UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Macronuclear development in the ciliate *Oxytricha trifallax*

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MOLECULAR, CELL and DEVELOPMENTAL BIOLOGY

by

Zachary T. Neeb

June 2016

The Dissertation of Zachary T. Neeb
is approved:

Professor Alan Zahler, Ph.D.

Professor Melissa Jurica, Ph.D.

Professor Jeremy Sanford, Ph.D.

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

mRNA Expression Profiles During Macronuclear Development in the Ciliate *Oxytricha trifallax*

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 2

Chapter 3

Abstract

Macronuclear development in the ciliate *Oxytricha trifallax*

by

Zachary T. Neeb

Oxytricha trifallax, like all ciliated protozoans, possess two distinct types of nuclei: a germline micronucleus that is transcriptionally silent and a somatic macronucleus used for vegetative growth of the cells that is transcriptionally active. While the micronucleus resembles a typical eukaryotic nucleus with DNA organized on long chromosomes, micronuclear genes are interrupted by nongenic DNA and often exist in a non-linear, scrambled order. After ciliate mating, the parental macronucleus provides genetic information for the formation of a new macronucleus, which is derived from a newly formed diploid, micronuclear precursor. During this micronucleus to macronucleus transition, the micronuclear genome is modified drastically through various processing events to yield small, gene-sized molecules called nanochromosomes. During the multi-stage macronuclear development process, the near 1 Gb micronuclear genome is reduced to roughly 5% of its original DNA sequence complexity, with the resulting macronuclear nanochromosomes varying greatly in copy number. In addition to excision and removal of nongenic DNA and transposable elements (TEs), the developing macronucleus must sort and reorder the remaining coding segments into functional genic open reading frames for healthy vegetative growth of the cells. Here, I investigate the potential roles of small RNAs (sRNAs) in these macronuclear development processes, with an emphasis on a class called 27macRNAs, highly expressed after *Oxytricha* mating, during the early stages of macronuclear development. In addition, to explore the genes required for this complex macronuclear development process, I performed next generation sequencing of *Oxytricha* mRNAs from vegetative cells and from various timepoints during ciliate mating and macronuclear development. We have identified at least 5 regulatory groups or modules consisting of genes whose expression is highly co-regulated during the distinct stages of this process. We find that a disproportionate number of the mRNAs upregulated during this process encode for proteins involved in DNA and RNA metabolism, with the majority of these genes encoding evolutionarily conserved proteins across species.

Dedication

This dissertation is dedicated to my late grandparents Dr. Richard K. Morgan and Mrs. Chris B. Morgan, who were a driving force and constant inspiration throughout the work performed. In addition, I would also like to dedicate this to my loving parents Cindy Neeb and Tom Neeb and to my wonderful sister Micaela Neeb, without whom none of this work would have been possible.

Acknowledgements

I would like to properly acknowledge those that played a critical role in making this research a possibility. First, I would like to thank Dr. Alan M. Zahler for being an outstanding and patient mentor throughout my thesis work. He has trained me to think critically about science and about life and has offered his expertise to help me become a better scientist than I ever thought possible. Second, I would like to thank Dr. Yaeta Endo, who provided hands on training that contributed to my research as well as always offering very helpful and thoughtful advice. Dr. Daniel Hogan was integral to the bioinformatic analyses of the mRNA sequencing performed in Chapter 3 of this dissertation as well as contributing incredible insight to the molecular processes occurring throughout ciliate development. Dr. Melissa Jurica and Dr. Jeremy Sanford, my other thesis committee members, contributed great advice and participated in interesting, thought provoking conversations throughout my thesis work as well. I would like to thank the other graduate students in my cohort, especially Dr. James Matt Ragle, with whom I shared a lab for the last five years. Matt was a significant inspiration to me during my time in the Zahler lab and always encouraged me not to give up, especially during the troublesome and frustrating times. I would also like to acknowledge two former undergraduate researchers turned research technicians, Athena Lin and Cameron Ferguson, who were instrumental in the development of the model system and maintenance of our model organism *Oxytricha trifallax*. Lastly, I would like to thank Dr. Sol Katzman for help with our bioinformatic analyses, John Paul Donohue and Hiram Clawson for

their help setting up our UCSC *Oxytricha* Genome Browser, and the rest of the previous and current Zahler lab mates I have had the pleasure to work with over the last five years.

Chapter 1

Introduction: Novel roles for RNAs in the ciliated protozoans

In accordance with the central dogma of molecular biology, RNA has long been thought of as simply an intermediate by which genes are expressed. However, in recent years, evidence has emerged to illustrate the incredibly diverse functionality of small RNAs (sRNAs) in many different aspects of gene regulation and development. These 20–30 nucleotide (nt) RNAs can silence the expression of genes by interactions with mRNAs (microRNAs and silencing RNAs) and are important for gene regulation through chromatin modification (piRNAs and siRNAs) (Luteijn and Ketting, 2013; Wilson and Doudna, 2013). In the ciliated protozoans, sRNAs have been shown to be involved in the epigenetic transmission of information between maternal nuclei and their derivatives (Nowacki et al., 2011).

