosomes are more diffusely arranged (Foissner 1996; Postberg, et al. 2005), *C. uncinata* and some other members of the class Phyllopharyngea possesses a heteromeric somatic macronucleus comprised of two distinct zones: 1) a DNA rich perimeter (orthomere) consisting of dense chromatin granules close to the nuclear envelope and 2) a DNA poor interior (paramere) with diffuse DNA (Bellec, et al. 2014; Pyne 1978). We combine fluorescent *in situ* hybridization methods and analyses of transcriptomic data to demonstrate the link between *C. uncinata*'s unusual nuclear architecture and patterns of molecular evolution.

2.3 Materials and Methods

2.3.1 Cell Lines and Culture

Chilodonella uncinata (Pol strain, ATCC PRA-257) was cultured in filtered and autoclaved pond water with a rice grain to support bacterial growth at room temperature and in the dark. Prior to fixation cells were collected from culture during exponential growth, centrifuged and then washed in sterile water and kept in the dark overnight.

2.3.2 Transcription Labeling

For pulse labeling of RNA synthesis, *C. uncinata* cells were incubated in filtered and autoclaved pond water containing 1mM 5-ethynyl uridine (EU; Invitrogen) for 30 minutes directly on Superfrost microscope slide (Fisher). Cells were then fixed in 2% paraformaldehyde (Venter, et al.) solution in phosphate buffer solution (PBS) for 30 minutes. Fixed cells were then washed in PBS and permeabilized with 0.5% Triton X-100 for 10 minutes at room temperature. EU labeling was carried out according to the manufacturer's instructions (Invitrogen; Click-iT RNA labeling kits). The cells were incubated in a 1x working solution of Click-iT reaction solution for 30 minutes at room temperature. Subsequently the slides were washed once with Click-iT reaction rinse buffer then once more with PBS. Following this, DNA was counterstained with 0.1 μg/mL 4',6-diamidino-2-phenyl-indole (DAPI) for 1 min in the dark. Cells were then washed twice with PBS and a drop of SlowFade Gold was added prior to sealing with nail polish.

2.3.3 Flourescence *in situ* hybridization (FISH)

Localization of macronuclear α-tubulin, β-tubulin paralogs and nSSU-rDNA genes was performed one at a time using oligonucleotide probes labeled at their 5'-ends with Alexa Fluor 488, 594, or 647. Probe sequences are as follows:

α-tubulin (5'-GTCGTCGATGAGGTCAGAACCGGAACCTACAGACAACTGTTCCAC-3')

β-tubulin P2 (5'-CGCGTGCAAGAGCGGTTTGTGGAACTGATGCGGGTCCGGGCGTAC-3')

β-tubulin P3 (5'-GCAGTCTCGTACTCAAAGCAGCCAGTAGATGGGAACCAAACCTCA-3')

nSSU (5'-CGGAGAGGCTAGGGAACTTTAATCGGAACTCTAGATGACCCAGCA-3')

Cells were fixed directly onto slides as previously described. Cells were then permeabilized in 0.5% Triton X-100 in PBS for 20 minutes at room temperature, washed briefly with PBS and incubated in 0.1 N HCl for 5 minutes at room temperature. Cells were treated with 100µg/mL of RNase One (NEB) for one hour at 37° C before being equilibrated overnight in a mix of 50% formamide in 2x SSC at room temperature. Oligonucleotide probes were dissolved in hybridization buffer (20% formamide, 4x SSC) with 50ng/µL of unlabeled *Chilodonella* DNA. Denaturation of nuclear DNA was performed in 70% formamide/2xSSC at 75°C for 5 minutes. The hybridization mix was denatured separately at 95°C for 10 minutes, snap cooled in an ice bath, loaded onto slides and incubated overnight at 37°C in a moist incubator. Post-hybridization washes

were performed in 2x, 1x and then 0.1x SSC at 42°C. Nuclei were counterstained and sealed as described above.

2.3.4 Confocal Laser Scanning Microscopy

Cells were analyzed using a Leica TCS SP5 confocal laser scanning microscope equipped with an oil immersion 63/1.4 objective lens (HPX PL APO). Fluorochromes were visualized with an UV laser with an excitation wavelength of 405 nm for DAPI, an argon laser with an excitation wavelength of 488 nm for Alexa Fluor 488™ and helium-neon lasers with excitation wavelengths of 594 for Alexa Fluor 594™ and 633 for Alexa Fluor 647™. Images were scanned sequentially, generating 8-bit grey scale images. All images were captured with a resolution of 1024 x 1024 pixels, an acquisition speed of 200 Hz and a line average of 8 to reduce noise. ImageJ (Rasband, W.S. ImageJ. U. S. National Institutes of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij/, 1997-2014) was used to convert 8-bit greyscale images to false RGB colors and for image analysis.

2.3.5 Image Analysis

For each nanochromosome probe and transcription labeling, z-stacks of 50 nuclei that were determined to be most circular by eye were taken for radial measurements (i.e. in 30 degree increments, see methods) using ImageJ. Measurements of fluorescent intensity were taken from the slice with the greatest diameter and the fluorescence profile was taken from the center of the macronucleus towards the nuclear perimeter every 30°. Once all measurements were made, they were normalized against each macronucleus' maximal fluorescent intensity and radial distance (as the size of each macronucleus is

variable depending on cell size) and then were averaged across all 50 nuclei before plotting.

2.3.6 GC3 and ENc Analysis

Calculations of GC content of third position four-fold degenerate sites and the effective number of codons were done through the use of custom python scripts (available: tbd) The analyses made use of the transcriptome assembly of the Pol strain of *C. uncinata* (Grant, et al. 2012) and *T. thermophila* (Miao, et al. 2009).

2.4 Results and Discussion

2.4.1 Transcription is concentrated in chromatin poor areas

We used fluorescent microscopy to assess the distribution of RNA transcripts within the somatic macronucleus of *C. uncinata.* Such analyses must be interpreted in light of the heteromeric nature of the macronucleus in this ciliate: the thousands of somatic nanochromosomes are arranged into a DNA-rich peripheral orthomere and a DNA-poor central paramere. To detect newly-synthesized RNA, we measured the incorporation of the uridine analog EU over a 30-minute interval, revealing that the majority of transcripts accumulate in the central paramere as compared to the peripheral orthomere (Fig. 2.1). These analyses contrast with observations made by Radzikowski (1976), which suggested that transcription was greatest in the DNA rich orthomere as compared to the paramere itself. An explanation for the difference in our findings and those observed by Radzikowski (1976) may be related to the choice of probes and overall technique: after incubation with radioactive uridine for ‘a long time’, the rRNAs that are heavily transcribed likely provided the clearest signal in autoradiographic studies by

Radzikowski (1976) occurring in nucleoli, which are often nestled in close proximity to the orthomere and the nuclear envelope (i.e. DNA poor gaps near nuclear perimeter; Fig. 2.1A, 2.2A). In contrast, our approach reveals the accumulation of transcripts both in putative nucleoli and throughout the large DNA-poor paramere. Moreover, Radzikowski (1976) isolated only nuclei through additional manipulations that altered the morphology of macronuclei (i.e. figure 7 and 8 in Radzikowski 1976), which may also contribute to differences between the studies.

Transcriptional activity corresponds to nuclear architecture in diverse eukaryotes, though the heteromeric nature of nuclei is unique to ciliates within the class Phyllopharyngea (Hausmann and Bradbury 1996; Raikov 1982). In lineages such as animals and plants transcriptionally active regions of chromosomes are either recruited to DNA poor foci of intense transcription (e.g. transcription factories) or near nuclear pores, facilitating rapid exportation of nascent RNAs (Pombo, et al. 1997; Straatman, et al. 1996). In *C. uncinata* there is a large transcriptional neighborhood lacking the distinct foci typical of transcription factories, suggesting that the small size and high abundance of nanochromosomes makes transcription factories unnecessary in *C. uncinata.*

2.4.2 Distinct organization of somatic nanochromosomes

We investigated the spatial distribution of specific nanochromosomes within the heteromeric macronucleus of *C. uncinata*. Using Oligo-FISH (Zwirglmaler, et al. 2003), we captured the spatial distribution of nSSU-rDNA and three protein-coding nanochrosomes using 45-mer probes. Two of these genes, nSSU-rDNA and α-tubulin, represent at leaset an order of magnitude difference in nanochromosome copy number ($5.9x10^4$ and $8.5x10^3$ copies respectively) and relative expression ($5.6x10^5$ and $1.3x10^3$

transcripts respectively) as estimated from qPCR analyses (Bellec and Katz 2012; Huang and Katz 2014). The other two genes, paralogs P2 and P3 of β-tubulin, share similar nanochromsome copy numbers ($6.4x10^4$ and $3.2x10^3$ copes respectively) to the two highly expressed genes, yet have no measureable transcription during vegetative growth (Bellec and Katz 2012; Huang and Katz 2014).

The distribution of highly expressed nSSU-rDNA and α-tubulin nanonchromosomes is distinct from the lowly-expressed β-tubulin paralogs P2 and P3. The highly expressed nSSU-rDNA nanochromosomes are found enriched in the paramere as well as in putative nucleoli nestled within the orthomere (Fig. 2.2A), while highly expressed α-tubulin nanochromosomes have a more uniform distribution throughout the entire macronucleus (Fig. 2.2B). In contrast, both of the lowly expressed β-tubulin paralogs are restricted to the orthomere of the macronucleus (Fig. 2.2C & D), with almost no fluorescent signal measureable in the DNA poor paramere during vegetative growth. Quantifying the distribution of nanochomosomes along the macronuclear radius (i.e. from macronuclear center to envelope), we show that highly expressed nanochromosomes are significantly enriched in the paramere compared to the lowly expressed β-tubulin paralogs (Fig. 2.2 & 2.3). The relationship between the distributions of nanochromosomes is related to the distinct localization of transcription described above. Both of the lowly expressed nanochromosomes (β-tubulin P2/P3) are enriched in the DNA rich orthomere near the nuclear envelope where transcription appears absent (Fig. 2.2C & D, 2.3C & D, 2.4B).

Despite the differences in genome architecture among eukaryotic lineages (i.e. the unique heteromeric arrangement in *C. uncinata),* the recruitment of highly expressed

genes to DNA poor regions appears common across eukaryotes (Navarro, et al. 2007; Osborne, et al. 2004; Postberg, et al. 2006). Postberg et al. (2006) found α-tubulin nanochromosomes in close proximity to DNA poor areas, presumably transcriptionally active, in the somatic nucleus (i.e. macronucleus) of the ciliate *Stylonychia lemnae*. Similarly, highly expressed genes in *C uncinata* are found in the DNA-poor paramere (Fig. 2.2 & 2.3), presumably a means for ensuring that these genes are accessible for transcription. In contrast, nanochromomes with low expression but high copy number that are enriched in the heterochromatin-rich orthomere may serve a skeletal role, maintaining nuclear shape and volume. This structural role is analogous to the positioning of gene-poor and silent loci of animal and plant chromosomes that form the core of chromosome territories (Bickmore and van Steensel 2013; Fransz, et al. 2002) and perhaps also the existence of condensed chromosomes found in interphase in 'core dinoflagellates' (Bachvaroff, et al. 2014).

2.4.3 Transcriptional activity is related to degrees of codon usage bias

We assessed the relationship between patterns of genome evolution and gene expression by examining patterns of codon bias of genes from the published transcriptome of *C. uncinata* (Grant, et al. 2012). Specifically, we examined the relationship between the GC content at four-fold degenerate third positions (GC3s) and codon bias in 974 protein-coding genes. Estimates of GC3s based on the *C. uncinata* transcriptome show a relatively high average GC content (53.6%) in protein coding genes as compared to other ciliates such as *Ichthyophthirius multiformis* (15.9%; Coyne, et al. 2011), *Tetrahymena thermophila* (16.1%; Eisen, et al. 2006), *Stylonychia lemnae* (23.0%; Aeschlimann, et al. 2014), and *Oxytricha triffallax* (24.9%; Swart, et al. 2013). The range

in GC3s for *C. uncinata* (~30-70%; Fig. 2.4A) is very broad compared to protein coding genes among other ciliate lineages such as in *T. thermophila* (~10-25%) and in *O. trifallax (*~15-35%), which may be due to the unusual genome architecture in *C. uncinata*. This variance is also reflected in the codon bias of protein-coding genes in *C. uncinata*, ranging from 27 to 61 (Fig. 2.4A).

Despite the large variance in GC content at four-fold degenerate sites, we found a weaker relationship between codon usage bias and gene expression as compared to *T. thermophila*. To determine this relationship, we examined the correlation between codon usage bias (strength and direction) and expression levels as determined from previous transcriptome data for *C. uncinata* (Grant, et al. 2012; Miao, et al. 2009) and *T. thermophila* (Miao, et al. 2009). Using the number of reads from the *C. uncinata* and *T. thermophila* transcriptomes as a proxy for gene expression reveals that genes that are more highly expressed typically have the greatest codon bias whereas genes with low codon bias appear to be lowly expressed (Fig. 2.4B & C). Transcriptomes of *Tetrahymena thermophila* (Class Oligohymenophorea) have been generated for all major life stages (asexual growth, starvation and sexual conjugation). From these data sets, we examined over 100 protein-coding genes from the available transcriptomes of *T. thermophila* focusing on the relationship between peak expression, and patterns of codon bias (Miao, et al. 2009). Analyses of these genes demonstrate the relationship between peak gene expression and codon bias (Fig. 2.4C; $R = -0.785$, $P << 0.05$); highly expressed genes have great codon bias. The precise relationship between codon bias in *C. uncinata* and expression is weak (Fig. 2.4B; $R = -0.261$, $P = 1.262e\text{-}6$). Unlike *T. thermophila*, transcriptome data for *C. uncinata* are from unsynchronized cultures in which the bulk of

cells are vegetative and ~5% are in conjugation; the lack of synchronized cultures in *C. uncinata* may explain the variability in the relationship between codon bias and expression (Fig. 4B).

Analyses of protein coding genes in animals (Duret and Mouchiroud 1999; Ma, et al. 2014; Zhang and Li 2004), plants (Amanda, et al. 2015; Feng, et al. 2013) and fungi (Duret and Mouchiroud 1999) have shown that codon usage bias correlates with gene expression for many of genes, where highly expressed genes are the most biased in codon usage (Hershberg and Petrov 2008). Greater codon bias in plants and animals is typical of developmentally important genes, highlighting the increased expression of these genes during brief developmental time periods followed by large periods of decreased expression (Chavez-Barcenas, et al. 2000; Schmid, et al. 2005). Similarly, we found that numerous conserved proteins (e.g. histones and macronuclear development protein) in *C. uncinata* comprise the fraction of lowly expressed and highly biased genes in the *C. uncinata* transcriptome. Examination of expression profiles of homologous conserved genes from *T. thermophila* (e.g. histones, elongation factors, epigenetic proteins – DNA methyltransferase) reveal that these genes are often constitutively expressed throughout all major life stages, at relatively low levels, experiencing brief periods of intense expression during specific events, such as conjugation (Forcob, et al. 2014; Miao, et al. 2009).

2.5 Synthesis

Combining analysis of the transcriptome of *C. uncinata* with fluorescence microscopy reveals: 1) there exists a distinct organization of *C. uncinata*'s 'gene-size' nanochromosomes relative expression levels: highly expressed genes are enriched in the

transcriptionally active and DNA poor paramere of the macronucleus; 2) gene expression is linked to patterns of codon usage bias as protein-coding genes with the greatest bias are more highly expressed; and 3) taken together observed patterns of molecular evolution appear to be intrinsically linked to the nuclear architecture of *C. uncinata*. Our conclusions can be combined with insights from other eukaryotic lineages as highly expressed genes are typically under more evolutionary constraint and have significantly fewer nucleotide substitutions at silent sites, a signatures of codon bias (Amanda, et al. 2015; Duret and Mouchiroud 1999; Feng, et al. 2013; Hershberg and Petrov 2008). Highly expressed genes are often found in close proximity to chromatin poor areas or recruited to these areas in numerous eukaryotes, including ciliates (this study; Postberg, et al. 2006), dinoflagellates (de la Espina, et al. 2005; Figueroa, et al. 2014), trypanosomes (Navarro, et al. 2007), plants (Fransz, et al. 2002; Schubert and Shaw 2011) and animals (Mahy, et al. 2002; Osborne, et al. 2004; Pombo, et al. 1997; Postberg, et al. 2006). This interplay between molecular evolution and nuclear architecture may be common to eukaryotes, though it may be more exaggerated in unusual nuclear architectures of lineages such as is found in *C. uncinata*.

We further hypothesize that the heteromeric nuclear architecture in *C. uncinata* provides a 'gene bank' (Fig. 2.5). Under this model, the DNA-rich peripheral orthomere harbors the bulk of high copy number nanochromosomes that have low expression in vegetative cells. By having this envelope of nanochromosomes surrounding the transcriptionally active paramere, there may be rapid transitions in transcriptional states by small-scale adjustments in nanochromosome position in response to developmental and environmental cues (Fig. 2.5). Despite occurring at different scales, *C. uncinata*'s

gene bank shares similarities with the well characterized resting egg banks described in copepods (Metazoa) whereby a large number of dormant eggs can remain viable for large periods of time, becoming active during optimal hatching periods (e.g. Drillet, et al. 2011; Marcus, et al. 1994). Just as these animals essentially move from their egg bank to the water column (upon activation), the gene bank in *C. uncinata* consists of inactive chromosomes that can rapidly move into transcriptionally active areas.

2.6 Acknowledgements

We thank Rachel O'Neill (UConn) for advice in transcription labeling and FISH techniques, and both Judith Wopereis and Nathan Derr (Smith College) for valuable discussion on fluorescent microscopy.

Figure 2.1 RNA transcription is predominantly found in the DNA poor regions of the macronucleus. **A.** Location of transcripts determined with 'click' chemistry (Green – RNA, Blue – DAPI, Yellow – overlay**.** Scale bar is 5 μm. **B.** Distribution of fluorescent intensity estimated radially in 30 degree increments for each nucleus and averaged over 50 cells; nascent RNA (Green) and DNA (Bluemel, et al.)

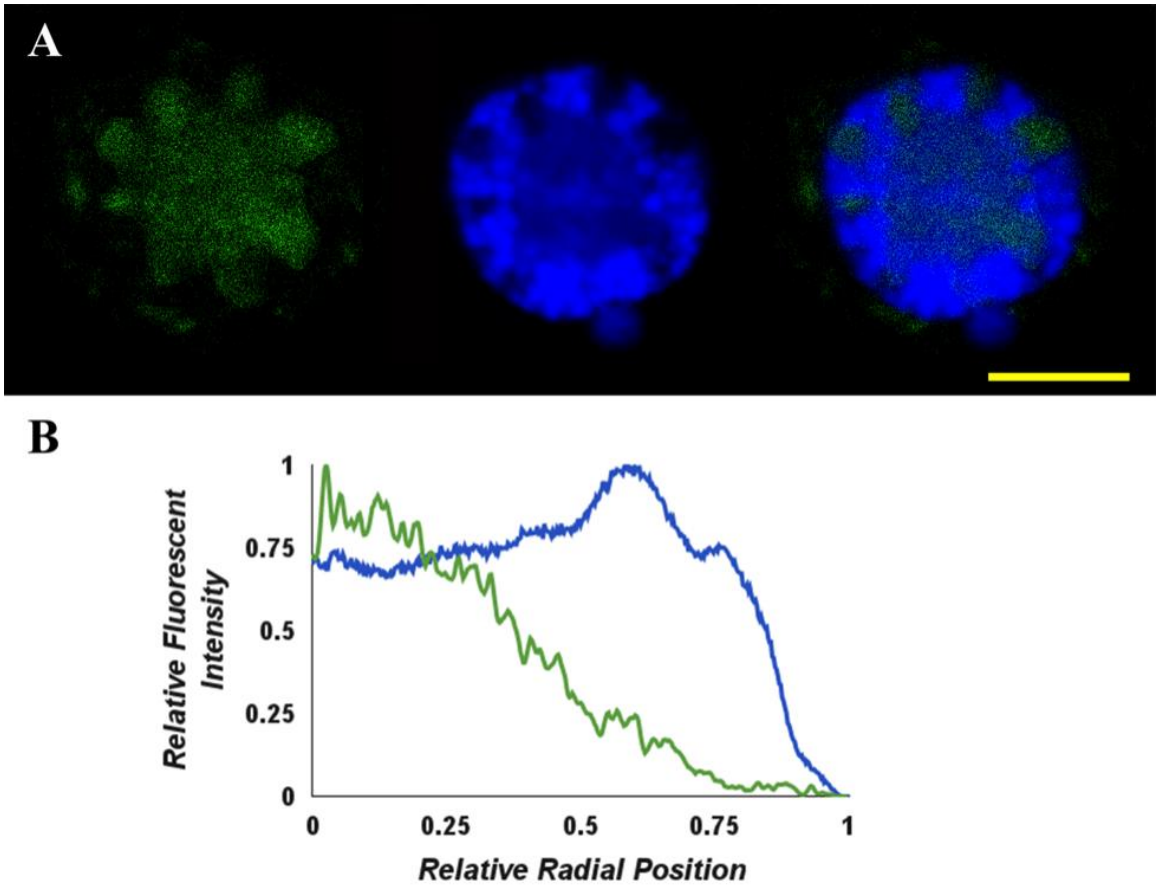
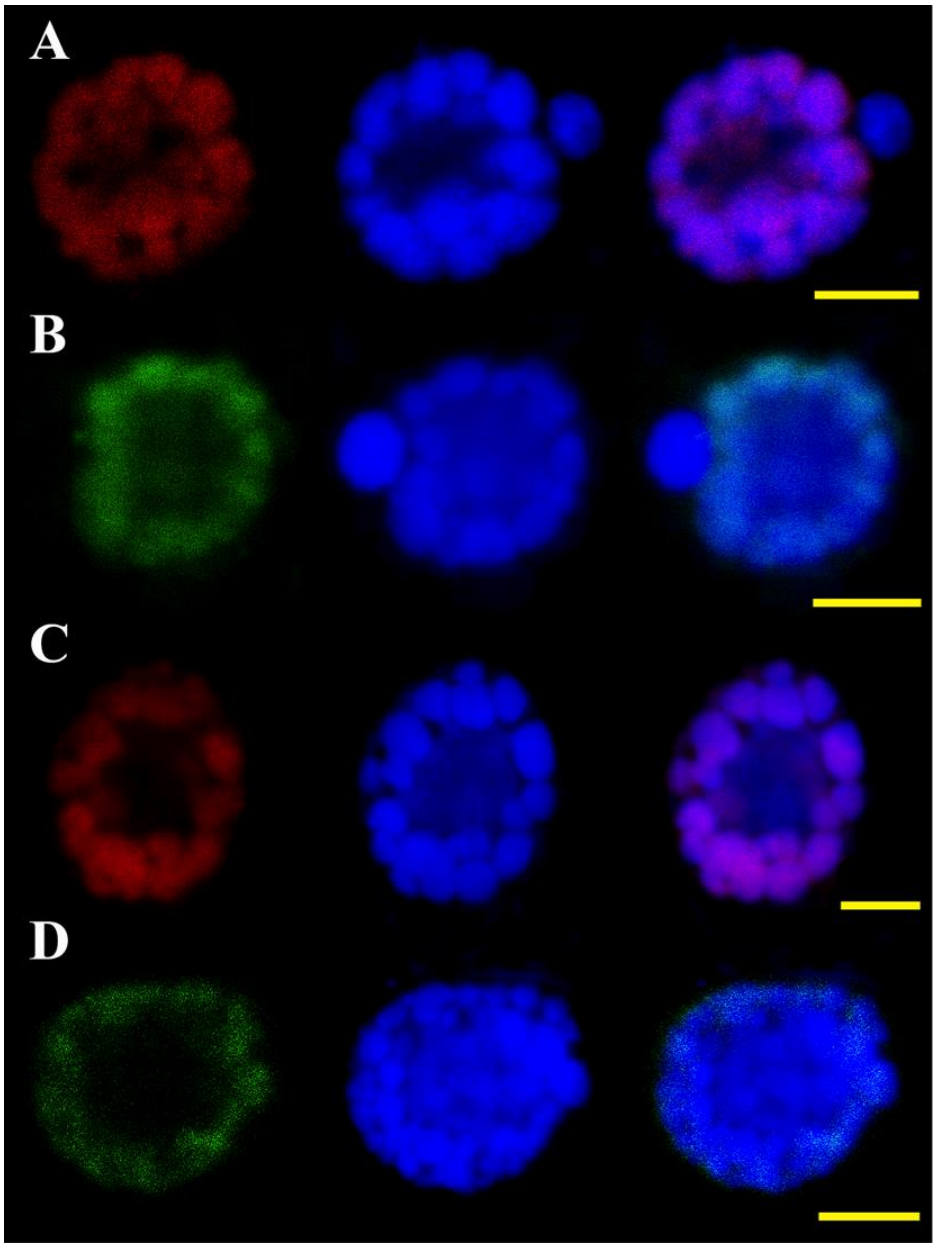


Figure 2.2 Nanochromosomes are distributed non-randomly and in distinct patterns related to levels of expression. **A.** nSSU-rDNA nanochromosomes (red) are found throughout the macronucleus (Blue – DAPI, Purple – overlay). **B.** α-tubulin chromosomes (green) are also distributed throughout the macronucleus despite lower copy number (Blue – DAPI, Yellow – overlay). **C.** Nanochromosomes of β-tubulin P2 (red) are restricted to the orthomere despite similar copy number to nSSU-rDNA nanochromosomes (Blue – DAPI, Purple – overlay). **D.** Similarly β-tubulin P3 chromosomes (green) are also limited to the orthomere of the macronucleus (Blue – DAPI, Yellow – overlay). Scale bars are 3 μm.



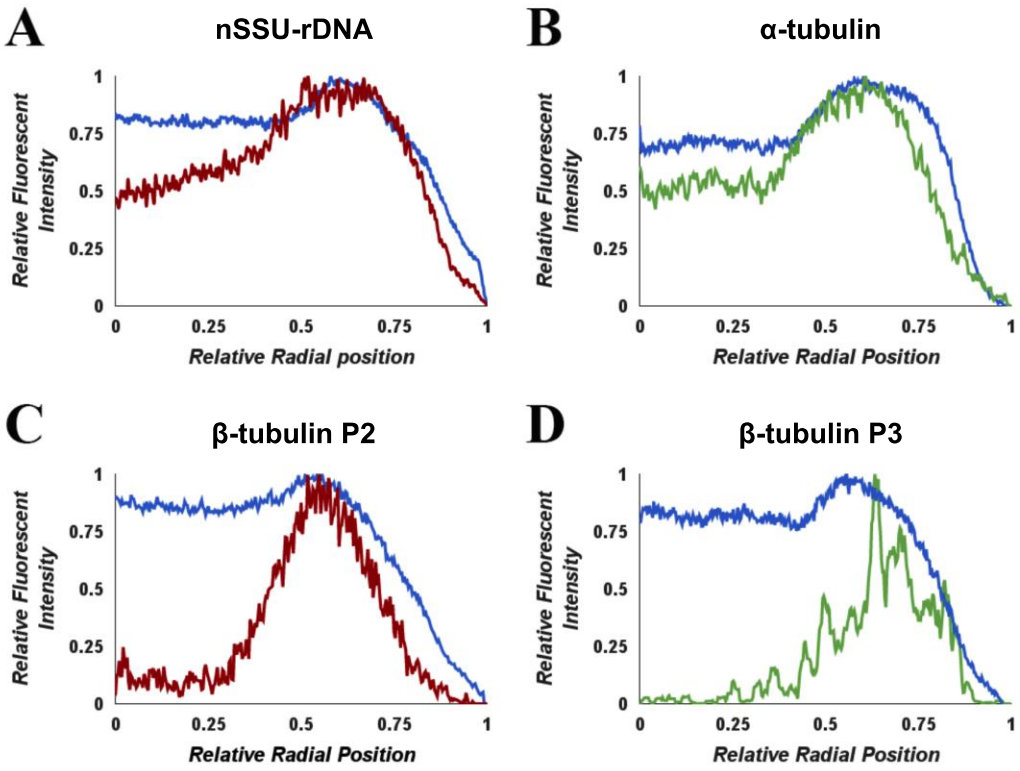


Figure 2.3 Radial distribution of fluorescent intensity of probes in the *C. uncinata* macronucleus shows distribution of nanochromosomes. Fluorescent intensity of nanochromosomes (Red – high copy number A,C; Green – low copy number B,D) and bulk DNA (Bluemel, et al.) are measured along the radius of the macronucleus, from center to the nuclear envelope and at 30 degree increments. **A.** nSSU-rDNA **B.** α–tubulin **C.** β–tubulin P2 **D.** β–tubulin P3.

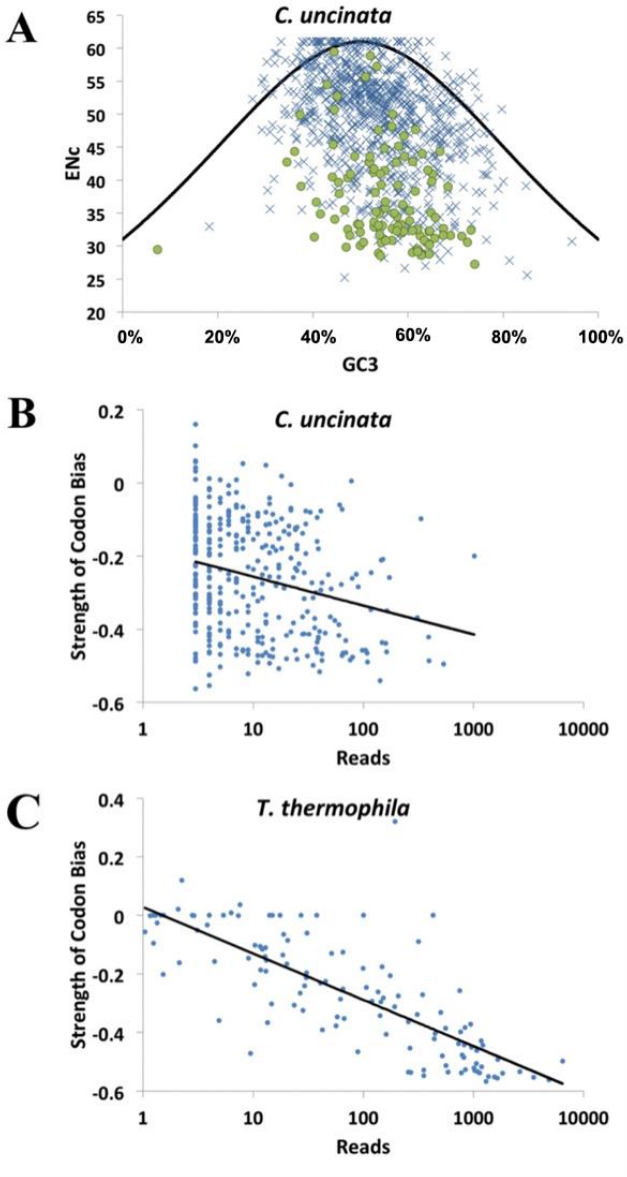


Figure 2.4 Codon bias and gene expression are linked in *Chilodonella uncinata* and *Tetrahymena thermophila*. A. Highly expressed genes (green circles) are typified by greater codon bias than lowly expressed genes (Blue x's). **B.** Vegetative gene expression in *C. uncinata* is somewhat correlated to the degree of codon bias (R = -0.261, P = 1.262e-6). **C.** Peak gene expression in *T. thermophila* is strongly correlated to codon bias (R = -0.785, P << 0.05).

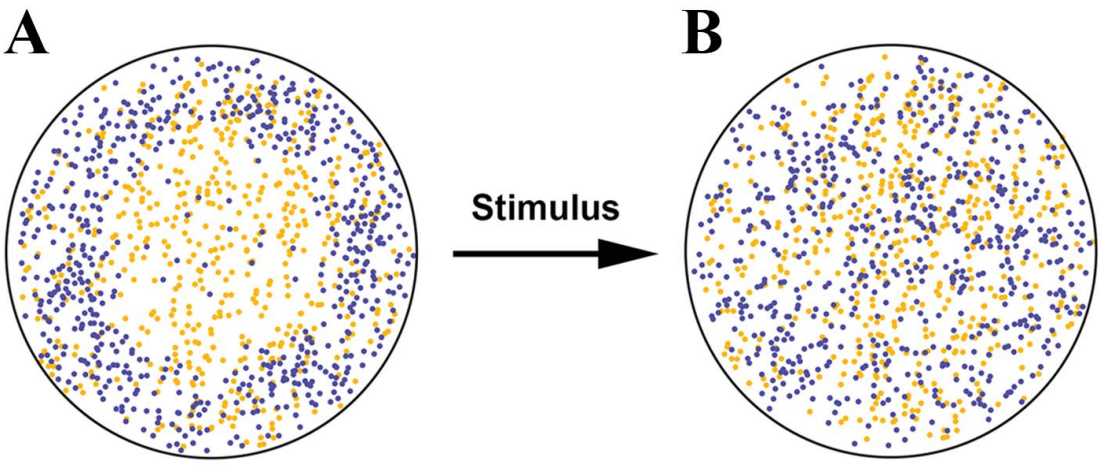


Figure 2.5. We hypothesize that a 'gene bank' in Chilodonella uncinata, whereby genes that are lowly expressed in vegetative cells are concentrated near the nuclear envelope of the marconucleus, permits rapid changes in transcriptional activity in response to environmental and/or developmental cues. A. Transcriptionally active nanochromosomes (Orange) are enriched in the nuclear center, while lowly expressed nanochromosomes (Bluemel, et al.) are mostly distributed near the nuclear envelope (Black) where they comprise the gene bank. **B.** In response to developmental or environmental cues, previously lowly expressed genes (Bluemel, et al.) can quickly move from the gene bank to the transcriptionally active center, resulting in a rapid transition from low to high expression.

CHAPTER 3

EXPLORING THE GERMLINE GENOME OF THE CILIATE *CHILODONELLA UNCINATA* THROUGH SINGLE-CELL 'OMICS (TRANSCRIPTOMICS AND GENOMICS)

3.1 Abstract

Separate germline and somatic genomes are found in numerous lineages across the eukaryotic tree of life, often separated into distinct tissues (e.g. plants, animals and fungi) or distinct nuclei sharing a common cytoplasm (e.g. ciliates and some foraminifera). In ciliates, germline-limited (i.e. micronuclear-specific) DNA is eliminated during the development of a new somatic (i.e. macronuclear) genome in a process that is tightly linked to large-scale genome rearrangements such as deletions and reordering of protein coding sequences. Most studies of germline genome architecture in ciliates focused on the model ciliates *Oxytricha trifallax*, *Paramecium tetraurelia* and *Tetrahymena thermophila* that now have complete germline genome sequences. Outside of these model taxa, only a few dozen germline loci are characterized from a limited number of cultivable species, which is likely due to difficulty in obtaining sufficient quantities of 'purified' germline DNA in these taxa. Combining single cell transcriptomics and genomics, we overcome these limitations and provide the first insights into the structure of the germline genome of the ciliate *Chilodonella uncinata,* a member of the understudied class Phyllopharyngea. Our analyses reveal: 1) large gene families contain a disproportionate number of genes from scrambled germline loci; 2) germline-soma boundaries in the germline genome are demarcated by substantial shifts in GC content; 3) single-cell 'omics' techniques provide large-scale quality germline

genome data with limited effort, at least for ciliates with extensively fragmented somatic genomes. Our approach provides an efficient means to better understand the evolution of genome rearrangements between germline to soma in ciliates.

3.2 Importance

Our understanding of the distinctions between germline and somatic genomes in ciliates has largely relied on studies of a few model genera (e.g. *Oxytricha, Paramecium, Tetrahymena*). We use single-cell 'omics to explore germline-soma distinctions in the ciliate *Chilodonella uncinata,* which likely diverged from the better-studied ciliates ~700 million years ago. The analyses presented here indicate that developmentally-regulated genome rearrangements between germline and soma are demarcated by rapid transitions in local GC composition and lead to diversification of protein families. The approaches used here provide the basis for future work aimed at discerning the evolutionary impacts of germline-soma distinctions among diverse ciliates.

3.3 Introduction

For most 'textbook' eukaryotes, the genome is often viewed as identical in every cell. However, any organism with established germline and somatic cells harbors numerous distinct genomes in part due to the potential differences in ploidy (e.g. N in germline-nuclei compared to 2N in somatic tissues for diploid eukaryotes). Differences between germline and soma extend beyond ploidy with numerous studies documenting the developmental genome rearrangements (e.g. changes in genome architecture) that occur during cellular differentiation into specific tissues such as the V(D)J recombination

in the immune system of vertebrates (Alt, et al. 1986; Mani and Chinnaiyan 2010). Additional examples of types of developmentally regulated genome rearrangements include formation of extra-chromosomal rDNAs and antigen switching in parasites, and such processes are found throughout the eukaryotic tree of life (Li 2015; Maurer-Alcalá and Katz 2015; Nieuwenhuis and Immler 2016; Parfrey, et al. 2008; Smith, et al. 2009; Wang and Davis 2014; Zufall, et al. 2005).

In ciliates, a clade of microbial eukaryotes that is estimated to be about 1 billion years old (Parfrey, et al. 2011), germline and somatic functions are isolated into distinct nuclei within a single cell/individual. As in animals, the germline remains quiescent throughout much of a ciliate's life, only becoming transcriptionally active during conjugation (i.e. sex in ciliates). In *Chilodonella uncinata* (in the class Phyllopharyngea), the germline genome is composed of more 'traditional' chromosomes (Gao, et al. 2015; Gao, et al. 2014; Katz and Kovner 2010) while the somatic chromosomes are present as 'gene-sized' nanochromosomes that are maintained at variable copy numbers. As a result, this ciliate, described as having an extensively fragmented genome, has a somatic nucleus that harbors >20 million nanochromsomes (Bellec and Katz 2012; Huang and Katz 2014; Riley and Katz 2001).

Because of difficulties in culturing and the high level of amplification of somatic genomes compared to the germline (which contributes to contamination in germline DNA preps), traditional methods for sequencing germline-limited DNA are fairly laborious and costly in terms of time and bench work. This has led to limitations in the phylogenetic breadth of explorations of ciliate germline genomes to a few model species, where cultures can provide sufficient numbers of cells (often in the millions) and where

time-tested germline isolation and purification techniques exist. The limitations on the ability to extract quality germline micronuclear DNA with sufficient yields for high-throughput library construction, especially considering the loss of DNA associated with each manipulation and purification (Miller, et al. 1999), has likely been the greatest barrier to studies of germline genomes in non-model ciliates.

The emergence of single-cell 'omics techniques enables us to employ single-cell genomics and transcriptomics for the first large-scale exploration of germline genome architecture in the extensively fragmenting ciliate, *Chilodonella uncinata* in the class Phyllopharyngea. By taking advantage of the biochemical bias in multiple displacement amplification towards large chromosomes (i.e. long template DNA) during whole genome amplification reactions (Gawryluk, et al. 2016; Roy, et al. 2014; Yoon, et al. 2011), we have been able to assemble and explore a substantial portion of the germline genome of *C. uncinata*.

In this study, we demonstrate the power of single-cell 'omics to provide insights into germline genomes in ciliates with gene-sized chromosomes. In addition to providing a summary of general features of the *C. uncinata* germline genome architecture, we use the data generated here and those from other ciliate species to show how dramatic shifts in local GC content distinguish somatically-destined DNA from germline-limited DNA. We also describe how germline genome architecture is associated with gene family size; in *C. uncinata* the largest gene families, which appear *Chilodonella*-specific, are enriched with scrambled genes. This supports the model that scrambling and alternative processing are ways that ciliates increase protein diversity (Gao, et al. 2014; Katz and Kovner 2010).

3.4 Methods and Materials

3.4.1 Ciliate Culturing and DNA extraction

A clonal line of *Chilodonella uncinata* (Pol strain, ATCC PRA-257) was cultured in filtered and autoclaved pond water at room temperature and in the dark, with a sterilized rice grain to support bacterial growth following published protocols (Bellec, et al. 2014; Gao, et al. 2014; Maurer-Alcalá and Katz 2016). Following traditional protocols, micronuclear-enriched DNA extraction started with ~400,000 cells and relied on gel isolation of high molecular weight molecules as described in (Gao, et al. 2015; Gao, et al. 2014; Katz and Kovner 2010). Briefly, after purification of DNA from the agarose gel, the enriched high-molecular weight DNA are digested with Bal-31 for up to 5 minutes, yielding greater micronuclear-enriched DNA that was used for further analyses. Bal-31 is an enzyme that digests double stranded DNA at a rate of ~100bp per minute per end (Rittie, 2008 #10). Given the time required in generating sufficient number of cells, the 5 minute Bal-31 incubation, which equates to ~ 2Kbp of degraded DNA) is our 'best guess' for sufficient MAC degradation with limited MIC destruction.

3.4.2 Single-cell Whole Genome Amplification

For single-cell genomics protocols, we selected 'vegetative' cells (i.e. those not undergoing conjugation or division) from a rapidly growing population. Each cell was washed 5 times in 0.2 µm filtered pond water to dilute any bacteria that may have been carried over. For whole genome amplification (WGA), we placed each cell in an individual sterile 0.2 mL tube and followed the manufacturer's instructions (Qiagen; Repli-g Single Cell Kit, catalog number 150343).