The mechanism of biogenesis of many of the different classes of small RNAs are varied and many are still being worked out, but they are often processed from a double-stranded region of RNA that is generated either by fold-back secondary structure, by RNA-dependent RNA polymerases or by annealing of complementary RNAs (Luteijn and Ketting, 2013; Wilson and Doudna, 2013). Generation of many classes of small RNAs from double-stranded regions involves the function of a double-stranded RNA endonuclease of the Dicer family, which produces cleavage product RNAs with characteristic 2 nt 3' overhangs.

Unlike typical eukaryotic genomes that generally code for one or two Dicer/Dicer-like proteins, the ciliated protozoans often have many more homologs.

Paramecium tetraurelia for example has 8 different Dicers; three from the Dicer class and five from the Dicer-like class, which contains only RNAse III domain pairs and lacks the N-terminal helicase domain of dicers (Sandoval et al., 2014). Previous work with Dicer-like proteins lacking the N-terminal helicase domains of traditional Dicers suggests that these domains are not required for catalytic activity and may not be necessary for sRNA biogenesis (MacRae et al., 2006). These expansions of Dicer and Dicer-like genes in *Paramecium* appear to arise from gene duplication during evolution (Sandoval et al., 2014). While some are ubiquitously expressed and seem to function in general RNAi mechanisms, specific expression of particular Dicer/Dicer-like proteins at distinct stages of development has been observed, and may provide clues to their potential distinct functions in sRNA biogenesis and function during the large-scale genomic rearrangements that occur after ciliate mating (Neeb and Zahler, 2014).

Stichotrichous ciliates, such as *Oxytricha trifallax,* a large unicellular ciliated protist, contain two separate caches of genetic information: a transcriptionally silent germline micronucleus that is exchanged during matings and a transcriptionally active somatic macronucleus containing hundreds to thousands of amplified gene-sized DNA molecules called "nanochromosomes", from which genes are transcribed during vegetative growth of the cells (Prescott, 1994). These macronuclear nanochromosomes are the smallest known DNA molecules in nature, on average 2 kb, and are present at 100-100,000 copies per macronucleus (Aeschlimann et al., 2014; Swart et al., 2013). The micronuclear genome closely resembles that of a canonical eukaryotic genome with many genes organized on long chromosomes. However, micronuclear genes are typically interrupted by many short non-genic DNA sequences called internally eliminated sequences (IESs). For approximately 3,500 of these genes (~20% of genes in the genome), the macronuclear destined sequences (MDSs) that are connected upon IES removal exist in a non-linear, scrambled order (Chen et al., 2014). For example, the actin I gene exists as ten separate germline segments whose order and orientation must be unscrambled to produce a functional gene-coding sequence (Greslin et al., 1989; Prescott 1999). When a mating occurs under the desired environmental conditions, which typically involves the presence of complementary mating types under starvation, two ciliates conjugate and each cell's micronucleus undergoes a series of meiotic divisions. One of the four meiotically derived haploid micronuclei is exchanged between each of the two mating cells, while the other three are broken down and degraded by an unknown mechanism (Adl and Berger, 2000). This newly acquired haploid micronucleus fuses with an existing haploid micronucleus. This new diploid micronucleus then divides by mitosis. The parental macronucleus then breaks down and one of the two newly formed diploid micronuclei develops into a new macronucleus through a distinctive stage known as the anlagen (Adl and Berger, 2000). It is at this point that the ciliates undergo a polytene chromosome stage, eliminate more than 90% of their noncoding germline genome, fragment their chromosomes, and then sort and reorder the many thousands of non-linear macronuclear destined sequences (MDSs) that remain into functional genes. *De novo* telomere addition and amplification of macronuclear nanochromosomes to the appropriate high copy number, completes the development of a new, functional macronucleus (Figures 1 and 2) (For general reviews of the process of macronuclear development see Chalker and Yao, 2011; Jahn and Klobutcher, 2002; Nowacki and Landweber, 2002; Prescott, 2000).

Macronuclear development has been more extensively studied in the distantly related ciliates *Tetrahymena* and *Paramecium,* where it has been shown that epigenetic information from the parental macronucleus guides the elimination and retention of DNA sequences in the developing macronucleus. In these ciliates, the entire parental micronuclear genome is transcribed bi-directionally to produce long, double-stranded RNAs early on in macronuclear development (Chalker and Yao, 2001). These double-stranded RNA precursors are cleaved by Dicer-like enzymes, to produce a class of small RNAs, called scan RNAs (scnRNAs) (26-31 nt in *Tetrahymena* and 25 nt in *Paramecium*) (Mochizuki and Gorovsky 2004, 2005; Malone et al., 2005). These scnRNAs are proposed to be transported to the parental macronucleus, where those with homologous macronuclear sequence are degraded. The remaining scnRNAs that survive this filtering step, corresponding to micronuclear specific sequences, are then transported to the developing macronucleus where, in association with a PIWI protein, they "scan" the genome and mark IESs for excision and elimination through Histone H3 lysine 9 and lysine 27 methylation (Chalker and Yao, 1996; Duharcourt et al., 1996; Kataoka and Mochizuki, 2011; Liu et al., 2007; Mochizuki et al., 2002; Mochizuki and Gorovsky, 2004; Taverna et al., 2002; Yaho and Chao, 2005). Although the mechanism of DNA excision and elimination remains poorly understood, it has been shown to involve a "domesticated" *piggyBac* transposase called PiggyMac (Baudry et al., 2009). A role for iesRNAs, small RNAs complementary to scnRNAs that peak in expression later in macronuclear development, has also been implicated in genome quality control, helping to ensure the full removal of all IESs matching these sequences from the amplified chromosomes in *Paramecium* (Sandoval et al., 2014). It is also worth noting that while *Tetrahymena* and *Paramecium* eliminate IESs during macronuclear development, these ciliates do not possess scrambled micronuclear genes like the stichotrichs and their macronuclear chromosomes are much larger, coding for hundreds of genes, instead of just one or two, typical of *Oxytricha* nanochromosomes (Prescott, 1994; Swart et al., 2013).

Previous work has illustrated the potential roles of RNAs in mediating IES recognition/removal and the unscrambling events that ultimately take place during *Oxytricha* macronuclear development as well. Although the junctions of MDSs and IESs contain short "pointer" sequences that are likely involved, they seem to act as more of a structural requirement for unscrambling and DNA splicing, rather than for recognition by the necessary protein machinery (Prescott and Dubois, 1996). Instead, maternal guide RNA templates that are transcribed in the maternal macronucleus from the nanochromosomes have been hypothesized to mediate this massive genomic rearrangement process (Prescott et al., 2003). Long sense and antisense RNA transcripts, corresponding to entire macronuclear DNA molecules, can be detected for a brief, 24-hour period of time post-conjugation and these are transported to the newly developing macronucleus to provide guide templates for the correct rearrangement, deletion and sometimes inversion of the micronuclear DNA sequences (Nowacki et al., 2008). Microinjection of synthetic double stranded DNA or RNA versions of alternatively rearranged nanochromosomes into the macronucleus of mating cells leads to changes in the reordering of MDSs, not only in the injected cells, but in offspring as well, suggesting epigenetic inheritance through these RNA templates (Nowacki et al., 2008). In our study, we have identified a novel class of macronuclear-derived 27 nt small RNAs, called 27macRNAs, that are highly upregulated after *Oxytricha* conjugation, peaking at 24 hours post-mixing of complementary mating types (Fang et al., 2012; Zahler et al., 2012). Although these 27mers share some common characteristics with the *Tetrahymena* and *Paramecium* scnRNAs, such as their size and a strong 5' U bias, they are unique in that they are derived from the parental macronucleus as opposed to the micronucleus, and do not possess a 2'-O-CH_3 group modification at their 3' end (Fang et al., 2012; Kurth and Mochizuki, 2009; Zahler et al., 2012). These 27macRNAs have been shown to associate with a PIWI homolog called Otiwi1 and may specify which segments of micronuclear DNA will remain protected from degradation throughout macronuclear development, distinct from the role of scanRNAs that denote which regions will be eliminated (Fang et al., 2012). This makes sense when considering that while *Oxytricha* eliminate upwards of 90% of their germline genome, *Tetrahymena* are only responsible for the elimination of approximately 30% (Chen et al., 2014). The relationship between 27mer PIWI-associated RNAs and the long dsRNA “guide templates” implicated in MDS rearrangements remains unknown.

Within the last five years, both the *Oxytricha* macronuclear and micronuclear genomes have been sequenced and made publicly available. The macronuclear genome consists of approximately 50 Mb of DNA on close to 16,000 individual and distinct nanochromosomes (Swart et al., 2013). The vast majority of these nanochromosomes only code for a single protein, although in rare cases as many as eight, comprising the ~18,500 total genes present in the genome. The micronuclear genome on the other hand, contains roughly 1 Gb of DNA, most of which is made up of IESs, transposable elements and repetitive sequences that are eliminated during the micronuclear to macronuclear transition (Chen et al., 2014). The micronuclear genome is fragmented into over 200,000 MDSs, with more than 18% of macronuclear-destined genes being scrambled. There is also evidence to suggest that there are some micronuclear genes expressed solely during the later stages of macronuclear development that are not present on the nanochromosomes that ultimately make up the macronuclear genome (Chen et al., 2014).