3.4.3 PCR-based Confirmation of Whole Genome Amplification

We utilized the inherent template length bias of the WGA reaction, which better amplifies "long" (<2kbp) template DNA (according to the manufacturer) to selectively amplify the long chromosomes of the germline genome. To confirm these results, we used PCR primers designed to specifically amplify macronuclear or the scrambled micronuclear forms of Actin (based on data from Katz and Kovner (2010); Table 3.5) for all the WGA products. All WGA products sequenced demonstrated substantial enrichment of the micronuclear arrangement of Actin, with no observable amplification with macronuclear-specific Actin primers demonstrating the preference of the WGA reaction for germline DNA templates. In contrast, PCR of the traditional DNA isolation products were far more variable, with substantial PCR amplification of micronuclear Actin as well as some reduced amplification of the somatic arrangement of Actin (as compared to non-Bal31 treated DNA preparations).

3.4.4 Single-cell Whole Transcriptome Amplification

For whole transcriptome amplification (WTA) we followed the same cleaning protocol but also selected individual cells undergoing division (amitosis) and conjugation (sex) within the clonal cultures to assess major variations in transcription. After washing, the WTA reactions were carried out following the manufacturer's protocols (Clontech; SMART-Seq v4 Ultra Low Input RNA Kit), though we used only ¼ reactions. These single-cell transcriptomes (representing 3 major life-cycle stages) were used for our analyses.

3.4.5 Genome and Transcriptome Sequencing

We sequenced three types of material: 1) micronuclear-enriched DNA isolated by gel electrophoresis, 2) WGAs from four individual cells to capture micronuclear DNA, and 3) 12 WTA from single cells (5 vegetative, 3 dividing, 4 in conjugation). The micronuclear-enriched DNA, from gel isolation, was sequenced on a single channel on an Illumina HiSeq2500 at the Yale Center for Genome Analysis. The four individual WGAs were later sequenced on a single channel of an Illumina HiSeq4000 at the Genome Resource Center at the University of Maryland at Baltimore. Libraries of the WTAs were constructed using the NexteraXT kit, following manufacturer's instructions (Illumina) and then sequenced at the IGM Genome Center at the University of California at San Diego on a portion of a single channel of a HiSeq4000. Description of raw data can be found in Table 3.1.

3.4.6 Genome and Transcriptome Assembly

Raw reads for both genomes and transcriptome assemblies were assessed and trimmed using the BBTools (Package (http://sourceforge.net/projects/bbmap; Bushnell 2015) with a minimum quality score 28 and minimum length 125 bp. Following quality trimming, genome data for all four individuals were pooled and assembled using SPAdes (v3.5.0; Bankevich, et al. 2012) and MaSuRCA (Zimin, et al. 2013). As the continuity of the SPAdes assembly was greater than that of the MaSuRCA assembly (determined as the number of transcripts mapped to the assembly per kilobase), we used the SPAdes assembly for all data analyses reported here. Each single-cell transcriptome was assembled independently using rnaSPAdes (v0.1.1) due to the likely heterogeneity in

exact timing for each life-stage. Reads were deposited in GenBank's Short Read Archive (SRA) under BioProject number PRJNA413041.

3.4.7 Preparation of Single-cell Transcriptome Data

Each of the assembled transcriptomes were processed through a series of custom python scripts, which includes updating the name of the transcripts to include their representative life-cycle stage (e.g. conjugation) and the removal of contaminating rRNA and bacterial transcripts (github.com/maurerax/KatzLab/tree/HTS-Processing-PhyloGenPipeline). We then pooled these transcriptomes to remove transcripts of near identity (e.g. > 98% identical) across ≥70% of their length to larger transcripts. This reduced pool was considered as the "core" *C. uncinata* transcriptome that we used for subsequent analyses.

3.4.8 Identification of Putative Germline Loci

To identify germline genome regions, we mapped the prepared 'core' transcriptome (a proxy for macronuclear gene-sized chromosomes) to the long contigs generated from both the gel isolated high-molecular weight DNA (from a culture) and the assembled pool of the four single-cell WGAs. To distinguish putative germline loci from bacterial contaminants in the WGA assemblies, we used AUGUSTUS (v3.2.1; Stanke and Morgenstern 2005) to predict open reading frames under the available *E. coli* K-12 model. Due to the expected complexity in the germline genome architecture of *C. uncinata* (i.e. ORFs tend to contain internally eliminated sequences demarcated by variable pointer sequences; some ORFs being scrambled), complete ORFs should be difficult to identify. For characterization of ciliate germline scaffolds, we considered both lower numbers of ORFs, as well as higher numbers of matches to the core *C. uncinata*

transcriptome: scaffolds ≥ 10kbp with few predicted ORFs and numerous (> 3) mapped transcripts were considered putative germline loci and used for further analyses.

3.4.9 Identification of MDS structure

After identifying a set of putative *C. uncinata* germline (micronuclear) scaffolds, we used BLAST (v2.4.0; Camacho, et al. 2009), with parameters of -ungapped –perc_identity 97 -outfmt 6, to map transcriptome data along germline scaffolds. Custom python scripts (github.com/maurerax/KatzLab/tree/SingleCellGermSoma) analyzed the output from BLAST and categorized the loci and transcriptome data into three broad categories: non-scrambled, scrambled, and unmapped. A range from 30-90% of mapped transcript length was explored, with greater % mapped values biased against scrambled gene data, where 60% of mapped provided the clearest evidence for germline genome architectures. Therefore, only transcripts with ≥60% of their length mapped to the germline assembly were used for subsequent analyses.

To ensure that the single-cell assembly was not generating chimeric scaffolds, we checked read coverage maps for multiple genomic scaffolds associated with different germline architectures (scrambled and non-scrambled). We found no evidence to suggest our assemblies were chimeric (e.g. germline-limited DNA between pointer sequences with abnormally low coverage) and we thus used this assembly for further analyses. To ensure that potential MDS-IES boundaries were not intron-exon boundaries (considering our use of transcripts as a proxy for the somatic genome), to characterize a transcript as harboring an IES, the IES must be flanked by identical pointer sequences and not be nearly identical to the canonical GT-YAG intron-exon boundaries.

3.4.10 Analyses of GC Composition at Germline-Soma Boundaries

To assess GC composition at MDS-IES boundaries, we used the most recent versions of *Tetrahymena thermophila* and *Oxytricha trifallax*'s macronuclear and micronuclear genomes (micronuclear germline assemblies for *Tetrahymena* and *Oxytricha* are available from GenBank under AAGF00000000 and ARYC00000000, with their corresponding macronuclear assemblies: AAGF00000000 and AMCR00000000, respectively). Germline data for *Paramecium tetraurelia* was downloaded from: http://paramecium.cgm.cnrs-gif.fr/. For *Tetrahymena* and *Oxytricha*, telomere sequences were removed and whole macronuclear chromosomes were mapped to their respective germline genome assemblies using BLAST as described above. For *Chilodonella*, we used the BLAST report for confirmed germline loci. For *Paramecium*, transitions from MDSs to germline-limited sequences in the available assembly are marked by the shift from upper-case to lower-case characters which we processed into genomic scaffold coordinates. With the coordinates for these transitions from soma to germline for each taxon, custom python scripts were then used to assess local changes in average GC composition over a sliding 3 bp window with a 2bp step at MDS-IES boundaries.

3.4.11 Identification of Somatic Contamination from Germline Genome Assemblies

For identification (and removal) of somatic chromosomes from our germline genome assemblies, we removed all scaffolds capped with *Chilodonella*'s telomeric repeat: "CCCCAAA" (McGrath, et al. 2007). Specifically, any scaffold with "CCCCAAACCCC" or "AAACCCCAAA" found within the first and last 30bp of the scaffold (allowing for a single mismatch) were characterized as somatic and isolated prior

to our analyses of the germline genome architecture using custom python scripts. These data are summarized in TableS1.

3.4.12 Comparison of Germline DNA Isolation Methods

To compare traditionally-isolated germline DNA (i.e. isolated from cultured cells by gel electrophoresis and treatment with Bal-31 nuclease; following protocols from Gao, et al. 2015; Gao, et al. 2014; Katz and Kovner 2010) to single-cell genome amplification, we evaluated the putative germline assembly sizes for both methods as well as the proportion of the transcriptome data that were mapped to the respective germline assemblies. Because of its superior performance, only the single-cell WGA assembly was used for further analyses; basic statistics and comparisons found in the supplement (TableS1). Statistical analyses comparing different criteria of the different germline DNA isolation approaches were performed using R (v3.2.3; R_Core_Team 2013) and custom python scripts (github.com/maurerax/KatzLab/tree/SingleCellGermSoma).

3.4.13 Gene Family Identification

We used OrthoMCL (v5.0; Chen, et al. 2006) for identification of gene families from the 'core' *C. uncinata* transcriptome using default parameters (minimum similarity: 50%; minimum *e*-value 1e-5). This involves an initial all *versus* all blast, followed by MCL clustering, which ultimately provided a set of gene families and a list of their members. Using custom python scripts, germline mapped members of gene families were binned into different categories (scrambled and non-scrambled).

3.4.14 Estimation of Gene Family Enrichment

To test the distribution of scrambled transcripts' contribution to gene family sizes, we calculated the expected frequency of scrambled members based on the overall

proportion of gene scrambling in the *Chilodonella* germline genome. We used these values to estimate the expected proportions of gene scrambling in each multi-member gene family and a Chi-Square test compared the observed and expected proportions of gene family members that are scrambled. The life-cycle stage (found in the updated transcript names, see **3.4.6 Preparation of Single-Cell Transcriptome Data**) were used to identify the potential enrichment of a given life history stage in a particular gene family.

3.5 Results

3.5.1 Recovery of germline sequences from single-cell 'omics

To explore germline-soma differences we compared the characterization of germline sequences from a 'traditional; gel based method and a single-cell 'omics approach in the ciliate *Chilodonella uncinata*. Our traditional method requires the cultivation of large numbers of cells, total DNA isolation, enrichment for large germline chromosomes, and treatment with Bal31 to remove somatic contaminants; the last step of this process is difficult to optimize given the time required to obtain sufficient number of cells (~2-3 weeks). In contrast, the single cell 'omics approached used the Repli-g Single-Cell kit to amplify the germline genome; here the reliance on the high-fidelity Phi-29 polymerase provided selectivity for larger germline chromosomes over short somatic chromosomes (more in **Methods and Materials**). Our pilot assessment of the traditional DNA isolation and single-cell approaches revealed substantially more 'somatic' contamination in the traditional approach (> 2 orders of magnitude) as measured by the number of assembled scaffolds that are bounded by 1 or more telomeres (Table 3.1).

Similarly, we were able to identify a far greater number of putative germline loci using the single-cell genomic assembly compared to the traditional approach (> 5,000 loci versus ~400 loci; Table 3.1). Given these data, we proceeded to further analyze only the single-cell 'omic derived data.

3.5.2 Patterns of genome rearrangements inferred from germline sequences

To assess the resulting germline sequences, we mapped transcripts, which are a proxy for the gene-sized macronuclear chromosomes of *Chilodonella uncinata*, to putative micronuclear scaffolds generated using single-cell 'omics approach. Using our requirement of ≥ 60% of coverage for each transcript, we mapped 5,019 transcripts (~40% of the total assembled *C. uncinata* transcriptome) to over 32.7 Mbp of the germline genome. A total of 7,448 transcripts remain unmapped to the germline assembly, indicating that additional sequencing effort is required to completely sequence the germline genome. Nevertheless, we estimate the size of the germline genome based on gene-number estimates of ~22,500 from the somatic genomes of *Oxytricha* (Swart, et al. 2013) and *Stylonychia* (Aeschlimann, et al. 2014; distantly related ciliates with extensive fragmentation). Using estimates for overall gene content (~15,000 – 22,500 genes) and our ability to map ~5,000 transcripts across ~33 Mbp (~150 genes per Mbp), we estimate a germline genome size of ~99-149 Mbp for *Chilodonella uncinata.* This estimate will be refined with additional sequencing as we expect variation among ciliates in the proportion of repetitive regions (e.g. microsatellites, transposons and centromeres).

Mapping transcripts allows us to identify the proportion of genes from non-scrambled *versus* scrambled germline loci. Non-scrambled loci are those whose transcripts map to macronuclear destined sequences (MDSs) maintained in consecutive

order, and those lacking evidence of internally eliminated sequences (IES; i.e. germline-limited DNA; Fig. 3.1A). We identify scrambled loci as those meeting two criteria: 1) MDS-IES boundaries with identifiable pointer sequences (i.e. short direct repeats required for unscrambling); 2) MDSs in a non-consecutive order and/or MDSs are found on both strands of the germline scaffolds (i.e. some are inverted; Fig. 3.1B-D). Of these mapped transcripts, we find 3,475 (69%) cases of non-scrambled loci in the germline (Fig. 3.1A; Table 3.2) while 1,544 (31%) loci show strong evidence of scrambling (including alternative processing of germline loci; Fig. 1B-D; Table 3.2).

Scrambled and non-scrambled germline loci differ in several key features (Table 3.2). Scrambled genes tend to be more fragmented in the germline – composed of a greater number of MDSs – than non-scrambled transcripts (3.29 and 2.46 respectively; $p << 0.05$). Moreover, these MDSs are also significantly shorter in length compared to non-scrambled loci (161.0 bp, 212.2 bp respectively; $p << 0.05$). Similarly, scrambled gene loci tend to have longer pointers (8.59 bp, 6.55 bp respectively; $p << 0.05$). We find that the consecutive MDSs of scrambled germline loci (found on the same germline scaffold) are separated by far greater distances than their non-scrambled counterparts (1,454.89 bp, 136.78 bp respectively; $p << 0.05$).

3.5.3 GC composition at MDS-IES boundaries

We examined the distribution of GC content at both small scales, focusing on identifiable MDS-IES boundaries, and broad scales, assessing fluctuations across entire assembled scaffolds. Average GC content at MDS-IES boundaries in *C. uncinata* do not differ between scrambled and non-scrambled MDSs (41.25% and 39.61% respectively; $p > 0.05$; Table 3.2) so we combined these data for further comparisons. By focusing on a

40 bp window on both the 5' and 3' ends of MDSs, we observe a substantial change in GC composition (~12% difference) at MDS-IES boundaries in *C. uncinata*, with greater GC content in MDSs than in the neighboring micronuclear-limited sequences (Fig. 3.2).

We also looked at this small-scale relationship in the few other ciliates with either complete germline genomes (e.g. *Oxytricha trifallax* and *Tetrahymena thermophila*; Fig. 3.2) or with thousands of inferred MDS-IES boundaries (e.g. *Paramecium tetraurelia*; Fig. 3.2). Despite relatively large differences in overall GC content in the germline genome data among these divergent taxa (e.g. ~20.67% in *Tetrahymena* and ~ 49.44% in *Chilodonella*), the boundaries between germline-limited and somatic destined DNA are marked by sharp changes in GC content (~10-14%).

Deploying this knowledge of changes in GC content between germline and somatic regions across broader scales allows identification of coding domains that do not map to our transcript libraries. Given that sharp transitions in base composition likely delineate MDSs from neighboring germline-limited regions among diverse ciliate taxa, we identified regions (>40 bp) in the *C. uncinata* germline scaffolds that had significantly greater or lower in GC content (>2 standard deviations) compared to the average GC content of the assembly. We used BLAST to determine if these regions with extreme composition bias had homologs in other organisms. Of the 250 largest regions with atypically high GC content (average: 1,065 bp), 136 regions (54.4%) have significant BLAST hits (E-value < $1e^{-10}$) to other eukaryotes, predominantly ciliates, whereas only 1 of the 250 largest regions (< 1%; average: 580 bp) with significantly lower GC content has a homolog to another organism (Table 3.3); the functional significance (if any) of regions with very low GC content remain to be discovered.

3.5.4 Gene Scrambling and Gene Family Size Evolution

To assess the impact of gene scrambling on gene family size, we classified the transcriptome data from *C. uncinata* into gene families using OrthoMCL's clustering algorithms (Chen, et al. 2006). We used the number of unique transcripts within a given gene family (referred to as transcript diversity) as an approximation of gene family size given the potential for partial ORFs encoded from the non-exhaustive transcriptomic data. When considering only mapped transcripts, the gene families with the greatest observed transcript diversity are disproportionately composed of transcripts with strong signatures of scrambling (Fig. 3.3). Gene families containing scrambled transcripts are also disproportionately larger (often double in size) than other gene families with ~2.93 members in scrambled gene families compared to ~1.29 members in non-scrambled gene families ($p << 0.05$). Using the observed overall frequencies of scrambled and non-scrambled transcripts (31% and 69% respectively) to generate expected proportions of scrambling, we find that the largest gene families are often significantly more enriched with scrambled gene family members than expected ($p << 0.05$).

3.6 Discussion

In this study, we use single-cell 'omics to compare the germline and somatic genome of *C. uncinata*, and demonstrate that: 1) germline genome architecture and subsequent processing (e.g. DNA elimination, unscrambling, etc) impact gene family sizes and patterns of molecular evolution in the somatic genome; 2) substantial shifts in composition (i.e. GC content) in the germline micronucleus demarcate boundaries between somatic coding sequences and germline-limited DNA; 3) the use of single-cell

molecular approaches is able to provide a robust preliminary look at the germline genome of ciliates with extensively fragmented somatic genomes.

3.6.1 Feasibility and Use of Single-Cell 'Omics' for Germline Genomes

In this study, we demonstrate that single-cell 'omics efficiently provides quality insights into the germline genome architecture of *Chilodonella uncinata*. Currently, the majority of data on germline genome rearrangements and architecture in ciliates is limited to three model ciliates: *Oxytricha trifallax* (Chen, et al. 2014), *Paramecium tetraurelia* (Arnaiz, et al. 2012), and *Tetrahymena thermophila* (Hamilton, et al. 2016). Yet these well-studied taxa come from only 2 of the 11 ciliate classes (cl: Spirotrichea and Oligohymenophorea). Reasons for this limitation have been the inability to gather enough starting material for high-throughput sequencing efforts, as well as potential bioinformatic bottlenecks (e.g. assembly related issues such as low sequencing coverage). Our combination of single-cell genome (from four individual cells) and transcriptome amplification outperformed traditional germline DNA isolation in terms of the number of identifiable germline loci and exploration of general germline features (Table 3.1). Similarly, the gel-isolation based approach for enrichment of micronuclear DNA is also considerably time inefficient, requiring robust and dense cultures (which may be currently impossible for some organisms), whereas the single-cell approaches used in this study can be performed within several days, requiring very few cells and relatively low effort for robust results. Hence, single-cell 'omics methods provide the means to move beyond the confines of the bench and explore the overall complexity and impacts of genome architectures in uncultivable ciliates and perhaps other microbial eukaryotes.

3.6.2 Impact of Germline Genome Architecture on Evolutionary Patterns

Genome architecture and processing (e.g. DNA elimination, genome rearrangements and amplification during generation of somatic chromosomes) appears to play a role in gene family evolution in ciliates. Gao, et al. (2014) hypothesized that the patterns of gene family evolution in ciliates (few unique families with large numbers of members) may be a consequence of genome processing, which is further supported by our analyses of *C. uncinata*'s germline genome. We find that gene families with the greatest transcript diversity are enriched for genes scrambled in the germline. Intriguingly the largest gene families are rich with transcripts present only during conjugation as estimated by single-cell transcriptomic of conjugating pairs (Table 3.4). These large gene families also appear *Chilodonella*-specific as they lack homologs in other eukaryotes.

Compared to other eukaryotic lineages, ciliate genomes tend to be composed of fewer but large gene families (e.g. gene families with > 15 members). For example, the model ciliate *Tetrahymena thermophila*'s somatic genome contains 26,992 protein coding genes comprising 8,826 gene families (3.04 members per family) as estimated from OrthoMCL's gene family clustering. In contrast, other eukaryotes tend to have many more gene families with fewer members. For example, the estimate for *Drosophila melanogaster* is that its 14,422 protein coding genes fall within 12,925 gene families (1.11 members per family; Hahn, et al. 2007), and for *Arabidopsis thaliana* the 25,498 genes fall into 11,601 different gene families (2.31 members per family; Guo 2013).

In *C. uncinata*, estimates of gene family sizes based on our transcriptomic data are consistent with data from *T. thermophila*, with *C. uncinata*'s 12,467 transcripts comprising 4,153 families (3.00 transcripts per family). While this may be an

overestimate for gene family sizes (given the incomplete nature of transcriptomic data), the lack of major differences in gene family sizes between *T. thermophila* and *Chilodonella* is fairly striking as our evidence implicates a close relationship between scrambled germline loci and gene family size. This lack of clarity may be due to the bias in the expansion of *Chilodonella*-specific gene families (through gene scrambling), which would not be accounted for in the above estimates. This may be common among ciliates with highly-scrambled germline genomes, although this may depend on the evolutionary origins of gene-scrambling which remains uncertain.

Although ciliates in the class Phyllopharyngea (e.g. *C. uncinata*) and the species-rich class Spirotrichea (e.g. *O. trifallax, S. lemnae*) harbor scrambled loci, the large-scale arrangement of MDSs in their germline genomes differ. While non-scrambled and scrambled genes are often found interdigitated in germline loci in both *O. trifallax* (Chen, et al. 2014) and *C. uncinata*, the somatically destined DNA in the *O. trifallax* germline genome tends to be present in far more tightly compact genomic 'islands' (Chen, et al. 2014); the degree of proximity is so close that the typical distance between neighboring MDSs is nearly non-existant. From our observations, this is not the case in *C. uncinata* as distances between neighboring MDSs are often relatively large (often > 1kbp apart; Table 3.2). This difference is consistent with the proposed independent origins of germline genome scrambling in these divergent taxa (Katz 2001).

3.6.3 Compositional Bias Demarcates Germline-Soma Boundaries

We demonstrate that MDS-IES boundaries are delineated by rapid shifts in GC content with germline-limited DNA being GC-poor compared to somatic-destined sequences (Fig. 3.2; Table 3.2). Using biases in GC content (MDSs being GC rich and

germline-limited DNA being GC poor) as a tool to understand germline genome architecture, we observe visual evidence for well-known differences in the developmental process (e.g. precision of DNA elimination) among ciliates. For example, almost all IES excision in *T. thermophila* is known to be imprecise and is marked by the greater variability in GC content associated with MDS-IES boundaries within the inferred MDS itself (~10bp from the inferred MDS-IES boundary; Fig. 3.2). However, in *Paramecium tetraurelia*, which undergoes precise IES excision during development, we observe the opposite: a substantial decrease in GC content in much closer proximity to its MDS-IES boundaries (Fig. 3.2).

The role of compositional bias in marking important genomic features has been well described in model plants and animals with major transitions in GC richness associated with transcriptional start sites (Calistri, et al. 2011; Fujimori, et al. 2005) and recombination hot spots (Polak, et al. 2010). As somatic chromosomes in ciliates are far more streamlined (e.g. smaller intergenic regions, lacking centromeres, and intron-poorer genes; Aeschlimann, et al. 2014; Aury, et al. 2006; Eisen, et al. 2006; McGrath, et al. 2007; Swart, et al. 2013), selection may be maintaining the strong clines in GC content associated with MDS-IES boundaries as a means of identifying transcriptionally active sequences (soma) within potentially large regions of non-protein coding DNA (germline-limited DNA). These observations from highly processed ciliate chromosomes are consistent with data from diverse eukaryotes, where GC content in coding domains differs substantially from neighboring intergenic regions (Eichinger, et al. 2005; Haerty and Ponting 2015; Kaul, et al. 2000; Venter, et al. 2001; Zhu, et al. 2009), implicating the role of shifts in GC content as a means for demarcating coding domains despite major

differences in genome architecture (e.g. single-gene nanochromosomes *versus* traditional "long" multi-gene chromosomes).

3.7 Acknowledgements

This work was supported by an NIH award (1R15GM113177) and NSF Go-LIFE (DEB-1541511) to LAK and a Blakeslee award to Smith College. We are grateful to four reviewers for their comments on an earlier version of this manuscript. We also thank members of the Katz Lab for frequent and valuable discussion and members of the Knight Lab for their technical guidance.

Table 3.1. Comparisons of germline genome assemblies based on germline DNA gel-isolation method and single-cell techniques demonstrates superiority of single-cell WGA. Putative germline scaffolds are those with predicted ORFs across < 20% of their length while supported germline scaffolds had at least 3 transcripts that aligned to the scaffold. Somatic contamination (e.g. presence of telomere-containing scaffolds) are far more common and problematic in assemblies of Gel-isolated germline DNA. Similarly, using BLAST to map transcripts to the independent assemblies further demonstrates the superiority of the single-cell approach.

	Gel-isolated DNA	Single-Cell WGA
Number of reads	136,790,808	246,944,949
Number of Scaffolds	49,551	24,881
Putative Germline Scaffolds	420	2,751
Supported Germline Scaffolds	26	1,022
Scaffolds with Telomeres	9,222	57
Mapped Transcripts	468	5,019
Average Scaffold Length	195,422	25,975

Table 3.2. Non-scrambled and scrambled germline loci are substantially different in numerous basic features. All values in parentheses represent the median values for that category.

	Scrambled	**Non-Scrambled**
Mapped transcripts	1,544	3,475
MDSs number	3.29* (4)	2.46* (2)
MDS length	160.96* (133bp)	212.20* (179bp)
Pointer length	8.59 bp* (8bp)	6.55 bp* (6bp)
GC MDS-IES	41.25% (41.09%)	39.61% (39.80%)
Bp between pointers	1,454.89 bp* (805 bp)	136.78bp* (104 bp)

MDS: Macronuclear Destined Sequences (soma)
MDS-IES: Germline-Soma boundaries

'*' denotes significant differences between scrambled and germline loci ($p < 0.05$)

Table 3.3. Top BLAST-hits for largest 250 regions of germline scaffolds without mapped transcriptome data that are significant above or below the average GC content. The majority of these atypically GC rich regions from the *C. uncinata* germline genome had homologs in other eukaryote taxa, predominantly other ciliate taxa and alveolates.

Eukaryote Hit	Germline Regions 2 S.D. Above Mean
Tetrahymena thermophila	43
Paramecium tetraurelia	36
Stylonychia lemnae	17
Oxytricha trifallax	11
Apicomplexa	4
Stramenopila	4
Other	21

Table 3.4. The largest gene families in *C. uncinata* are disproportionately composed of transcripts found during conjugation.

Gene Family	Vegetative	Conjugation	Amitosis
GFam 1	2	145	0
GFam 2	5	64	22
GFam 3	13	46	11
GFam 4	5	16	21
GFam 5	4	24	6
GFam 6	1	17	11
GFam 7	1	19	4
GFam 8	1	17	4
GFam 9	2	10	5
GFam 10	2	8	2
GFam 11	6	4	1
GFam 12	1	6	4
GFam 13	2	6	2
GFam 14	3	5	2
GFam 15	2	4	4
GFam 16	1	2	7
GFam 17	2	6	2
GFam 18	1	3	5
GFam 19	5	3	1
GFam 20	0	6	2
GFam 21	1	7	0
GFam 22	1	7	0
GFam 23	0	4	3
GFam 24	1	3	2
GFam 25	6	1	0

Table 3.5. PCR primers used to discriminate between macro- and micronuclear copies of Actin.

Primer Name	Target Genome	Sequence (5' – 3')
Blue_MAC_Actin_53F	Soma	GGTACCGGTATGATCAAGGC
Actin_1080Rext	Soma	GTGATCCACATYTGYTGRAANGT
Blue_MIC_Actin_164F	Germline	GTACCATTGTCGATGACCACAG
Blue_MIC_Actin_913R	Germline	TTCCAGATCTTCTCCATGTAGTC

Figure 3.1. Exemplar patterns of genome architecture from the germline-mapped transcriptome data of *Chilodonella uncinata*. Germline loci are represented as a single line harboring MDSs (colored rectangles). A) Typical non-scrambled germline genome architecture. B) Exemplar scrambled germline locus. C) Processing of two distant germline loci into single somatic sequence. D) Alternative processing of a single germline locus produces two distinct somatic sequences. Arrows indicate directionality of macronuclear destined sequences.

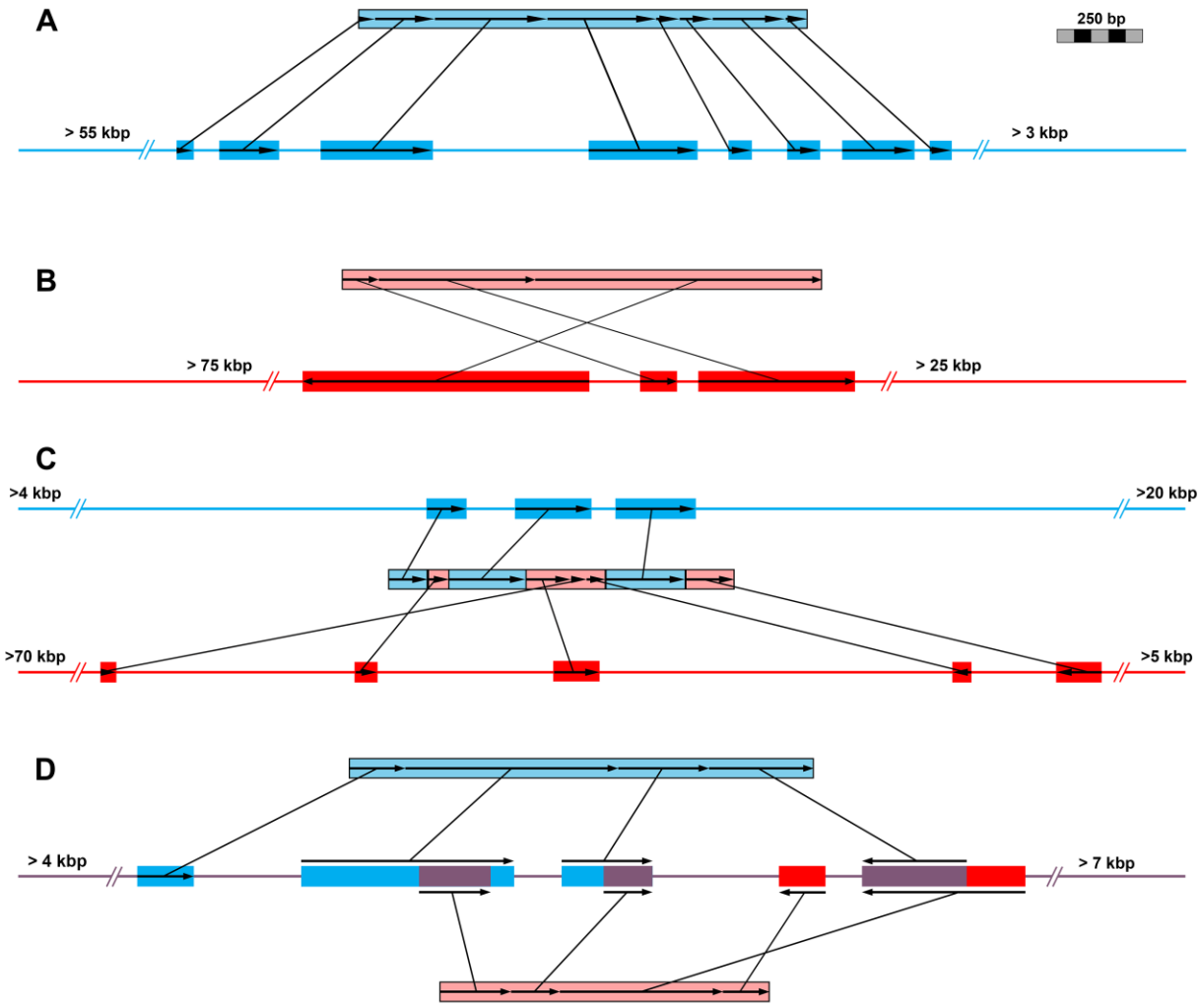


Figure 3.2. Sharp increases in local GC content are associated with germline-soma boundaries in diverse ciliates. Sliding window average (3bp; black) of GC content with 95% confidence intervals (red). Values under taxon names indicate the number of MDS-IES boundaries examined. Data for *C. uncinata* is from this study and data from other ciliates are from GenBank (see methods).

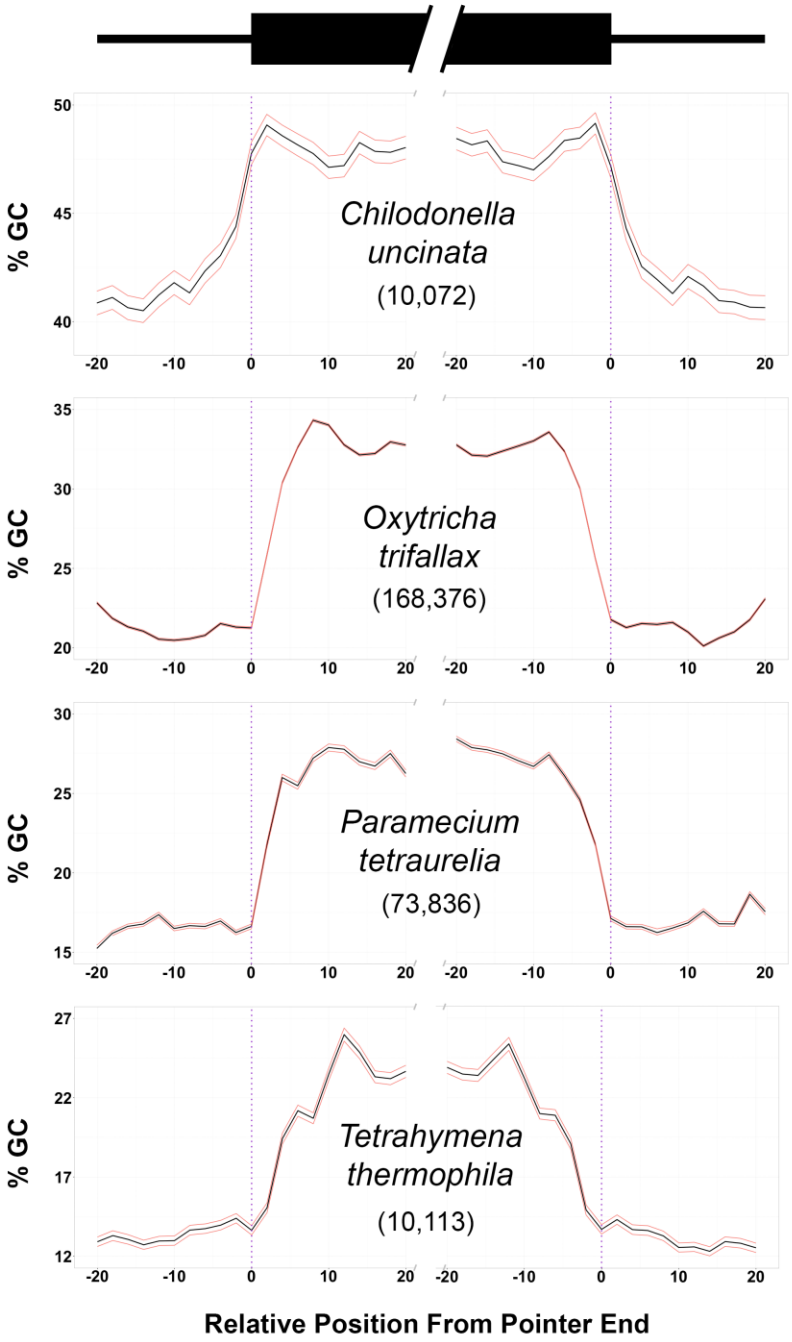
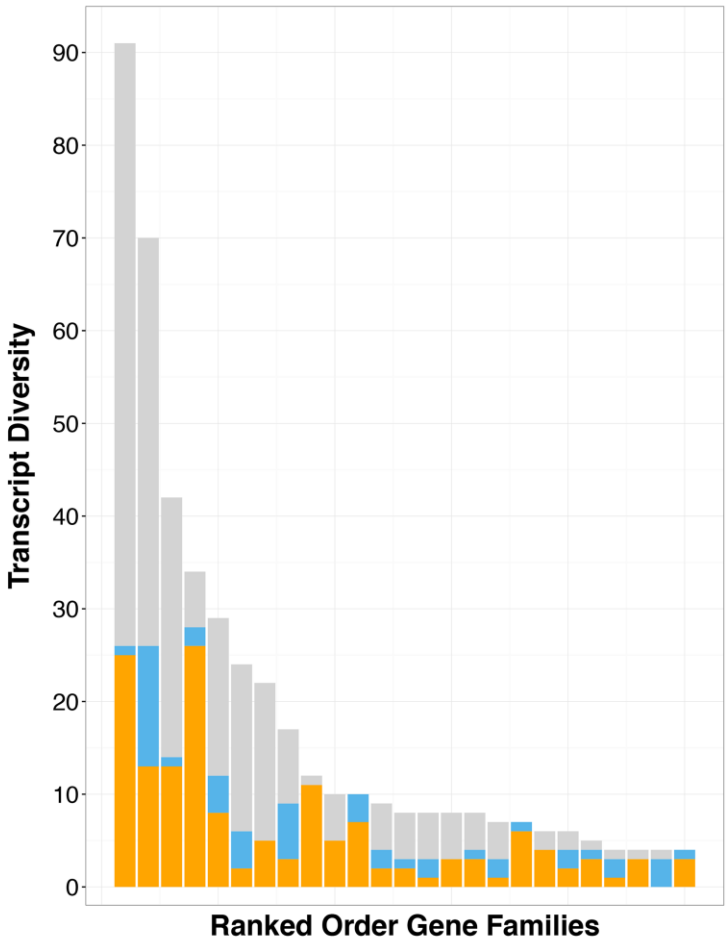


Figure 3.3. *Chilodonella uncinata*'s largest (most diverse) gene families are composed of scrambled genes. Contributions to gene family size by scrambled genes (orange) is typically far greater than non-scrambled genes (Bluemel, et al.), despite the large number of unmapped transcripts (grey). The proportion of scrambled transcripts in each of these large families are significantly greater than expected ($p << 0.05$) given their overall abundance.



CHAPTER 4

TWISTED TALES: INSIGHTS INTO GENOME DIVERSITY OF CILIATES USING SINGLE-CELL GENOMICS

4.1 Abstract

The emergence of robust single-cell ‘omics techniques enables studies of uncultivable species, allowing for the (re)discovery of diverse genomic features. In this study, we use single-cell ‘omics to explore the genome biology of diverse ciliates and to evaluate long-standing assumptions about genome evolution in this ~1 billion year old clade. With these tools, our analyses show: 1) the description of the ciliates in the class Karyorelictea as ‘primitive’ is inaccurate; 2) gene-sized somatic chromosomes exist in the class Litostomatea, consistent with the 1890 observation of giant chromosomes in this lineage; and 3) the presence of gene scrambling in the underexplored Postciliodesmatophora, one of two major clades of ciliates. Together these data highlight the complex germline genome architectures among ciliates. These data also provide the basis for further exploration of diverse ciliates, as well as other microeukaryotes, to evaluate the limitations on our understanding of genome biology built on limited information from model organisms.

4.2 Introduction

Although genomes are often described as being static, where changes in structure (and composition) are presumably catastrophic, there exist a plethora of data demonstrating their inherently dynamic nature (Oliverio and Katz 2014; Parfrey, et al.

2008). In eukaryotes, the focus of this study, examples of dynamic genomes include the separation of germline and soma, and genetic variation throughout life cycles (such as changes in ploidy or DNA content). These changes are often regulated through epigenetic mechanisms and tightly linked to changing developmental stages. Interestingly, many of these epigenetic mechanisms are involved in highly analogous processes between anciently diverged groups of eukaryotes (e.g. RNA-directed DNA methylation in land plants is also found in diatoms; Matzke and Mosher 2014; Rogato, et al. 2014) and has been hypothesized to have evolved near the evolutionary origin of eukaryotes (Maurer-Alcalá and Katz 2015).

Dynamic genomes, including the separation of germline and somatic DNA, are are present ciliates, a group of single-celled eukaryotic microorganisms whose germline and somatic genomes are isolated into distinct nuclei. Unlike multi-cellular organisms, where germline and somatic genomes are in distinct cell-types (e.g. gametes and leaves, hyphae or skin), the germline micronucleus (MIC) and somatic macronucleus (MAC) share a common cytoplasm (Prescott 1994; Raikov 1982). As in other eukaryotes, the germline genome differentiates into a new somatic genome after sex. However, the development of a new somatic genome includes complex epigenetically guided processes (i.e. large-scale genome rearrangements, DNA elimination, chromosome fragmentation, *de novo* telomere addition, and chromosome amplification). In the context of germline-soma and its differentiation during development, the vast focus of the ciliate community is limited to a few model ciliates – *Oxytricha trifallax* (Chen, et al. 2014; Swart, et al. 2013), *Paramecium tetraurelia* (Arnaiz, et al. 2012; Aury, et al. 2006; Guerin, et al. 2017), and *Tetrahymena thermophila* (Eisen, et al. 2006; Hamilton, et al. 2016). All of

these models fall within the 'Intramacronucleata' (referred to as the Im-clade for this study), which is one of the two major clades of ciliates (Fig. 4.1). The other major clade, being the Postciliodesmatophora (the Po-clade in this study), which last shared a common ancestor with the model ciliates over 800 MYA (Parfrey, et al. 2011), and for which very few molecular data have been made publicly available.