To date, although the general timing of events involved in macronuclear development has been fairly well characterized, the molecular mechanisms underlying many of these processing events remain poorly understood. To examine this, we have performed high throughput mRNA sequencing of the genes expressed during a time course of macronuclear development. In this study, we have identified at least 5 distinct regulatory modules of highly co-regulated genes differentially expressed during macronuclear development, corresponding to the timing of specific events. We find that a disproportionate number of the genes identified as upregulated encode proteins that are involved in DNA and RNA metabolism processes. The majority of these genes encode evolutionarily conserved proteins. Paralogous gene sets were also identified in which one paralog is expressed during macronuclear development and the other only during vegetative growth, indicating the presence of gene duplication events that ultimately led to functional specialization of these genes. Paralogous gene families include components of the RNA Polymerase II transcriptional machinery, transcription factors, replication factors, histones, and DNA and RNA helicases, among others. In addition, a striking number of differentially expressed macronuclear development genes are preferentially expressed in animal germline cells, illustrating that we have identified a highly conserved and primordial set of factors involved in germline and stem cell maintenance.

Figure 1: *Oxytricha* sexual life cycle. 1. Two vegetative *Oxytricha* cells of different mating types (represented by the difference in nuclei colors). **2.** Under starvation conditions the two ciliates fuse and begin to conjugate. **3.** The parental micronucleus in each cell undergoes meiosis. **4-6.** Three of the newly formed haploid micronuclei will break down while the one remaining will undergo mitosis. One mitosis-derived haploid micronucleus is exchanged between the mating cells. **7.** The newly acquired micronucleus fuses with the remaining maternal micronucleus to become diploid. **8.** The newly formed diploid micronucleus undergoes mitosis. **9.** One newly formed micronucleus develops into a new macronucleus while the maternal macronucleus is broken down and degraded. **10.** Two genetically identical exconjugant *Oxytricha* cells. Inside the circle is a graph showing the general timing of macronuclear development events in hours and the corresponding DNA content of the developing macronucleus (anlagen). Outside of the circle are the alternate MIC and MAC versions of two hypothetical genes, one scrambled and one unscrambled (MDSs in orange, IESs and nongenic DNA in blue and telomeres in black).

Oxytricha **Sexual Life Cycle**

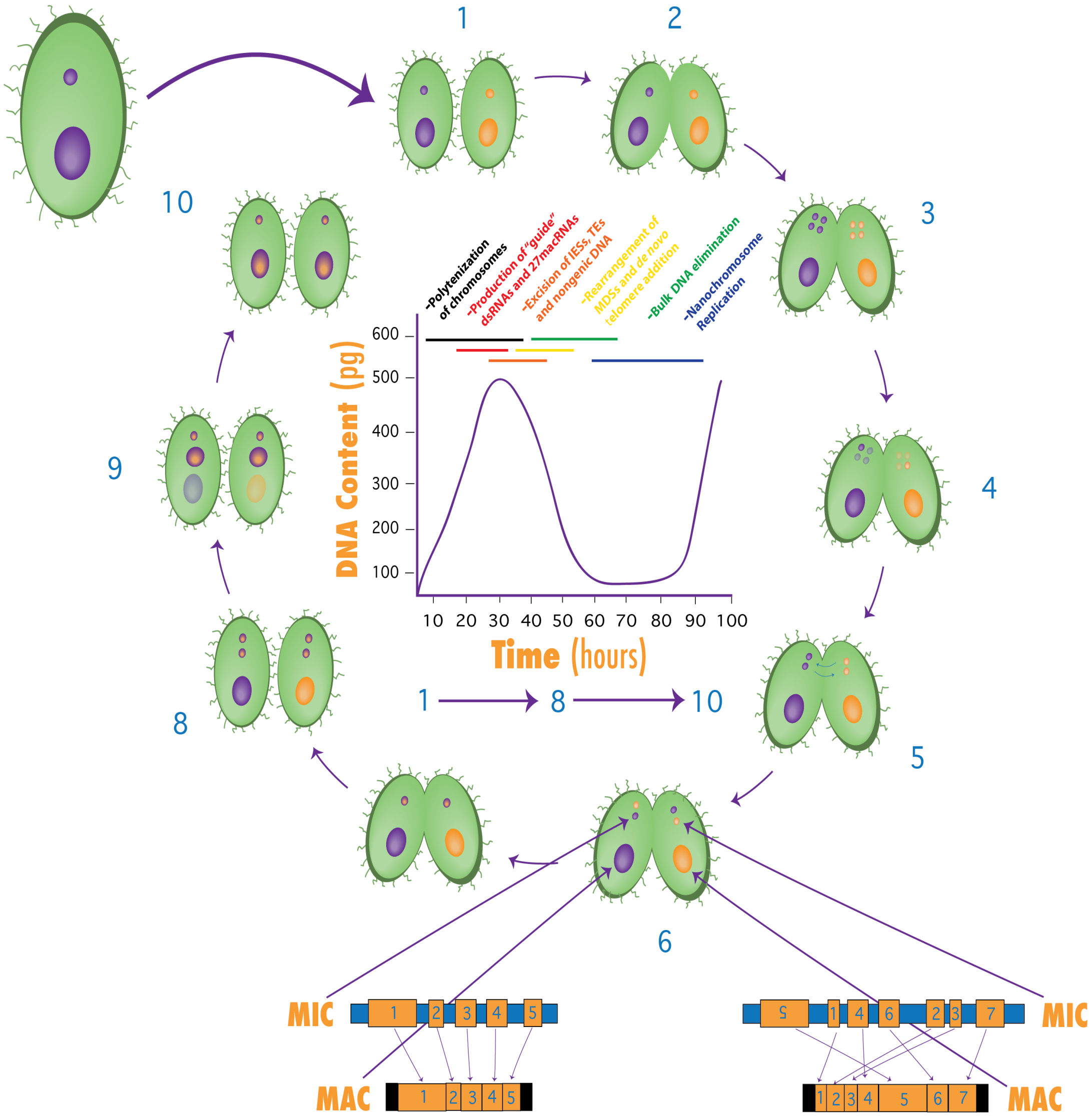
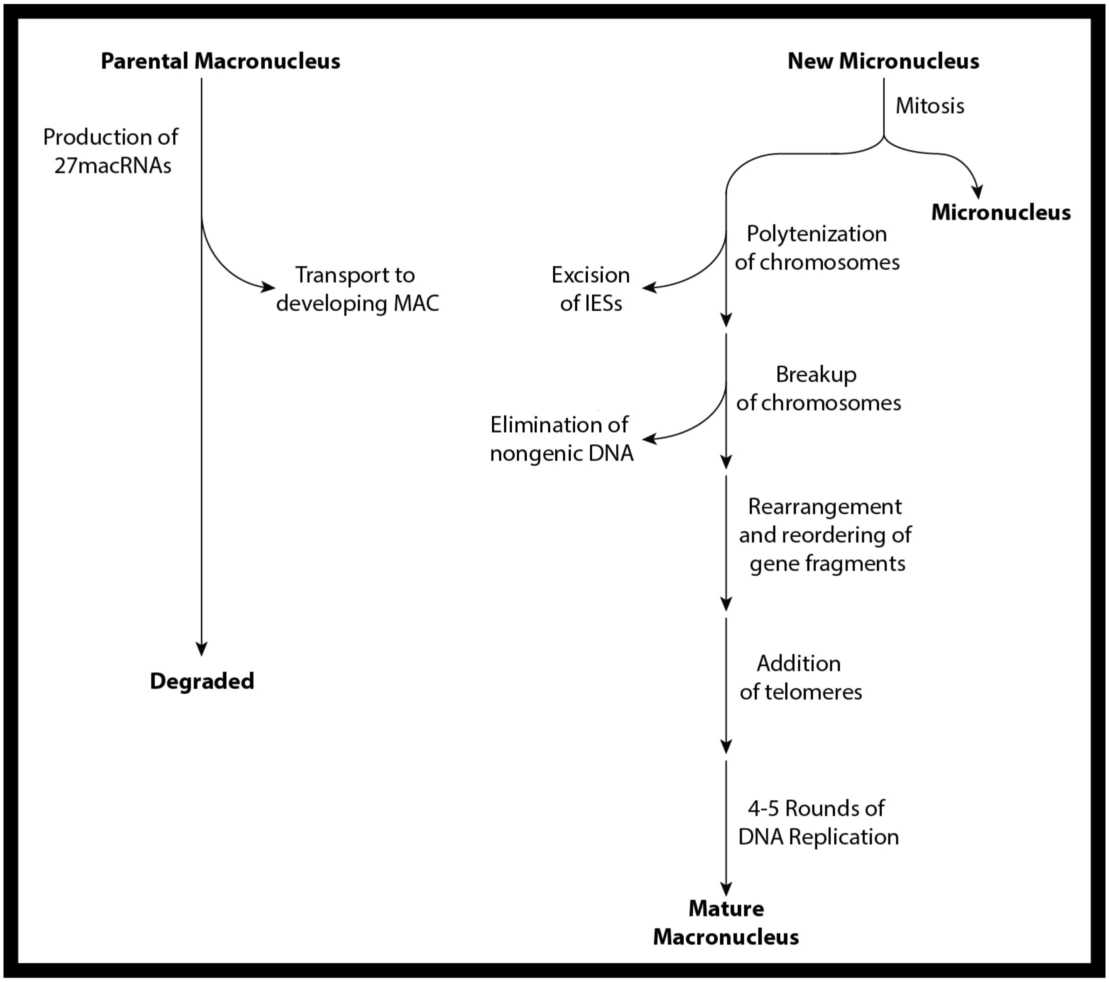

1
2
3
4
5
6
7
8
9
10
-Polytenization of chromosomes
-Production of "guide" dsRNAs and 27macRNAs
-Excision of IESs, TEs and nongenic DNA
-Rearrangement of MDSs and *de novo* telomere addition
-Bulk DNA elimination
-Nanochromosome Replication
DNA Content (pg)
600
500
400
300
200
100
10 20 30 40 50 60 70 80 90 100
Time (hours)
1 → 8 → 10
MIC
1 2 3 4 5
MAC
1 2 3 4 5
5 1 4 6 2 3 7
MIC
1 2 3 4 5 6 7
MAC

Figure 2: Flowchart of *Oxytricha* Macronuclear Development.