Arguably, one of the most notable differences among the model ciliates is the dramatic variation in somatic genomic architecture. In the models *T. thermophila* and *P. tetraurelia*, the somatic chromosomes are 'large' (by ciliate standards) being on average 100's of kilobases to ~ 1-2 megabases in length and are substantially gene-rich (~60-80% of their length composed of open reading frames) and lack centromeres (Aury, et al. 2006; Eisen, et al. 2006). Unlike those ciliates, *O. trifallax*'s somatic genome is predominantly composed of ~16,000 unique tiny 'nano-chromosomes', most of which contain a single ORF (ranging from < 1Kbp to ~ 66Kbp; Swart, et al. 2013). Direct evidence for the phylogenetic distribution of somatic nano-chromosomes is limited to only three ciliate classes: Spirotrichea, Armophorea and Phyllopharyngea (Fig. 4.1).

In addition to variable chromosome number and size distinguishing the somatic genome architecture of ciliates into 2 broad categories ('long' vs 'nano' sized chromosomes), there are differences in patterns of chromosome copy number. For example, in *Tetrahymena thermophila*, each of its 225 unique somatic chromosomes are maintained at ~45 copies each in the somatic nucleus (Doerder, et al. 1992; Eisen, et al. 2006). In the ciliates *Chilodonella uncinata* and *Oxytricha trifallax*, both with gene-sized somatic chromosomes, the macronuclei contain millions of nano-chromosomes maintained at independent copy numbers (Bellec and Katz 2012; Huang and Katz 2014;

Xu, et al. 2012). The range of copy numbers of these chromosomes can span multiple orders of magnitude from several hundred to > 50,000 (Bellec and Katz 2012; Huang and Katz 2014; Xu, et al. 2012). Currently, all evidence suggests that differential chromosome amplification is limited to those ciliates with macronuclear nano-chromosomes (Fig. 4.1).

Ciliates in the Po-clade represent two presumed extremes in genome architectures, which partially 'define' the two ciliate classes within this larger clade. Ciliates in the Heterotrichea are often very large (some species are > 1 mm in length) with correspondingly large somatic nuclei that contain from ~ 1,000 to > 13,000 times more DNA than their germline nuclei (Ovchinnikova, et al. 1965; Wancura, et al. 2017). The second class, the Karyorelictea, can be of similar sizes yet often have numerous clusters of somatic nuclei with relatively low DNA content (~1.1 to 12 times more DNA in their somatic nuclei); based on this observation, Karyorelictea are the only group of ciliates to be dubbed as paradiploid (i.e. nearly diploid) and their name (karyo = nucleus; relictea – relicted) suggests a primitive state (Bobyleva, et al. 1980; Kovaleva and Raikov 1978; Raikov 1985; Raikov 1982; Raikov and Karadzhan 1985). Currently, data on chromosome copy numbers (which could address the hypothetical paradiploidy of Karyolictea) is practically non-existant for ciliates in the Po-clade and is limited even within the far better sampled Im-clade.

The transformation from traditional to fragmented genome architectures (i.e. the development of a new somatic nucleus from the zygotic nucleus) in ciliates relies on the elimination of germline-limited DNA (i.e. internally eliminated sequences; IESs) and the accurate 'assembly' of functional somatic chromosomes (i.e. macronuclear-destined

sequences; MDS). The removal of IESs during the development of the somatic genome is analogous to a 'permanent' intron-splicing during mRNA maturation, as IES excision occurs within the DNA (Allen and Nowacki 2017; Jonsson, et al. 2009; Wahl, et al. 2009). The organization of MDS/IES in germline genomes fall into two major categories: scrambled and non-scrambled. We define non-scrambled germline loci as those with MDSs that are on the same DNA strand and joined in 'order' during DNA elimination in ciliates (Fig 4.4A). In contrast, scrambled germline loci are characterized by MDSs being found on opposing DNA strands and/or in non-consecutive order (Fig. 4.4B). Germline scrambling has only been documented in the Phyllopharyngea and Spirotrichea clades (Fig. 1; Ardell, et al. 2003; Chen, et al. 2014; Gao, et al. 2015; Katz and Kovner 2010; Maurer-Alcalá, et al. in review; Wong and Landweber 2006).

The details on germline genome architecture and the transformations that underlie the development of the somatic genome have only been deeply explored in only four ciliates, representing only three classes of ciliates (Oligohymenophorea, Phyllopharyngea, Spirotrichea; Fig 4.1). Taking advantage of single-cell genomics and transcriptomics technologies, we explore the genomes of *Blepharisma americanum* and several *Loxodes* spp. (Po-Clade) and the large *Bursaria truncatalla* and voracious predatory ciliate *Didinium nasutum* (Im-Clade), capturing ~800 MYA divide between the Im and Po clades (Fig. 4.1; Parfrey, et al. 2011). We also emphasize the necessity for focused work on non-traditional models as the data we present here demonstrate a greater diversity of genomic architectures than has been expected.

4.3 Materials and Methods

4.3.1 Ciliate Culturing and Isolation

Blepharisma americanum, Bursaria truncatella, and *Didinium nasutum* cultures were ordered from Carolina Biological whereas *Loxodes* spp. were collected from a small pond in Hawley Bog (Hawley, MA; 42°35'N, 72°53'W) by collecting water at the sediment-water column interface. From these wild-caught *Loxodes* spp., we observed two dominant morphospecies which we used for our analyses in this study. Cultures of *B. americanum* were maintained in filtered pond water with a sterilized rice grain to support bacterial growth. For isolation, individual cells were picked from cultures and then washed through a series of dilutions with filtered pond or bog water to dilute any contaminating bacteria and micro-eukaryotes that may have been carried over with the cell.

4.3.2 Total DNA Extraction

For *Blepharisma americanum*, approximately 1,300 cells were collected on a 10 µm filter and rinsed thoroughly with filtered pond water. DNA extraction from the filter was done using the ZR Soil Microbe DNA MiniPrep kit (Zymo Research, catalog number D6001) following the manufacturer's instructions. The eluted gDNA was stored at -20ºC prior to the qPCR analyses performed, described below.

4.3.3 Single-cell Whole Genome Amplification

For whole genome amplification (WGA), each washed cell was placed into a minimal volume of media in an individual sterile 0.2 mL tube containing 1 µL of molecular grade water. For each morphospecies this was done in triplicate. Cells lysis and genome amplification were then carried out following the manufacturer's instructions

(Qiagen; Repli-g Single Cell Kit, catalog number 150343). Of the resulting WGA products, we selected the most robust products (e.g. with the best amplification plots over time) for high-throughput sequencing and subsequently used in our analyses. In the end, we used a single WGA product for *B. americanum*, *B. truncatella* and *D. nasutum*. For the 2 distinct *Loxodes* spp. morphospecies, several WGAs were produces, although only 2 WGA products for each of the morphospecies were used in our study. Of *Loxodes* WGAs, only a portion of a single WGA product for each morphospecies was used for high-throughput sequencing, but all four products were used for the qPCR analyses in this study (detailed below).

4.3.4 Whole Transcriptome Amplification of Individual Cells

For the morphospecies with successful whole genome amplifications, freshly isolated (and washed) individual cells of the same morphospecies were placed in a minimal volume of their media in individual sterile 0.2 mL centrifuge tubes containing 1 µL of molecular grade water. The WTA reactions for each of the cells, followed the manufacturer's protocols (Clontech; SMART-Seq v4 Ultra Low Input RNA Kit, catalog number 634888) adjusting all volumes to ¼ reaction volumes. For *B. americanum*, 5 WTA products were prepared, 3 of which were from 'typical' individuals from a log-phase culture and the remaining 2 from 'giant' individuals with obvious signs of predation on other *B. americanum* (e.g. bright red vacuoles). For *B. truncatella*, *D. nasutum*, and each of the 2 morhospecies of *Loxodes*, 2 WTA products from 'vegetative' individuals (e.g. no apparent signs of conjugation, division or gigantism) were used for downstream analyses. Overall 13 WTA products were sequenced and used in this study.

4.3.5 Library Preparation, Genome and Transcriptome Sequencing

Libraries of the amplified WGAs and WTAs were constructed using the Nextera XT DNA Library Preparation kit, following the manufacturer's instructions (Illumina). The prepared libraries were sequenced at the IGM Genome Center at University of California at San Diego on a portion of a single channel of a HiSeq4000. For *Loxodes* spp., WGA and WTAs were also later sequenced at the IGS Genome Resource Center at the University of Maryland on a portion of a single channel of a HiSeq4000.

4.3.6 Genome and Transcriptome Assembly

The raw reads from all data sources were processed using BBDuK (Bushnell 2015) with a minimum quality score of 24 and minimum length 120 bp. Single-cell genomes were assembled with SPAdes (v3.10.0; Bankevich, et al. 2012) using the single-cell and careful parameters. For *Loxodes* spp. WGAs, we pooled the raw reads by morphospecies prior to assembly as they had been re-sequenced at a later date. All single-cell transcriptomes were assembled individually using rnaSPAdes, which is part of the SPAdes package (v3.10.0; Bankevich, et al. 2012), using default parameters.

4.3.7 Post-assembly Preparation of Transcriptome Data

A suite of custom python scripts was used sequenced transcriptomic data generated from our single-cell WTAs (github.com/maurerax/KatzLab/tree/HTS-Processing-PhyloGenPipeline). In brief the processing includes: 1) the removal of contaminating rRNAs and bacterial transcripts; 2) the identification of putative ORFs from the transcripts; 3) the removal of transcripts of near identity (> 98% nucleotide identity) across ≥70% of their length to larger transcripts. For all of our taxa, the pooling

of ‘redundant’ transcripts were performed after we concatenated the assemblies by taxon, resulting in a single ‘core’ transcriptome for each.

4.3.8 Identification of Telomeric Repeats

Prior to the identification of potential telomeric repeats from the taxa whose genomes we partially sequenced, we also downloaded the genomes of *Entodinium caudatum*, *Stentor coeruleus* and *Condylostoma magnus* (NBJL00000000, MPUH00000000, and CVLX00000000 respectively) from GenBank. These additional taxa were downloaded as they represent the only currently available large-scale genomic data from the same classes of ciliates to those in our studies (with the exception of *B. truncatella* no genomic data for members of the Colpodea has currently been released). For all of the genome assemblies, we isolated the first and last 30bp of every scaffold. These scaffold ends were run through MEME (v4.11.4(Bailey, et al. 2009) twice to evaluate the presence (or absence) of repetitive motifs, once without shuffling the sequences of the scaffolds’ ends and the second that did shuffle the sequence. Putative telomeric ends (e.g. significant motifs that were not found in the ‘shuffled’ run of MEME) were only found for *Stentor coeruleus*, *Didinium nasutum*, and *Entodinium caudatum*. Afterwards, we used custom python scripts using these potential telomeric repeats to identify and extract scaffolds that were capped on both ends with telomeric repeats (allowing for a single mismatch; github.com/maurerax/KatzLab/tree/SingleCellGermSoma).

4.3.9 Evaluation of Putative Germline Genome Scaffolds

Genomic scaffolds of the taxa we sequenced in this study that were not capped by telomeric repeats were used to identify putative germline loci that may have been

amplified by the WGA reaction (given its previously demonstrated ability to amplify portions of the germline genome in ciliates (Maurer-Alcalá, in review #13428)). For the Identification of putative germline genome scaffolds and identification of germline-soma architecture, we previously outlined protocols (Maurer-Alcalá, et al. in review). Briefly, this includes identification of ORF-poor genomic scaffolds, alignment of transcripts to those scaffolds and evaluation of common signatures of germline-soma architectures found in other ciliates.

4.3.10 Evaluation of Germline Genome Architecture

After identifying a set of putative germline (micronuclear) scaffolds for *Blepharisma amercianum*, *Bursaria truncatella*, and a single *Loxodes* sp. (due to poor assembly of second morphospecies; fragmented and signatures of contamination), we used BLAST (v2.4.0; Camacho, et al. 2009), with parameters of -ungapped –perc_identity 97 -outfmt 6, to map each taxon's transcriptome data to its germline scaffolds. Custom python scripts (github.com/maurerax/KatzLab/tree/SingleCellGermSoma) analyzed the output from BLAST and categorized the loci and transcriptome data into three broad categories: non-scrambled, scrambled, and unmapped. Based on data from a previous study exploiting single-cell genomics and transcriptomics for analyses of germline architecture, we also only used germline loci where ≥60% of the length of a transcript was successfully mapped for subsequent analyses.

As a precaution to ensure that these loci were more likely germline than soma (which often comprised a substantial proportion of the overall initial genome assembly), we explored the portions of the mapped transcripts that represented transitions from

aligned with the genome assembly to gapped (e.g. genome assembly limited DNA). To be considered a true putative germline sequence these boundaries must not be nearly identical to the canonical GT-YAG intron-exon boundaries (which ciliates possess ref). Similarly, to characterize the genomic-loci as being germline (e.g. harboring an IES), the genome-limited DNA must be flanked by identical pointer sequences that are present at these mapped-unmapped boundaries.

4.3.11 Quantitative PCR Estimates of Copy Number Variation

Quantitiave real-time PCR (qPCR) was used to estimate patterns of gene copy number in *Loxodes* spp. and *Blepharisma americanum.* Ten-fold serially diluted plasmids (1ng/μL to 10^{-7}ng/μL) containing gene fragments of interest were prepared and used to generate the standard curve for each gene. Primers were designed using sequences obtained from both the WGA and WTA products (Table S3) of *B. americanum* and *Loxodes* spp. The DyNAmo Flash SYBR Green qPCR kit (Fisher Scientific, USA) was used for all quantitative PCR experiments in 96-well plates on an ABI StepOnePlus thermal-cycler. Reactions were conducted in a final volume of 20 μL, containing 10 μL 2×master mix, 150nM of each primer, 1 μL of template DNA (at 1ng/μL), and 8 μL of water. qPCR of each targeted gene fragment and WGA sample was performed in triplicate for each experiment. Each experiment was replicated 2 times. We mitigated the potential impact of genome amplification on absolute copy number by assessing relative copy number for each gene of interest by setting the nSSU-rDNA copy number to $1x10^6$ (see Results and Discussion).

4.3.12 Statistical Analyses

All statistical analyses were performed using R(v3.2.3; R_Core_Team 2013). For qPCR data, we used a mixed effects ANOVA evaluating patterns of copy number abundance between and within cells for both *B. americanum* and *Loxodes* spp.

4.3.13 Code availability

All custom python scripts used in this study are available from: github.com/maurerax/KatzLab/tree/SingleCellGermSoma and github.com/maurerax/KatzLab/tree/HTS-Processing-PhyloGenPipeline.

4.4 Results and Discussion

4.4.1 Differential chromosome amplification in Po-Clade

The separation of germline and somatic functions into distinct genomes enables some ciliates to differentially amplify somatic chromosomes. In fact, many eukaryotes extensively amplify their ribosomal rDNA genes (Cohen, et al. 2008; Sinclair and Guarente 1997; Zufall, et al. 2005), so we compare the nuclear small subunit ribosomal DNA (nSSU-rDNA) gene to several protein coding genes. To explore chromosomal amplification in members of the Po-Clade, we analyze chromosome copy number from total DNA (isolated from ~1300 *Blepharisma americanum* individuals), and compare this to copy number estimates from three individual *B. americanum* following whole genome amplification (WGA). These comparisons allow us to evaluate whether the WGA reactions produce significant bias as well as to explore potential inter-individual heterogeneity in chromosome copy number.

In the analyses from both total genomic DNA (total gDNA) and single-cells WGA (sc-WGA) of *B. americanum*, the nSSU-rDNA gene is characteristically high, with $2.55\times10^7 \pm 8.42\text{x}10^6$ copies and $7.90\times10^7 \pm 1.02\times10^7$ copies, respectively. Estimates of copy numbers for protein coding genes between the different preparations of *Blepharisma* (total gDNA and sc-WGA) are similarly consistent, ranging from $1.18\text{x}10^6 \pm 4.38\text{x}10^4$ copies and $8.45\text{x}10^5 \pm 1.14\text{x}10^5$ copies (for one α-tubulin paralog). The least abundant of the protein coding genes from the total gDNA and single-cells are $1.71\text{x}10^5 \pm 1.47\text{x}10^4$ copies and $3.01\text{x}10^5 \pm 3.51\text{x}10^4$ copies respectively (Table 4.1). By setting the nSSU-rDNA copy number to 10^6 (a values based on evidence from diverse ciliates; Gong, et al. 2013; Heyse, et al. 2010; Huang and Katz 2014), we find that the ranges of copy number for chromosomes containing protein coding genes (two paralogs of Actin and α-Tubulin, and EF-1α) in *B. americanum* span ~2 orders of magnitude (Fig. 4.2B) with the exception of actin paralog 2, which is consistently underrepresented across all samples. Despite greater variability in absolute copy numbers from the population of cells compared to the individual cells, we observe no significant biases between methods (total gDNA *versus* single-cell WGA; $p = 0.474$; Fig. 4.2B). In other words, the sc-WGA method provides the means to assess patterns of inter-individual chromosome copy numbers that can approximate entire populations of cells.

We then deployed the same methods to study the uncultivable genus *Loxodes* in the 'paradiploid' class Karyorelictea. We performed a similar qPCR experiment using 5 genes (nSSU-rDNA, EF-1α, Actin, Rs11 and α-Tubulin) from sc-WGAs of wild-caught individuals of *Loxodes* spp., representing two distinct morphospecies. As we only have relative numbers here, we again set the nSSU-rDNA to 10^6 copies to allow comparison of

patterns of chromosome copy numbers with *B. americanum* (raw data in Table 4.2). In contrast to the stochastic patterns of chromosome copy number in *B. americanum*, the differences among copy number for protein coding genes in *Loxodes* spp. consistently spanned a far greater range (~4 orders of magnitude; Table 4.2; Fig. 4.2D). We observe significant differences in gene copy number within each cell of *Loxodes* spp. ($p << 0.05$), implicating the differential amplification of chromosomes. Interestingly, for both of the distinct morphospecies, gene copy numbers are maintained in a mostly-conserved order: nSSU-rDNA >> Actin > Rs11 > α-Tubulin > EF-1α (Fig. 4.2D), which contrasts with the stochastic pattern in Heterotrichea.

The contrasting pattern of stochasticity in chromosome copy number in *B. americanum* and the predictability in chromosome number in *Loxodes* spp. likely reflects differences genome architecture of their somatic nuclei. The macronuclei of *Blepharisma* house large quantities of DNA and possess the ability to divide, while *Loxodes* spp.' macronuclei are DNA poor and do not divide with cell division. The stochasticity in chromosome copy number for *Blepharisma* may be a byproduct of the massive genome amplification that occurs during development (Santangelo and Barone 1987), as the somatic nucleus is estimated to have > 1000x more DNA than the germline nucleus (Ovchinnikova, et al. 1965; Wancura, et al. 2017). Variable chromosome copy number among individuals is likely an inherent feature of *Blepharisma* and its relatives (in the class Heterotrichea; Fig. 1), exemplified by *Stentor coeruleus*, whose chromosome copy numbers of the nSSU-rDNA are clearly correlated to cell size (Slabodnick, et al. 2017) as well as nuclear volume (Cavaliersmith 1978). This suggests that the observed

stochasticity from our measurements rises from a combination of biological differences (e.g. cell volume or life-cycle stages; Fig. 4.2 A&B) and inherent stochasticity.

Although *Loxodes* spp. are found in the sister class to *B. americanum* (both in the Po-Clade), *Loxodes* and its relatives have long been considered as 'primitive' ciliates (Orias 1991; Raikov 1994, 1985; Raikov and Karadzhan 1985). This presumption arose from early studies that found that the somatic macronucleus is unable to divide (needing to be differentiated from a germline nucleus with each cell division) as well as from estimates of DNA content based on autoradiographic measurements from the somatic and germline nuclei of *Loxodes* and its relatives (Bobyleva, et al. 1980; Kovaleva and Raikov 1978). From these early measurements, where the somatic nuclei typically harbor only ~1.1 to ~12 times the amount of DNA compared to the germline nuclei, these taxa were labelled as paradiploid ('nearly-diploid'). This has led to the expectation that the relative copy number among protein-coding genes would be approximately equal in this class of ciliate (Fig. 2C). Such low ploidy is unusual among ciliates; for example, ploidy is species dependent and ranges from ~45N in *Tetrahymena thermophila* (Woodard, et al. 1972) to ~800N in *Paramecium tetraurelia* (Duret, et al. 2008).

Surprisingly, our data demonstrate that *Loxodes* spp. is neither paradiploid nor are all chromosomes equally amplified. Our estimates of relative chromosome copy number show that instead of being present in roughly equal abundance, chromosomes containing our target genes differ by several orders of magnitude (Fig. 4.2C & D). Though non-dividing macronuclei in *Loxodes* spp. (and other membersof the class Karyorelictea) age over time (at most 7 generations; Raikov 1994, 1985; Raikov 1982; Yan 2017), we do not believe aging along is sufficient to explain our data. This is because replicability of

estimates across cells would indicate we picked cells of similar ages, and the high variability across genes would suggest dramatic changes in copy number from diploidy to > 1000 copies in only seven generations (Bobyleva, et al. 1980; Kovaleva and Raikov 1978; Raikov 1985; Raikov 1982; Raikov and Karadzhan 1985). These copy number data suggest that the long-held description of *Loxodes* spp. as 'primitive', based upon DNA content estimates and the inability to divide their macronuclei, are inaccurate.

4.4.2 Unexpected extensive fragmentation of somatic genomes from the Im-Clade

Extensive fragmentation of chromosomes into gene-sized 'nano-chromosomes' during the development of somatic macronuclei is well documented in only three ciliate classes (e.g. in *Chilodonella uncinata* (cl: Phyllopharyngea; McGrath, et al. 2007), *Oxytricha trifallax* (cl: Spirotrichea; Swart, et al. 2013; Xu, et al. 2012), and *Nycotherus ovalis* (cl: Armophorea; McGrath, et al. 2007; Ricard, et al. 2008; Fig. 1). We searched for evidence of extensive fragmentation in the class Litosomatea (Fig. 4.1), analyzing a single-cell WGA assembly for *Didinium nasutum* and the recently released genome assembly of *Entodinium caudatum* (a distantly-related member of the same class). We evaluated the ends of scaffolds for both *D. nasutum* and *E. caudatum*, to look for telomeres as no record of telomeres has been reported for members in this class. This approach resulted in a common strong (and repetitive) motif in both taxa, C_4A_2T. As telomeric sequences seem well conserved over broad phylogenetic scales in ciliates (Aeschlimann, et al. 2014b; Aury, et al. 2006; Eisen, et al. 2006; McGrath, et al. 2007; Swart, et al. 2013), this simple repeat may be Litostome-specific.

To assess the size distributions of somatic chromosomes, we use the telomeric motif to identify scaffolds bounded by repeats at both ends (e.g. complete assembled

chromosomes) for both *D. nasutum* and *E. caudatum*. To our surprise, we identified 321 complete nano-chromosomes in *D. nasutum*'s telomere-bound scaffolds and 7,528 complete chromosomes from the released *E. caudatum* genome assembly. To check that these were not simply assembly artefacts, we mapped transcripts from single *D. nasutum* individuals to the pool of 321 putatively complete chromosomes, for which we observe 316 (98.4%) chromosomes that contain nearly complete transcripts, with 254 (80.4%) chromosomes harboring a single ORF. As no transcriptome data is publicly available for *E. caudatum*, we mapped 5,692 translated ORFs from our *D. nasutum* transcriptome to 5,293 (70.3%) of *E. caudatum*'s complete chromosomes.

Having affirmed the presence of nano-chromosomes in the *D. nasutum* and *E. caudatum* genome assemblies, we find that the size range of these complete chromosomes are nearly identical for both, ranging from ~0.4 Kbp to ~ 26 Kbp, despite differences in the methods used to obtain the genomic data (e.g. use of sc-WGA techniques for *D. nasutum* and more traditional DNA isolation approaches used for *E. caudatum*). Interestingly, previous work using pulsed-field gel electrophoresis of total gDNA from *D. nasutum* did not observe chromosomes below 50 kbp (Popenko, et al. 2015), which suggests that the nano-chromsomes may be present at relatively low copy numbers and/or that the retention of these chromosomes is strongly dependent on the DNA isolation approaches. However, comparisons of the size distribution of these complete chromosomes for *D. nasutum* and *E. caudatum* to genomic data from diverse taxa, demonstrate that these chromosomes' sizes are consistent with the 'gene-sized' chromosomes found in divergent ciliate taxa (e.g. Chilodonella uncinata and Oxytricha trifallax; McGrath, et al. 2007; Swart, et al. 2013; Fig. 3 and Fig. S1).

The data on nano-sized chromosomes in the class Litostomatea are consistent with the 1890 description of giant germline chromosomes (presumably through endoreplication) during development of a new macronucleus (Balbiani 1890). The correspondence between the appearance of giant chromosomes during development and the presence of nano-sized chromosomes in somatic genomes has been extensively documented (most notably Chilodonella and Stylonychia, classes Phyllopharyngea and Spirotrichea respectively; Ammermann 1986; Juranek, et al. 2005; Katz 2001; Katz and Kovner 2010; Postberg, et al. 2008; Pyne 1978; Riley and Katz 2001). In these ciliates, polytenization occurs just prior to the extensive genome remodeling that ultimately leads to the formation of the thousands of unique nano-chromosomes through epigenetically guided DNA elimination, large-scale genome rearrangements and *de novo* telomere addition (Ammermann 1986; Chen, et al. 2014; Fuhrmann, et al. 2016; Postberg, et al. 2008; Pyne 1978; Spear and Lauth 1976). The absence of polytenization of germline chromosomes from the model ciliates *Paramecium tetraurelia* and *Tetrahymena thermophila*, which possess 'large' macronuclear chromosomes (ranging from ~0.2Mbp to several Mbp in size; Aury, et al. 2006; Eisen, et al. 2006; Fig. 3 and Fig. S1), further implicates this step as being limited to nano-chromosome formation.

Well over 100 years ago, Éduoard-Gérard Balbiani, who provided the original description of polytene chromosomes in the dipteran *Chironomus* (Balbiani 1881), described the presence of polytene chromosomes in the ciliate *Loxophyllum meleagris* (a relative of *Didinium* and *Entodinium*; Balbiani 1890). Unfortunately, there had been little work able to corroborate the observations of Balbiani (1890). However, given the possible sister-relationships between the classes Litostomatea, Spirotrichea and

Armophorea, all of which have both nano-chromosomes (Ricard, et al. 2008; Riley and Katz 2001; Swart, et al. 2013) and giant chromosomes (Golikova 1965; Wichterman 1937), these unusual genome architectural features could be a synapomoprhy that unites this portion of the Im-clade (Fig. 4.1). These early observations, coupled to the single-cell approaches to analyses of chromosomes in *D. nastum*, highlight the unexpected presence of nano-chromosomes in the Litostomatea and draw attention to the limitations in data for ciliates outside of model systems.

4.4.3 Germline genome architecture from diverse ciliates

Studies of germline genome architecture in ciliates are phylogenetically limited, predominantly to the few model species in the classes Oligohymenophorea and Spirotrichea (Arnaiz, et al. 2012; Chen, et al. 2014; Gao, et al. 2015; Guerin, et al. 2017; Hamilton, et al. 2016; Landweber, et al. 2000; Maurer-Alcalá, et al. in review; Nowacki, et al. 2008). This has largely been a result of issues surrounding cultivability, as well as the lack of robust methods for the efficient extraction of high-quality germline DNA from uncultivable lineages. To overcome these limitations, we use a combination of single-cell genomics and transcriptomics to gain insights into the germline genome organization of three ciliate taxa, representing members of both the Im (*Bursaria truncatella*; cl: Colpodea) and Po clades (*B. americanum*; cl: Heterotrichea and *Loxodes* sp.; cl: Karyorelictea; Fig. 4.1).

To explore the germline genome architecture of these ciliates, we map transcripts from single-cell transcriptome assemblies to the putative germline scaffolds. By following established methods for identifying and characterizing germline scaffolds (Maurer-Alcalá, et al. in review), we are able to identify numerous putative germline

scaffolds for all three ciliates. For this study, we define non-scrambled germline loci as those where macronuclear destined sequences (MDSs; soma) are maintained in consecutive order (e.g. "MDS 1 – MDS 2 – MDS 3"; Fig. 4.4A). Scrambled loci meet at least one of two criteria: 1) MDSs are present in a non-consecutive order (e.g. "MDS 2 – MDS 3 – MDS 1") and/or 2) MDSs can be found on both strands of the germline scaffolds (i.e. some are inverted; Fig. 4.4B). In both *B. americanum* and *Loxodes* sp. we find several scrambled germline loci (24 and 23 respectively; Fig. 4.4B; Table 4.4) and easily recognizable non-scrambled germline loci (15 and 11 respectively; Table 4.4).

We find that the germline genome architecture of the members of the Po-clade are atypical from the expectations based on *C. uninata* and *O. trifallax* (members of the Im-clade). For example, the data on gene scrambling in the classes Spirotrichea and Phyllopharyngea (Im-clade) reveal small MDSs separated by relatively large distances in the germline genome (Chen, et al. 2014; Maurer-Alcalá, et al. in review). This is not the case in *B. americanum* and *Loxodes* sp. (Po-clade), where differences in the distances between MDSs for both scrambled and non-scrambled germline loci were insignificant ($p = 0.301$). Similarly, in both *C. uncinata* and *O. trifallax*, scrambled germline loci are composed of a greater number of MDSs than non-scrambled loci (Chen, et al. 2014; Maurer-Alcalá, et al. in review), yet for both *B. americanum* and *Loxodes* sp. nearly all germline loci (i.e. scrambled and non-scrambled) are composed of only two large MDSs (Table 4.4). The prevalence of two-MDS loci in the Po-clade suggests an as yet unknown link between germline genome architecture and macronuclear development in this clade.

The observations from the members of the Po-clade contrast with those from *Bursaria truncatella*, whose last common ancestor with the model ciliates *P. tetraurelia*

and *T. thermophila* was more recent (~500-700 MYA; Parfrey, et al. 2011). We did not find any evidence of scrambled germline loci from the mapping of transcriptomic data back to the putative germline scaffolds for *B. truncatella*, with all 162 identifiable germline loci being non-scrambled (Fig. 4.4A). This suggests that *B. truncatella*'s germline genome lacks substantial amounts of gene-scrambling and that the single-cell genomic methods used here do not introduce false evidence of scrambling.

Given the absence of scrambling, we sought to determine how similar the germline genome architecture of *B. truncatella* might be to *Paramecium* and *Tetrahymena*. The germline-limited IESs present in the *B. truncatella* germline genome do interrupt the protein-coding domains (Fig. 4.4A), as is the case in *P. tetraurelia* but not its close relative (Arnaiz, et al. 2012), *T. thermophila* where the majority of IESs occur in the intergenic regions (Hamilton, et al. 2016). Interestingly, the pointer sequences for *Paramecium tetraurelia* and *Tetrahymena thermophila*, which are involved in aiding the guided genome rearrangements during development, are redundant, being either 'TA' in *Paramecium* and 'TTAA' in *Tetrahymena* (Arnaiz, et al. 2012; Hamilton, et al. 2016). This contrasts with all the ciliates in our study, which possess unique pointer sequences for each germline locus.

4.5 Synthesis

The numerous subtle, yet impactful, differences in germline genome organization across both small (within class) and large (between class) phylogenetic distances call into question the existence of gross synapomorphies of the different ciliate genome architectures. Rather, with increasing evidence, there is a greater necessity to temper the expectations and 'rules' based on the phylogenetically limited data afforded from the

model organisms. As we have shown in this study, by addressing long-standing assumptions in a diverse clade of eukaryotes with emerging technologies, we can begin to shed light on the fact that the 'models', upon which so much information is built, may themselves be the exceptions to the rule.

4.6 Acknowledgements

The work in this study was supported by an NIH award (1R15GM113177) and NSF Go-LIFE (DEB-1541511) to LAK and a Blakeslee award to Smith College. We thank Kelsie Maurer-Alcalá for contributing artistic renderings of the organisms used in the figures of this manuscript. We also thank members of the Katz Lab for frequent and insightful discussion and members of the Knight Lab for their technical guidance.

Table 4.1. Raw estimates of chromosome copy numbers for several genes of *Blepharisma americanum* are incredibly variable and stochastic. Similarly, there is no significant difference in chromosome copy numbers that are attributable to the preparation of DNA for qPCR (e.g. total DNA extraction *versus* sc-WGA).

Sample	nSSU-rDNA	Actin P1	Actin P2	α-Tubulin P1	α-Tubulin P2	EF-1α
pop-DNA	$2.55x10^7 \pm 8.42x10^6$	$1.18x10^6 \pm 4.38x10^4$	$6.78x10^5 \pm 7.87x10^4$	$6.63x10^5 \pm 3.62x10^4$	$1.71x10^5 \pm 1.47x10^4$	$9.77x10^4 \pm 2.41x10^4$
WGA-1	$9.24x10^7 \pm 1.74x10^6$	$7.32x10^5 \pm 8.20x10^4$	$5.16x10^6 \pm 1.74x10^5$	$3.37x10^6 \pm 4.24x10^5$	$2.00x10^5 \pm 6.14x10^3$	$1.23x10^6 \pm 1.16x10^4$
WGA-2	$1.28x10^7 \pm 3.34x10^6$	$5.77x10^4 \pm 1.85x10^3$	$5.25x10^5 \pm 7.45x10^4$	$4.87x10^4 \pm 4.27x10^3$	$6.06x10^2 \pm 2.87x10^2$	$8.89x10^5 \pm 2.71x10^4$
WGA-3	$1.32x10^8 \pm 2.54x10^7$	$1.74x10^6 \pm 2.58x10^5$	$5.96x10^6 \pm 4.42x10^5$	$3.27x10^6 \pm 1.58x10^5$	$7.03x10^5 \pm 9.90x10^4$	$2.83x10^6 \pm 2.75x10^5$

Table 4.2. Relative qPCR-based estimates for chromosome copy numbers of several genes of *Loxodes* spp. reveal a 'semi-conserved' pattern of differential chromosome amplification. This pattern crosses the boundaries of 2 distinct morphospecies (WGA-1/2 and WGA-3/4).

Sample	nSSU-rDNA	Actin	Rs11	α-Tubulin	EF-1α
WGA-3	$1.00x10^{6} \pm 1.26x10^{5}$	$6.39x10^{3} \pm 1.27x10^{3}$	$1.85x10^{3} \pm 2.47x10^{2}$	$2.48x10^{1} \pm 1.99x10^{0}$	$6.34x10^{0} \pm 1.31x10^{0}$
WGA-4	$1.00x10^{6} \pm 8.31x10^{4}$	$1.61x10^{3} \pm 2.43x10^{2}$	$4.93x10^{1} \pm 5.47x10^{0}$	$7.46x10^{-1} \pm 2.26x10^{-1}$	$8.78x10^{-1} \pm 3.17x10^{-1}$
WGA-1	$1.00x10^{6} \pm 6.66x10^{4}$	$4.50x10^{3} \pm 1.47x10^{3}$	$1.99x10^{1} \pm 2.91x10^{0}$	$1.10x10^{2} \pm 1.46x10^{1}$	$4.39x10^{-1} \pm 9.85x10^{-2}$
WGA-2	$1.00x10^{6} \pm 1.68x10^{5}$	$2.87x10^{4} \pm 7.51x10^{3}$	$1.06x10^{3} \pm 3.46x10^{2}$	$4.38x10^{1} \pm 9.63x10^{0}$	$1.02x10^{1} \pm 3.71x10^{0}$

Table 4.3. List of qPCR primers used in this study for *Loxodes* spp.

Taxon	Gene	Primer	Primer sequence (5' - 3')	Expected Fragment Length
Loxodes spp.	nSSU-rDNA	Lox_qSSU_1F	TAGTTGGAGGAAGTGTGAGGC	216
		Lox_qSSU_1R	AGGGACTTAATCAGTGCAAGC	
	Actin	Lox_qACT_2F	TGAACCACCCGAGGAACATCC	190
		Lox_qACT_2R	CACCATCACCTGAGTCCATAAC	
	Alpha Tubulin	Lox_qATub_534F	CACTAACTCAGcTTTCGAACC	193
		Lox_qATub_727R	TAATACCGCATTTGAATCCAG	
	Elongation Factor 1α	Lox_qEF1a_1F	GGTGGTATCGGAACTGTCCC	170
		Lox_qEF1a_1R	GGTCATTCTTTGAGTCACCAC	
	Ribosomal protein s11	Lox_qRs11_80F	TGTCTTGCTTCCAGTTCCACC	143
		Lox_qRs11_223R	TGATGGTGAGAGCTGACAGAG	
Blepharisma americanum	nSSU-rDNA	Lox_qSSU_1F	TAGTTGGAGGAAGTGTGAGGC	216
		Lox_qSSU_1R	AGGGACTTAATCAGTGCAAGC	
	Actin P1	Bleph_qACT_P1_F	TGCTAAGTGCCAAGGGATAC	209
		Bleph_qACT_P1_R	GTCAAACATTGCTTCTGGGC	
	Actin P2	Bleph_qACT_P2_F	CATCCAAGCTCTGCTCTCTG	197
		Bleph_qACT_P2_R	TGGTTAGTGAGTAGCCTCTTGC	
	Alpha Tubulin P1	Bleph_qATub_P1_F	GCTTCACCGTTTACCCAAGC	194
		Bleph_qATub_P1_R	GGCAATAAGTCTGTTCAAGTTGG	
	Alpha Tubulin P2	Bleph_qATub_P2_F	GCTTCACCGTTTACCCATCT	194
		Bleph_qATub_P2_R	GGCAATAAGTCTGTTCAAGTTTG	
	Elongation Factor 1α	Bleph_qEf1a_491F	GATGTGTTGAGAGCTGAACC	211
		Bleph_qEf1a_701R	TATGTGAAGTATGGCAGTCC	

Table 4.4. Summary statistics of germline genome architecture for ciliates in this study. All values in parentheses represent the median values for that category.

	Bursaria truncatella	***Blepharisma americanum***	***Loxodes* sp.**
Scrambled Loci	–	24	23
Non-Scrambled Loci	162	15	11
Scrambled MDS	–	2.08 (2)	2.1 (2)
Non-Scrambled MDS	2.34 (2)	2.00 (2)	2.00 (2)
Scrambled MDS Length	–	274.48 (294)	304.81 (301)
Scrambled IES Length	–	520.23 (498)	552.45 (544)
Non-Scrambled MDS Length	244.08 (202)	272.53 (269)	341.11 (305)
Non-Scrambled IES Length	430.92 (419)	527.29 (513)	591.66 (560)

MDS: Macronuclear Destined Sequences (soma)
IES: Internally Eliminated Sequences (germline)

Figure 4.1. Summary of general ciliate features demonstrates large gaps in knowledge for many ciliate classes. Absence of available data is denoted as '–'. Novel data presented in this study are in Blue. Germline (Germ) genomes are denoted as either scrambled (Sc) or non-scrambled (NS). Somatic genomes (Soma) are marked as either extensively fragmented (EF) or non-extensively fragmented (NEF). Similarly, copy number variation of chromosomes containing protein coding genes are indicated as variable (V) or approximately equal (≈). The lineages in the Postciliodesmatophora (Po-clade) are highlighted by red. The remaining ciliate classes are found in the Intramacronucleata (Im-clade).