References

Adl SM, Berger JD (2000) Timing of life cycle morphogenesis in synchronous samples of *Sterkiella histriomuscorum*. II. The sexual pathway. *J Eukaryot Microbiol* 47: 443–449.

Aeschlimann, S.H., Jonsson, F., Postberg, J., Stover, N.A., Petera, R.L., Lipps, H.J., Nowacki, M., and Swart, E.C. (2014) The draft assembly of the radically organized Stylonychia lemnae macronuclear genome. *Genome Biol Evol* 6:1707-1723.

Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, Meyer E, Bétermier M (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. *Genes Dev* 23(21):2478-83.

Chalker D, Yao MC (1996) Non-Mendelian, heritable blocks to DNA rearrangement are induced by loading the somatic nucleus of *Tetrahymena thermophila* with germ line-limited DNA. *Mol Cell Biol* 16(7):3658.

Chalker D, Yao MC (2001) Nongenic, bidirectional transcription precedes and may promote developmental DNA deletion in *Tetrahymena thermophila. Genes Dev* 15:1287-1298.

Chalker D, Yao MC (2011) DNA elimination in ciliates: transposon domestication and genome surveillance. *Annu Rev Genet* 45:227–46.

Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM, Swart EC, Perlman DH, Doak TG, Stuart A, Amemiya CT, Sebra RP, Landweber LF (2014) The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell* 158(5):1187-98.

Duharcourt S, Butler A, Meyer E (1995) Epigenetic self-regulation of developmental excision of an internal eliminated sequence on *Paramecium tetraurelia*. *Genes Dev* 15;9(16):2065–77.

Fang W, Wang X, Bracht JR, Nowacki M, Landweber LF (2012) Piwi-interacting RNAs protect DNA against loss during Oxytricha genome rearrangement. *Cell* 151:1243–1255.

Greslin AF, Prescott DM, Oka Y, Loukin SH, Chappell JC (1989) Reordering of nine exons is necessary to form a functional actin gene in Oxytricha nova. *Proc Natl Acad Sci USA* 86(16):6264-8.

Jahn CL, Klobutcher LA (2002) Genome remodeling in ciliated protozoa. *Annu Rev Microbiol* 56: 489–520.

Kataoka K, Mochizuki K (2011) Programmed DNA elimination in *Tetrahymena*: a small RNA-mediated genome surveillance mechanism. *Adv Exp Med Biol* 722: 156–173.
Liu Y, Taverna SD, Muratore TL, Shabanowitz J, Hunt DF, et al. (2007) RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in *Tetrahymena. Genes Dev* 21: 1530–1545.

Kurth H, Mochizuki K (2009) 2'-O-methylation stabilizes PIWI-associated small RNAs and ensures DNA elimination in *Tetrahymena. RNA* 15:675-685.

Luteijn MJ, Ketting RF (2013) PIWI-interacting RNAs: from generation to transgenerational epigenetics. *Nature Reviews Genetics* 14:523–534.

MacRae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, Adams PD, Doudna JA (2006) Structural basis for double-stranded RNA processing by Dicer. *Science* 311:195–198.

Malone CD, Anderson AM, Motl JA, Rexer CH, Chalker DL (2005) Germ line transcripts are processed by a Dicer-like protein that is essential for developmentally programmed genome rearrangements of *Tetrahymena thermophila. Mol Cell Biol* 25:9151-9164.

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) Conjugation-specific small RNAs in *Tetrahymena* have predicted properties of scan (scn) RNAs involved in genome rearrangement. *Genes Dev* 18:2068-2073.

Mochizuki K, Gorovsky MA (2005) A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev* 19:77-89.

Neeb ZT, Zahler, AM (2014) An expanding world of small RNAs. *Dev Cell* 28:111-112.

Nowacki M, Landweber LF (2009) Epigenetic inheritance in ciliates. *Curr Opin Microbiol* 12: 638–643.

Nowacki M, Shetty K, Landweber LF (2011) RNA-Mediated Epigenetic Programming of Genome Rearrangements. *Annu Rev Genomics Hum Genet* 12:367-389.

Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, et al. (2008) RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature* 451: 153–158.

Prescott DM (1994) The DNA of ciliated protozoa. *Microbiological reviews* 58:233-267.

Prescott DM (1999) The evolutionary scrambling and developmental unscrambling of germline genes in hypotrichous ciliates. *Nucleic Acids Res* 27(5):1243-50.

Prescott DM (2000) Genome gymnastics: unique modes of DNA evolution and processing in ciliates. *Nat Rev Genet* 1: 191–198.

Prescott DM, Ehrenfeucht A, Rozenberg G (2003) Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates. *J Theor Biol* 222(3):323-30.

Prescott DM, Dubois ML (1996) Internal eliminated segments (IESs) of *Oxytrichidae. The Journal of eukaryotic microbiology* 43:432-441.

Sandoval PY, Swart EC, Arambasic M, Nowacki M (2014) Functional diversification of Dicer-like proteins and small RNAs required for genome sculpting. *Dev Cell* 28(2):174-88.

Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, Khurana JS, Goldman AD, Nowacki M, Schotanus K, Jung S, Fulton RS, Ly A, McGrath S, Haub K, Wiggins JL, Storton D, Matese JC, Parsons L, Chang WJ, Bowen MS, Stover NA, Jones TA, Eddy SR, Herrick GA, Doak TG, Wilson RK, Mardis ER, Landweber LF (2013) The Oxytricha trifallax macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biol* 11(1):e1001473.

Taverna SD, Coyne RS, Allis CD (2002) Methylation of histone h3 at lysine 9 targets programmed DNA elimination in *Tetrahymena. Cell* 110: 701–711.

Wilson RC, Doudna JA (2013) Molecular mechanisms of RNA interference. *Annual Review of Biophysics* 42:217–39.

Yao MC, Chao JL (2005) RNA-guided DNA deletion in *Tetrahymena*: an RNAi-based mechanism for programmed genome rearrangements. *Annu Rev Genet* 39: 537–559.

Zahler AM, Neeb ZT, Lin A, Katzman S (2012) Mating of the stichotrichous ciliate *Oxytricha trifallax* induces production of a class of 27 nt small RNAs derived from the parental macronucleus. *PLoS ONE* 7(8):e42371. doi: 10.1371/journal.pone.0042371.

CHAPTER 2

Mating of the Stichotrichous Ciliate *Oxytricha trifallax* Induces Production of a Class of 27 nt Small RNAs Derived from the Parental Macronucleus

Introduction

Ciliated protozoans are characterized by nuclear dimorphism. These large single-celled organisms possess two types of nuclei; the macronucleus undergoes active transcription (the somatic nucleus) while the micronucleus, which is not transcribed, serves as the genetic repository (the germ line nucleus). During sexual reproduction, the parental micronuclei undergo meiosis and haploid micronuclei are exchanged between conjugating (mating) cells to form a new diploid micronucleus. This new diploid micronucleus divides by mitosis, and one of its daughter nuclei develops into the new macronucleus. This new developing macronucleus is referred to as the anlage. While the new macronucleus develops from the anlage, the parental macronucleus is destroyed. The process of macronuclear development from a diploid micronucleus requires extensive DNA amplification and polytenization, followed by elimination of micronucleus-specific sequences. This elimination results in fragmentation to smaller chromosomes, to which telomeres are added *de novo*. DNA elimination also occurs within macronuclear-destined regions resulting in splicing out of internally eliminated sequences (IESs) and rejoining of macronuclear-destined sequences (MDSs) (for general reviews of the process of macronuclear development see Prescott, 2000, Jahn and Klobutcher, 2002, Nowacki and Landweber, 2002, Chalker and Yao, 2011).

Among the ciliates, the stichotrichous ciliates (including *Stylonichia*, *Euplotes* and *Oxytricha* species), take the processing of micronuclear sequences into macronuclear chromosomes to an extreme. Using re-association kinetics, it was measured that the macronuclear genome possesses only 5% of the sequence complexity of the micronuclear genome (Lauth et al., 1976). While the micronuclear chromosomes of this group are typical of eukaryotes in terms of chromosome length, structure and mitotic division, the macronuclear genome consists of over 20,000 different chromosomes with an average length of ~2,200 bp (Swanton et al., 1980). The majority of these "nanochromosomes" contain only a single gene (Prescott et al., 2002, Cavalcanti et al., 2004) and are present in over 1000 copies per nucleus Lauth et al., 1976). The micronuclear versions of macronuclear *Oxytricha* genes are not capable of being expressed without extensive DNA processing that occurs during macronuclear development. This includes the removal of internally eliminated sequences (IESs) followed by splicing together of the surrounding macronuclear destined sequences (MDSs). These IESs total over 100,000 in number (Prescott, et al., 2002). At the junctions at which the MDSs are joined, there are short direct repeats referred to as "pointers". Only one copy of each pointer pair from the micronuclear DNA is found in the macronucleus, suggesting a potential role for a homology-directed DNA repair mechanism in the process of IES elimination (Prescott and DuBois, 1996). In addition to IES elimination, an added complexity of macronuclear development in the stichotrichous ciliates is that some genes have a different linear ordering of MDS segments in the micronucleus than in the macronucleus. The actin gene was the first of these scrambled micronuclear genes to be identified, and it is interesting in that MDS segments are not only out of order in the micronuclear genome, but MDS2 is actually found on the opposite strand from the others (Greslin et al., 1989). The alpha telomere binding protein and DNA polymerase alpha are two other highly scrambled micronuclear genes that have been characterized (Mitcham et al., 1992, Hoffman and Prescott, 1996).