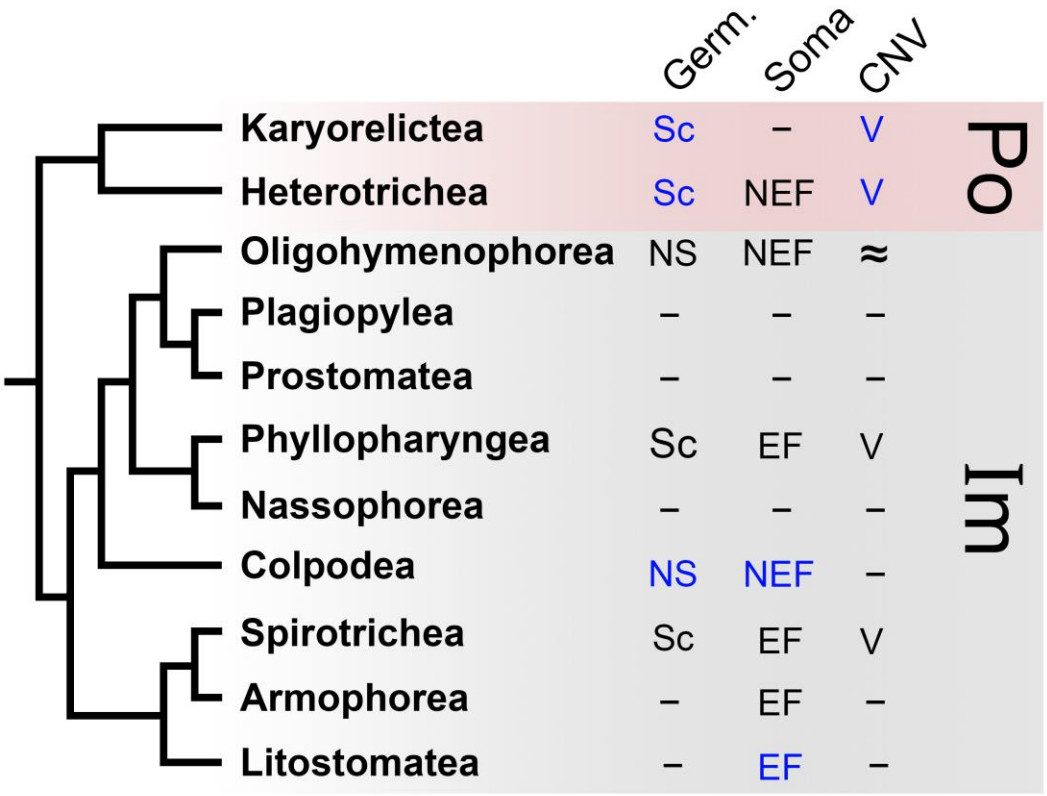


Figure 4.2. Relative chromosome copy numbers for members of the Po-Clade show contrasting patterns of high copy number but stochasticity in *Blepharisma* and variable but repeatable copy number in Loxosodes. Expected (Armstrong, et al.) plots of chromosome copy number for *Blepharisma americanum* (A) and *Loxodes* spp. (C) are based on previous studies. The observed variable copy number for *B. americanum* (B) corresponds to the expected results for both the population sample (pop-DNA) and the three individuals (WGA-#). However, for all four *Loxodes* spp. individuals, (WGA-1/2 and WGA-3/4 representing distinct morphospecies), the observed chromosome copy number (D) deviates substantially from the expected copy numbers (C). '*' indicate relative chromosome copy number values below '3'.

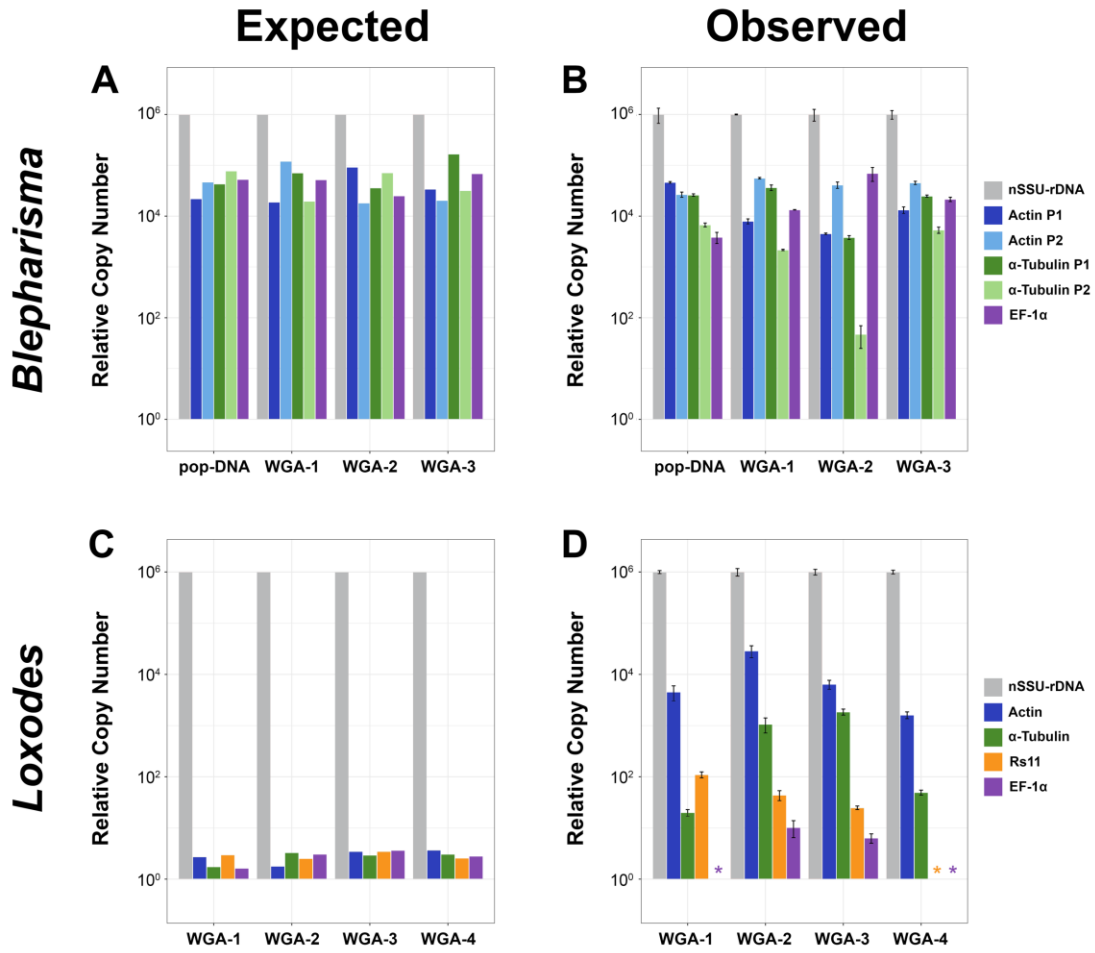


Figure 4.3. Distribution of chromosome lengths among diverse taxa reveals unexpected pool of minute nano-chromosomes in *Didinium nasutum* and *Entodinium caudatum*. Representative images of each taxon are next to their names and are not drawn to scale. *Tetrahymena thermophila*'s germline chromosomes are noted, whereas the ciliate's drawing is next to its somatic chromosomes.

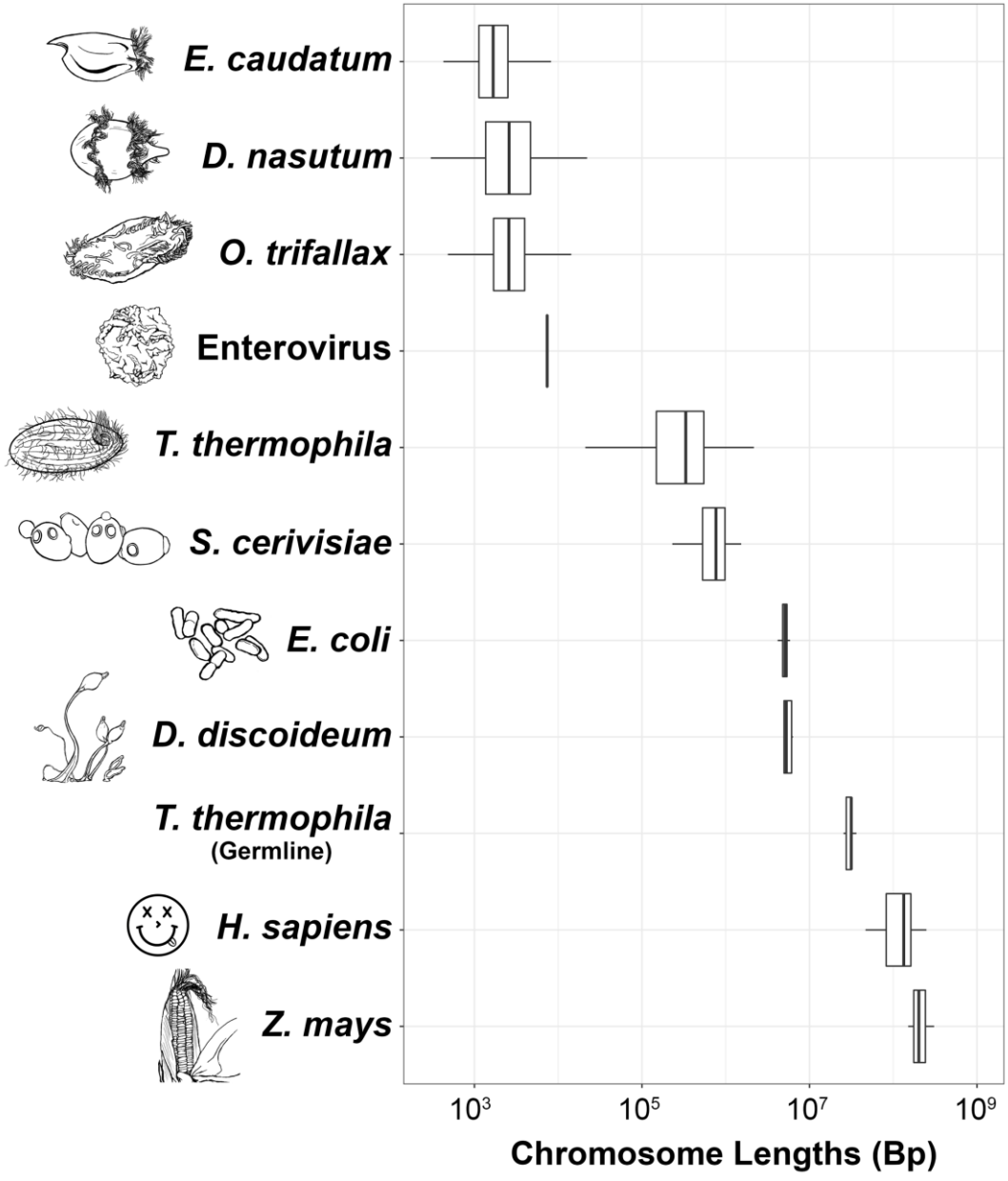
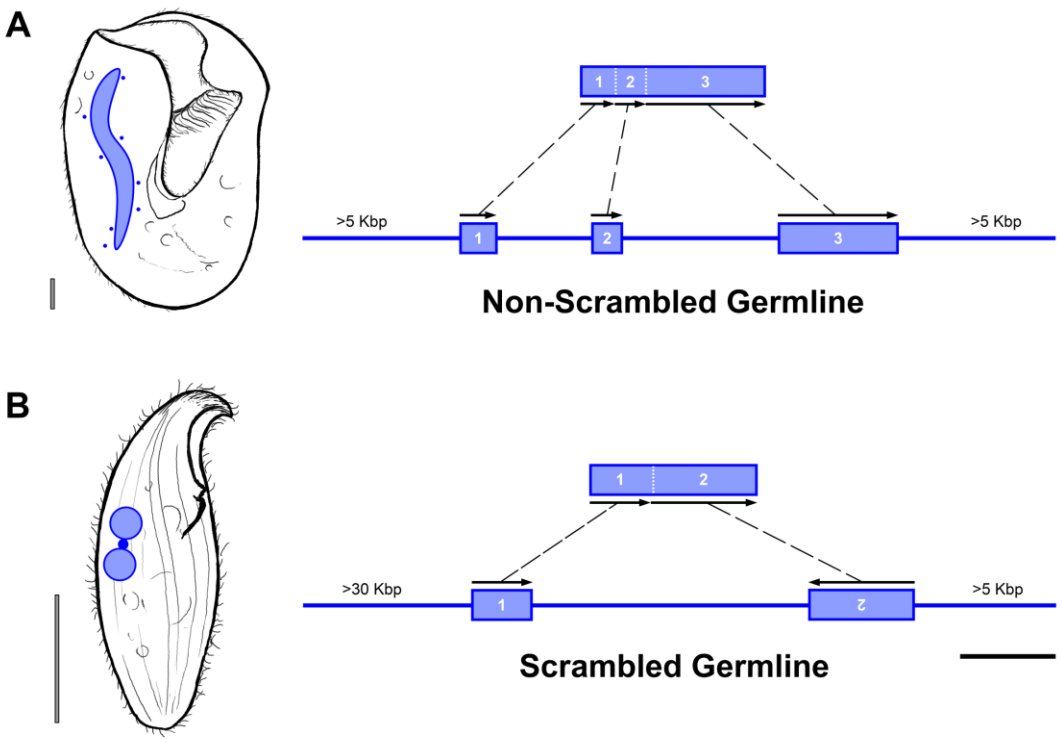


Figure 4.4. Exemplar cases of ciliate germline genome architecture from *Bursaria* and *Loxodes*. Left, representative images of *Bursaria truncatella* (A) and *Loxodes* sp. (B) with their germline (small blue circles) and somatic nuclei (blue-bordered). Right, germline loci are represented as a single line harboring MDSs (blue-bordered rectangles). All identifiable germline loci from *Bursaria truncatella* (A) were non-scrambled, whereas for *Loxodes* sp. (B) there is a mixture of scrambled and non-scrambled loci (only scrambled shown here). MDSs are numbered according to the order in which they are found in the soma and the corresponding arrows indicate their directionality in the germline genome. Bottom right scale bar (black) is 300bp. Scale bar (bottom right of each ciliate) is 25 µm.



APPENDIX

PRODUCTS RESULTING FROM THIS DISSERTATION

Thesis Chapters

Maurer-Alcalá XX, Yan Y, Pilling O, Knight R, Katz LA. *In Prep*. Twisted Tales: Insights into Genome Diversity of Ciliates Using Single-Cell Genomics.

Maurer-Alcalá XX, Knight R, Katz LA. *Accepted*. Exploring the Germline Genome of the Ciliate *Chilodonella uncinata* Through Single-cell 'omics (Transcriptomics and Genomics).

Maurer-Alcalá XX and Katz LA. 2016. Nuclear Architecture and Patterns of Molecular Evolution Are Correlated in the Ciliate *Chilodonella uncinata*. *Genome Biol. Evol.* 8(6)

Maurer-Alcalá XX and Katz LA. 2015. An epigenetic toolkit allows for diverse genome architectures in eukaryotes. *Curr. Opin. Genet. Dev.* 35:93-99

Other Publications

Wancura M*****, Yan Y, Katz LA, **Maurer-Alcalá XX**. 2017. Genome amplification, life cycle and nuclear inclusion in the ciliate *Blepharisma americanum*. *Journal of Eukaryotic Microbiology.*

Tekle YI, Anderson OR, Katz LA, **Maurer-Alcalá XX**, Cerón Romero MA, Molestina R. 2016. Phylogenomics of 'Discosea': A new molecular phylogenetic perspective on Amoebozoa with flat body forms. *Molecular Phylogenetics and Evolution*. 99:144-154

Bellec L, **Maurer-Alcalá XX**, Katz LA. 2014. Characterization of the Life Cycle and Heteromeric Nature of the Macronucleus of the Ciliate *Chilodonella uncinata* Using Fluorescence Microscopy. *Journal of Eukaryotic Microbiology*. 61(3):313-316

BIBLIOGRAPHY

Aeschlimann SH, et al. 2014. The draft assembly of the radically organized Stylonychia lemnae macronuclear genome. Genome Biology and Evolution 6: 1707-1723.

Allen SE, Nowacki M 2017. Necessity Is the Mother of Invention: Ciliates, Transposons, and Transgenerational Inheritance. Trends in Genetics 33: 197-207.

Alt FW, Blackwell TK, Depinho RA, Reth MG, Yancopoulos GD. 1986. Regulation of Genome Rearrangement Events during Lymphocyte Differentiation. Immunological Reviews 89: 5-30.

Ammermann D. 1986. Giant Chromosomes in Ciliates. Biological Chemistry Hoppe-Seyler 367: 1102-1102.

Andersson JO, et al. 2007. A genomic survey of the fish parasite Spironucleus salmonicida indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. Bmc Genomics 8: 25.

Aravin AA, et al. 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D-melanogaster germline. Current Biology 11: 1017-1027.

Ardell DH, Lozupone CA, Landweber LF. 2003. Polymorphism, recombination and alternative unscrambling in the DNA polymerase alpha gene of the ciliate *Stylonychia lemnae* (Alveolata; class Spirotrichea). Genetics 165: 1761-1777.

Arkhipova I, Meselson M. 2005. Deleterious transposable elements and the extinction of asexuals. Bioessays 27: 76-85.

Arkhipova IR, Morrison HG. 2001. Three retrotransposon families in the genome of Giardia lamblia: Two telomeric, one dead. Proceedings of the National Academy of Sciences of the United States of America 98: 14497-14502.

Arnaiz O, et al. 2012. The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. Plos Genetics 8.

10.1371/journal.pgen.1002984

Aury JM, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444: 171-178.

Ay F, et al. 2015. Multiple dimensions of epigenetic gene regulation in the malaria parasite Plasmodium falciparum. Bioessays 37: 182-194.

Bachmann-Waldmann C, Jentsch S, Tobler H, Muller F. 2004. Chromatin diminution leads to rapid evolutionary changes in the organization of the germ line genomes

of the parasitic nematodes A. suum and P. univalens. Molecular and Biochemical Parasitology 134: 53-64.

Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research 37: W202-W208.

Balbiani E. 1881. Sur la structure du noyau des cellules salivaires chez les larves de Chironomus. Zoologischer Anzeiger 4: 637-641.

Balbiani EG. 1890. Sur la structure intime du noyau du Loxophyllum meleagris. Zool. Anz 13: 110-115, 132-136.

Bankevich A, et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 19: 455-477.

Bellec L, Katz LA. 2012. Analyses of chromosome copy number and expression level of four genes in the ciliate *Chilodonella uncinata* reveal a complex pattern that suggests epigenetic regulation. Gene 504: 303-308.

Bellec L, Maurer-Alcalá XX, Katz LA. 2014. Characterization of the life cycle and heteromeric nature of the macronucleus of the ciliate *Chilodonella uncinata* using fluorescence microscopy. J. Euk. Micro. 61: 313–316.

Belyayev A. 2014. Bursts of transposable elements as an evolutionary driving force. J Evol Biol 27: 2573-2584.

Blackburn EH, Gall JG. 1978. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in *Tetrahymena*. J. Mol. Biol. 120: 33-53.

Bobyleva NN, Kudrjavtsev BN, Raikov IB. 1980. Changes of the DNA Content of Differentiating and Adult Macronuclei of the Ciliate Loxodes-Magnus (Karyorelictida). Journal of Cell Science 44: 375-394.

Bracht JR, Perlman DH, Landweber LF. 2012. Cytosine methylation and hydroxymethylation mark DNA for elimination in Oxytricha trifallax. Genome Biology 13.

Bushnell B. 2015. BBMap Short-Read Aligner, and Other Bioinformatics Tools.

Calistri E, Livi R, Buiatti M 2011. Evolutionary trends of GC/AT distribution patterns in promoters. Molecular Phylogenetics and Evolution 60: 228-235.

Camacho C, et al. 2009. BLAST plus : architecture and applications. Bmc Bioinformatics 10.

Cavaliersmith T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell-volume and cell-growth rate, and solution of DNA C-value paradox. Journal of Cell Science 34: 247-278.

Cerutti H, Casas-Mollano JA. 2006. On the oridin and functions of RNA-mediated silencing: from protists to man. Current Genetics 50: 81-99.

Chalker DL, Meyer E, Mochizuki K. 2013. Epigenetics of Ciliates. Cold Spring Harbor Perspectives in Biology 5.

Chalker DL, Yao MC. 2011. DNA Elimination in Ciliates: Transposon Domestication and Genome Surveillance. In: Bassler BL, Lichten M, Schupbach G, editors. Annual Review Genetics, Vol 45. p. 227-246.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research 34: D363-D368.

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

Cohen S, Houben A, Segal D. 2008. Extrachromosomal circular DNA derived from tandemly repeated genomic sequences in plants. Plant Journal 53: 1027-1034.

Corradi N, Pombert JF, Farinelli L, Didier ES, Keeling PJ. 2010. The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. Nature Communications 1.

Cortes A, Crowley VM, Vaquero A, Voss TS. 2012. A View on the Role of Epigenetics in the Biology of Malaria Parasites. Plos Pathogens 8.

Coyne RS, Lhuillier-Akakpo M, Duharcourt S. 2012. RNA-guided DNA rearrangements in ciliates: Is the best genome defence a good offence? Biology of the Cell 104: 309-325.

Croken MM, Nardelli SC, Kim K. 2012. Chromatin modifications, epigenetics, and how protozoan parasites regulate their lives. Trends in Parasitology 28: 202-213.

DeBolt S. 2010. Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate Family Generational Scales. Genome Biology and Evolution 2: 441-453.

Deshmukh AS, Srivastava S, Dhar SK. 2013. Plasmodium falciparum: epigenetic control of var gene regulation and disease. Sub-Cellular Biochemistry 61: 659-682.

Diez CM, Roessler K, Gaut BS. 2014. Epigenetics and plant genome evolution. Current Opinion in Plant Biology 18: 1-8.

Doerder FP, Deak JC, Lief JH. 1992. Rate of Phenotypic Assortment in Tetrahymena-Thermophila. Developmental Genetics 13: 126-132.

Drouin G. 2006. Chromatin diminution in the copepod Mesocyclops edax: diminution of tandemly repeated DNA families from somatic cells. Genome 49: 657-665.

Duret L, et al. 2008. Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia: A somatic view of the germline. Genome Research 18: 585-596.

Eichinger L, et al. 2005. The genome of the social amoeba Dictyostelium discoideum. Nature 435: 43-57.

Eisen JA, et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. Plos Biology 4: e286.

Espada J, Esteller M. 2007. Epigenetic control of nuclear architecture. Cellular and Molecular Life Sciences 64: 449-457.

Fedoroff NV. 2012. Presidential address. Transposable elements, epigenetics, and genome evolution. Science 338: 758-767.

Feng SH, Jacobsen SE, Reik W. 2010. Epigenetic Reprogramming in Plant and Animal Development. Science 330: 622-627.

Friz CT. 1968. The biochemical composition of the free-living Amoebae *Chaos chaos, Amoeba dubia* and *Amoeba proteus*. Comp. Biochem. Physiol. 26: 81-90.

Fuhrmann G, Jonsson F, Weil PP, Postberg J, Lipps HJ. 2016. RNA-template dependent de novo telomere addition. Rna Biology 13: 733-739.

Fujimori S, Washio T, Tomita M. 2005. GC-compositional strand bias around transcription start sites in plants and fungi. Bmc Genomics 6.

Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. Trends in Genetics 20: 417-423.

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

Gao F, Song WB, Katz LA 2014. Genome structure drives patterns of gene family evolution in ciliates, a case study using Chilodonella uncinata (Protista, Ciliophora, Phyllopharyngea). Evolution 68: 2287-2295.

Gawryluk RMR, et al. 2016. Morphological Identification and Single-Cell Genomics of Marine Diplonemids. Current Biology 26: 3053-3059.

Gijzen M. 2009. Runaway repeats force expansion of the Phytophthora infestans genome. Genome Biology 10.

Goday C, Esteban MR. 2001. Chromosome elimination in sciarid flies. Bioessays 23: 242-250.

Golikova M. 1965. Der Aufbau des Kernapparates und die Verteilung der Nukleinsäuren und Proteine bei *Nyctotherus cordiformis* stein. Archiv für Protistenkunde: 191-216.

Gomez-Diaz E, Jorda M, Angel Peinado M, Rivero A. 2012. Epigenetics of Host-Pathogen Interactions: The Road Ahead and the Road Behind. Plos Pathogens 8.

Gong J, Dong J, Liu XH, Massana R. 2013. Extremely High Copy Numbers and Polymorphisms of the rDNA Operon Estimated from Single Cell Analysis of Oligotrich and Peritrich Ciliates. Protist 164: 369-379.

Gregory TR. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biological Reviews 76: 65-101.

Guerin F, et al. 2017. Flow cytometry sorting of nuclei enables the first global characterization of Paramecium germline DNA and transposable elements. Bmc Genomics 18.

Guo YL. 2013. Gene family evolution in green plants with emphasis on the origination and evolution of Arabidopsis thaliana genes. Plant Journal 73: 941-951.

Haas BJ, et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature 461: 393-398.

Haerty W, Ponting CP. 2015. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. Rna-a Publication of the Rna Society 21: 320-332.

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 Drosophila genomes. Plos Genetics 3: 2135-2146.

Halic M, Moazed D. 2009. Transposon Silencing by piRNAs. Cell 138: 1058-1060.

Hamilton EP, et al. 2016. Structure of the germline genome of Tetrahymena thermophila and relationship to the massively rearranged somatic genome. Elife 5.

Hari Dass SA, Vyas A. 2014. Toxoplasma gondii infection reduces predator aversion in rats through epigenetic modulation in the host medial amygdala. Mol Ecol 23: 6114-6122.

Heard E, Martienssen RA. 2014. Transgenerational epigenetic inheritance: myths and mechanisms. Cell 157: 95-109.

Heyse G, Jonsson F, Chang WJ, Lipps HJ. 2010. RNA-dependent control of gene amplification. Proceedings of the National Academy of Sciences of the United States of America 107: 22134-22139.

Huang J, Katz LA. 2014. Nanochromosome Copy Number Does not Correlate with RNA Levels Though Patterns are Conserved between Strains of the Ciliate Morphospecies *Chilodonella uncinata*. Protist 165: 445-451.

Jahn CL, Klobutcher LA. 2002. Genome remodeling in ciliated protozoa. Annual Reviews in Microbiology 56: 489-520.

Jonsson F, Postberg J, Lipps HJ. 2009. The unusual way to make a genetically active nucleus. DNA and Cell Biology 28: 71-78.

Juranek SA, Rupprecht S, Postberg J, Lipps HJ. 2005. snRNA and heterochromatin formation are involved in DNA excision during macronuclear development in stichotrichous ciliates. Eukaryotic Cell 4: 1934-1941.

Kathiria P, et al. 2010. Tobacco Mosaic Virus Infection Results in an Increase in Recombination Frequency and Resistance to Viral, Bacterial, and Fungal Pathogens in the Progeny of Infected Tobacco Plants. Plant Physiology 153: 1859-1870.

Katz LA. 2001. Evolution of nuclear dualism in ciliates: a reanalysis in light of recent molecular data. Int. J. Syst. Evol. Microbiol. 51: 1587-1592.

Katz LA, Kovner AM. 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: 480-488.

Kaul S, et al. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796-815.

Klenov MS, et al. 2007. Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the Drosophila melanogaster germline. Nucleic Acids Research 35: 5430-5438.

Klobutcher LA, Herrick G. 1997. Developmental genome reorganization in ciliated protozoa: The transposon link. Progress in Nucleic Acid Research and Molecular Biology, Vol. 56 56: 1-62.

Kloc A, Zaratiegui M, Nora E, Martienssen R. 2008. RNA interference guides histone modification during the S phase of chromosomal replication. Current Biology 18: 490-495.

Kohno S, Kubota S, Nakai Y. 1998. Chromatin diminution and chromosome elimination in hagfishes. Biology of Hagfishes: 81-100.

Kovaleva VG, Raikov IB. 1978. Diminution and Re-Synthesis of DNA during Development and Senescence of Diploid Macronuclei of Ciliate Trachelonema-Sulcata (Gymnostomata-Karyorelictida). Chromosoma 67: 177-192.

Kruger K, et al. 1982. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. Cell 31: 147-157. ‘

Kumari V, et al. 2011. Differential distribution of a SINE element in the Entamoeba histolytica and Entamoeba dispar genomes: role of the LINE-encoded endonuclease. Bmc Genomics 12: 267.

Lai JS, Li YB, Messing J, Dooner HK. 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proceedings of the National Academy of Sciences of the United States of America 102: 9068-9073.

Landeira D, Navarro M. 2007. Nuclear repositioning of the VSG promoter during developmental silencing in Trypanosoma brucei. J Cell Biol 176: 133-139.

Landweber LF, Kuo TC, Curtis EA. 2000. Evolution and assembly of an extremely scrambled gene. Proceedings of the National Academy of Sciences of the United States of America 97: 3298-3303.

Law JA, Vashisht AA, Wohlschlegel JA, Jacobsen SE. 2011. SHH1, a Homeodomain Protein Required for DNA Methylation, As Well As RDR2, RDM4, and Chromatin Remodeling Factors, Associate with RNA Polymerase IV. Plos Genetics 7.

Le Blancq SM, Adam RD. 1998. Structural basis of karyotype heterogeneity in *Giardia lamblia*. Molecular and Biochemical Parasitology 97: 199-208.

Levy SF, Ziv N, Siegal ML. 2012. Bet Hedging in Yeast by Heterogeneous, Age-Correlated Expression of a Stress Protectant. Plos Biology 10.

Li BB. 2015. DNA Double-Strand Breaks and Telomeres Play Important Roles in Trypanosoma brucei Antigenic Variation. Eukaryotic Cell 14: 196-205.

Lin S. 2011. Genomic understanding of dinoflagellates. Res Microbiol 162: 551-569.

Lisch D. 2009. Epigenetic Regulation of Transposable Elements in Plants. Annual Review of Plant Biology 60: 43-66.

Liu Y, et al. 2007. RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in Tetrahymena. Genes & Development 21: 1530-1545.

Lopez-Gomollon S, et al. 2014. Global discovery and characterization of small non-coding RNAs in marine microalgae. Bmc Genomics 15.

Maatouk DM, et al. 2006. DNA methylation is a primary mechanism for silencing postmigratory primordial germ cell genes in both germ cell and somatic cell lineages. Development 133: 3411-3418.

Mani RS, Chinnaiyan AM. 2010. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. Nature Reviews Genetics 11: 819-829.

Marr AK, et al. 2014. Leishmania donovani Infection Causes Distinct Epigenetic DNA Methylation Changes in Host Macrophages. Plos Pathogens 10.

Matzke M, Kanno T, Huettel B, Daxinger L, Matzke AJM. 2007. Targets of RNA-directed DNA methylation. Current Opinion in Plant Biology 10: 512-519.

Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nature Reviews Genetics 15.

Maumus F, Rabinowicz P, Bowler C, Rivarola M. 2011. Stemming Epigenetics in Marine Stramenopiles. Current Genomics 12: 357-370.

Maurer-Alcalá XX, Katz LA. 2015. An epigenetic toolkit allows for diverse genome architectures in eukaryotes. Current Opinion in Genetics & Development 35: 93-99.

Maurer-Alcalá XX, Katz LA. 2016. Nuclear architecture and patterns of molecular evolution are correlated in the ciliate Chilodonella uncinata. Genome Biology and Evolution.

Maurer-Alcalá XX, Knight R, Katz LA. *in review*. Exploring the Germline Genome of the Ciliate Chilodonella uncinata Through Single-cell ‘omics (Transcriptomics and Genomics).

McGrath CL, Zufall RA, Katz LA. 2007. Variation in macronuclear genome content of three ciliates with extensive chromosomal fragmentation: a preliminary analysis. Journal of Eukaryotic Microbiology 54: 242-246.

Miller DN, Bryant JE, Madsen EL, Ghiorse WC. 1999. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. Applied and Environmental Microbiology 65: 4715-4724.

Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA. 2002. Analysis of a *piwi*-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. Cell 110: 689-699.

Mochizuki K, Gorovsky MA. 2004. Small RNAs in genome rearrangement in *Tetrahymena*. Current Opinion in Genetics & Development 14: 181.

Molinier J, Ries G, Zipfel C, Hohn B. 2006. Transgeneration memory of stress in plants. Nature 442: 1046-1049.

Morton NE. 1991. Parameters of the Human Genome. Proceedings of the National Academy of Sciences of the United States of America 88: 7474-7476.

Nakai Y, Kubota S, Kohno S. 1991. Chromatin Diminution and Chromosome Elimination in 4 Japanese Hagfish Species. Cytogenetics and Cell Genetics 56: 196-198.

Nieuwenhuis BPS, Immler S. 2016. The evolution of mating-type switching for reproductive assurance. Bioessays 38: 1141-1149.

Nowacki M, et al. 2008. RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature 451: 153-U154.

Oliverio AM, Katz LA. 2014. The dynamic nature of genomes across the tree of life. Genome Biology and Evolution 6: 482-488.

Orias E, 1991. Evolution of amitosis of the ciliate macronucleus: gain of the capacity to divide. J. Protozool 38: 217-221.

Ovchinnikova L, Cheissin E, Selivanova G. 1965. Photometric study of the DNA content in the nuclei of *Spirostomum ambiguum* (Ciliata, Heterotricha).

Parfrey LW, Lahr DJG, Katz LA 2008. The dynamic nature of eukaryotic genomes. Molecular Biology and Evolution 25: 787-794.

Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proceedings of the National Academy of Sciences of the United States of America 108: 13624-13629.

Polak P, Querfurth R, Arndt PF. 2010. The evolution of transcription-associated biases of mutations across vertebrates. Bmc Evolutionary Biology 10.

Popenko VI, Potekhin AA, Karajan BP, Skarlato SO, Leonova OG. 2015. The Size of DNA Molecules and Chromatin Organization in the Macronucleus of the Ciliate Didinium nasutum (Ciliophora). Journal of Eukaryotic Microbiology 62: 260-264.

Postberg J, Heyse K, Cremer M, Cremer T, Lipps HJ. 2008. Spatial and temporal plasticity of chromatin during programmed DNA-reorganization in Stylonychia macronuclear development. Epigenetics & Chromatin 1.

Poxleitner MK, et al. 2008. Evidence for karyogamy and exchange of genetic material in the binucleate intestinal parasite Giardia intestinalis. Science 319: 1530-1533.

Prescott DM. 1994. The DNA of Ciliated Protozoa. Microbiological Reviews 58: 233-267.

Pyne CK .1978. Electron-microscopic studies on macronuclear development in ciliate *Chilodonella uncinata*. Cytobiologie 18: 145-160.

R_Core_Team. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nature Reviews Microbiology 10: 417-430.

Raikov IB. 1994. The Nuclear-Apparatus of Some Primitive Ciliates, the Karyorelictids - Structure and Divisional Reorganization. Bollettino Di Zoologia 61: 19-28.

Raikov IB. 1985. Primitive Never-Dividing Macronuclei of Some Lower Ciliates. International Review of Cytology-a Survey of Cell Biology 95: 267-325.

Raikov IB. 1982. The Protozoan Nucleus: Morphology and Evolution. Wien: Springer-Verlag.

Raikov IB, Karadzhan BP. 1985. Fine-Structure and Cyto-Chemistry of the Nuclei of the Primitive Ciliate Tracheloraphis-Crassus (Karyorelictida). Protoplasma 126: 114-129.

Ramesh MA, Malik S-B, Logsdon JM. 2005. A Phylogenomic Inventory of Meiotic Genes: Evidence for Sex in Giardia and an Early Eukaryotic Origin of Meiosis. 15: 185.

Ricard G, et al. 2008. Macronuclear genome structure of the ciliate Nyctotherus ovalis: Single-gene chromosomes and tiny introns. Bmc Genomics 9.

Riley JL, Katz LA. 2001. Widespread distribution of extensive genome fragmentation in ciliates. Mol. Biol. Evol. 18: 1372-1377.

Robert VJP, Sijen T, van Wolfswinkel J, Plasterk RHA. 2005. Chromatin and RNAi factors protect the C-elegans germline against repetitive sequences. Genes & Development 19: 782-787.

Rogato A, et al. 2014. The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum. Bmc Genomics 15.

Ross L, Pen I, Shuker DM. 2010. Genomic conflict in scale insects: the causes and consequences of bizarre genetic systems. Biological Reviews 85: 807-828.

Rouxel T, et al. 2011. Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations. Nature Communications 2.

Roy RS, et al. 2014. Single cell genome analysis of an uncultured heterotrophic stramenopile. Scientific Reports 4.

Salcedo-Amaya AM, Hoeijmakers WA, Bartfai R, Stunnenberg HG. 2010. Malaria: could its unusual epigenome be the weak spot? Int J Biochem Cell Biol 42: 781-784.

Santangelo G, Barone E. 1987. Experimental Results on Cell-Volume, Growth-Rate, and Macronuclear DNA Variation in a Ciliated Protozoan. Journal of Experimental Zoology 243: 401-407.

Saze H, Tsugane K, Kanno T, Nishimura T. 2012. DNA Methylation in Plants: Relationship to Small RNAs and Histone Modifications, and Functions in Transposon Inactivation. Plant and Cell Physiology 53: 766-784.

Schatz DG, Swanson PC. 2011. V(D)J Recombination: Mechanisms of Initiation. Annual Review of Genetics, Vol 45 45: 167-202.

Shabalina SA, Koonin EV. 2008. Origins and evolution of eukaryotic RNA interference. Trends in Ecology & Evolution 23: 578-587.

Sinclair DA, Guarente L. 1997. Extrachromosomal rDNA circles - A cause of aging in yeast. Cell 91: 1033-1042.

Slabodnick MM, et al. 2017. The Macronuclear Genome of Stentor coeruleus Reveals Tiny Introns in a Giant Cell. Current Biology 27: 569-575.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics 8: 272-285.

Smith JJ, Antonacci F, Eichler EE, Amemiya CT. 2009. Programmed loss of millions of base pairs from a vertebrate genome. Proceedings of the National Academy of Sciences of the United States of America 106: 11212-11217.

Sonda S, et al. 2010. Epigenetic mechanisms regulate stage differentiation in the minimized protozoan Giardia lamblia. Molecular Microbiology 76: 48-67.

Spear BB, Lauth MR. 1976. Polytene Chromosomes of Oxytricha - Biochemical and Morphological Changes during Macronuclear Development in a Ciliated Protozoan. Chromosoma 54: 1-13.

Stancheva I. 2005. Caught in conspiracy: cooperation between DNA methylation and histone H3K9 methylation in the establishment and maintenance of heterochromatin. Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire 83: 385-395.

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research 33: W465-W467.

Sun C, Wyngaard G, Walton DB, Wichman HA, Mueller RL. 2014. Billions of basepairs of recently expanded, repetitive sequences are eliminated from the somatic genome during copepod development. Bmc Genomics 15.

Swart EC, et al. 2013. The Oxytricha trifallax Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes. Plos Biology 11.

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J. 2011. Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and Zea luxurians. Genome Biology and Evolution 3: 219-229.

Tricker PJ. 2015. Transgenerational inheritance or resetting of stress-induced epigenetic modifications: two sides of the same coin. Frontiers in Plant Science 6.

van West P, et al. 2008. Internuclear gene silencing in Phytophthora infestans is established through chromatin remodelling. Microbiology-Sgm 154: 1482-1490.

Veluchamy A, et al. 2013. Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. Nature Communications 4.

Venter JC, et al. 2001. The sequence of the human genome. Science 291: 1304-+.

Venter JC, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science (New York, N.Y.) 304: 66-74.

Vetukuri RR, et al. 2013. Phenotypic diversification by gene silencing in Phytophthora plant pathogens. Communicative & integrative biology 6: e25890-e25890.

Vogt G. 2015. Stochastic developmental variation, an epigenetic source of phenotypic diversity with far-reaching biological consequences. Journal of Biosciences 40: 159-204.

Volpe TA, et al. 2002. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science 297: 1833-1837.

Wahl MC, Will CL, Luhrmann R. 2009. The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell 136: 701-718.

Wancura MW, Yan Y, Katz LA, Maurer-Alcalá XX. 2017. Nuclear features of the heterotrich Ciliate *Blepharisma americanum*: genome amplification, life cycle, and nuclear inclusion. Journal of Eukaryotic Microbiology 0: 1-8.

Wang JB, Davis RE. 2014. Programmed DNA elimination in multicellular organisms. Current Opinion in Genetics & Development 27: 26-34.

Wassenegger M, Heimes S, Riedel L, Sanger HL. 1994. Rna-Directed De-Novo Methylation of Genomic Sequences in Plants. Cell 76: 567-576.

Weedall GD, Hall N. 2011. Evolutionary genomics of Entamoeba. Res Microbiol 162: 637-645.

Wegrzyn JL, et al. 2014. Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation. Genetics 196: 891-+.

Wichterman R. 1937. Division and conjugation in Nyctotherus cordiformis (Ehr.) Stein (Protozoa, Ciliata) with special reference to the nuclear phenomena. Journal of Morphology: 563-611.

Wong LC, Landweber LF. 2006. Evolution of programmed DNA rearrangements in a scrambled gene. Molecular Biology and Evolution 23: 756-763.

Woodard J, Gorovsky MA, Kaneshiro E. 1972. Cytochemical Studies on Problem of Macronuclear Subnuclei in Tetrahymena. Genetics 70: 251-+.

Wyngaard GA, Rasch EM, Connelly BA. 2011. Unusual augmentation of germline genome size in Cyclops kolensis (Crustacea, Copepoda): further evidence in support of a revised model of chromatin diminution. Chromosome research. 19: 911-923.

Xu K, et al. 2012. Copy number variations of 11 macronuclear chromosomes and their gene expression in *Oxytricha trifallax*. Gene 505: 75-80.

Yan Y, Rogers AJ, Gao F, Katz LA. 2017. Unusual features of non-dividing somatic macronuclei in the ciliate class Karyorelictea. European Journal of Protistology.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends in Genetics 13: 335-340.

Yoon HS, et al. 2011. Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332: 714-717.

Zhu LC, et al. 2009. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. Bmc Genomics 10.

Zilberman D, Henikoff S 2004. Silencing of transposons in plant genomes: kick them when they're down. Genome Biology 5.

Zimin AV, et al. 2013. The MaSuRCA genome assembler. Bioinformatics 29: 2669-2677.

Zufall RA, Robinson T, Katz LA. 2005. Evolution of developmentally regulated genome rearrangements in eukaryotes. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 304B: 448-455.