In stichotrichous ciliates, and in the Oligohymenophorea ciliates, which include *Tetrahymena* and *Paramecium*, there is strong evidence that the parental macronucleus provides information to guide the developing macronucleus (Chalker and Yao, 2011, Kataoka and Mochizuki, 2011). In essence, the parental macronucleus serves as an epigenetic guide to daughter macronuclear formation. In both groups of ciliates, RNA serves as the mediator of epigenetic control (Nowacki et al., 2011). In *Tetrahymena*, the diploid micronucleus is transcribed bi-directionally (Chalker and Yao, 2001), and transcripts are processed by a dicer homolog, Dcl1p, to produce a 28–29 nt long class of small RNAs, called scanRNAs (Mochizuki and Gorovsky, 2005, Mochizuki and Gorovsky, 2004, Malone et al., 2005). These scanRNAs are loaded into a complex with a Piwi homolog called Twi1p (Mochizuki and Gorovsky, 2004, Mochizuki et al., 2002). The scanRNA complexes are sent to the parental macronucleus, where those that have a match to sequences in the parental macronucleus are selectively removed from the pool that will guide IES excision

(Chalker and Yao, 2011, Kataoka and Mochizuki, 2011). Those that survive this parental macronuclear filtering step are then imported to the developing macronucleus, where they direct chromatin modifications that target micronuclear-specific sequences for degradation (Chalker and Yao, 2011, Kataoka and Mochizuki, 2011, Yao and Chao, 2005, Liu et al., 2007, Taverna et al., 2002). In the stichotrichous ciliate *Oxytricha trifallax*, it has been demonstrated experimentally that long RNAs are produced bi-directionally from the nanochromosomes of the parental macronucleus soon after mating. Experimental injection of single-stranded RNAs into the cytoplasm of mating cells can reprogram IES removal and MDS joining, suggesting that these long RNAs serve as a guide for MDS joining and gene unscrambling (Nowacki et al., 2008). The resulting nanochromosomes are differentially amplified in the developing macronucleus. Epigenetic information concerning nanochromosome amplification levels in the parental macronucleus is transferred to the developing macronucleus via RNA intermediates (Heyse et al. 2010, Nowacki et al., 2010).

We tested for the presence of small RNA species in the stichotrichous ciliate *Oxytricha trifallax* that are specifically produced during mating. We identified a class of 27 nt long RNAs that are expressed at high levels 24 hours after mating induction, and which decrease steadily during the subsequent steps of macronuclear development. While the scanRNAs of *Tetrahymena* are modified with a 2′O-methyl group at their 3′ ends (Kurth and Mochizuki, 2009), the 27 nt class produced in *Oxytricha* have 2′ and 3′ hydroxyl groups at their 3′ ends. We performed next generation sequencing of small RNAs from vegetative cells and mating cells over a time course after mixing cells of mating-competent strains. We demonstrate that the 27 nt RNAs originate from macronuclear chromosomes and are transcribed from both strands. Their distribution along the nanochromosomes is non-uniform, and for either strand the positions proximal to the telomeric repeats are much lower in small RNA coverage. We name these 27 nt mating-specific small RNAs “27macRNAs”. We propose several models for the roles that 27macRNAs may play during macronuclear development.

Materials and Methods

Vegetative growth of *Oxytricha trifallax*

Oxytricha trifallax strains JRB310 and JRB510 (Williams et al., 1993, Zoller et al., 2012) were kindly provided by Robert O. Hammersmith (Ball State University). Vegetative growth of cells was carried out in Pyrex dishes in inorganic salts media (Chang et al., 2004) using the algae *Chlorogonium elongatum* (UTEX Collection Strain B203) as a food source.

Mating of *Oxytricha trifallax*

Mating competent strains were grown separately until they had nearly exhausted their food. Cells were filtered through cotton to remove algal debris and were then concentrated on 10 µM Nitex filters. Cells were washed off of the filters into inorganic salts media and the individual cultures were counted. Cells of the different mating strains were mixed together in equal numbers, concentrated again on Nitex filters and washed on the filters into Pringsheim salts buffer. Timing of mating begins at this mixing step. Cells were put into Pyrex dishes at a concentration of 1200–2000 cells/ml in Pringsheim solution with 1 ml of stationary *Klebsiella pneumoniae* culture per 300 ml of mating cells as food. Cells began to show strong levels of aggregation by 5 hours and mating pairs were first visible after 8 hours. In order to refresh the JRB310 and JRB510 stocks to provide for a more robust mating, exconjugants with anlage were individually isolated from a JRB310 x JRB510 mating and these clonal isolates were grown as cultures. After six weeks of vegetative growth, we tested the mating compatibility of each of 12 clonal cultures by pairwise mixing. Two of the new strains, ALXC2 and ALXC9 showed strong mating ability with each other, and their progeny had high survival rates. ALXC2 cultures have a tendency to self mate if starved, and surprisingly these self-progeny are also viable. No ALXC9 selfing is evident. When equal numbers of ALXC2 and ALXC9 are mixed together, a highly efficient mating occurs, with ~70% of cells possessing anlage 48 hours after mixing.

Total RNA isolation

Oxytricha trifallax cultures were washed through cotton and then concentrated on 10 µM Nitex filters. Cells were put into microcentrifuge tubes and gently pelleted at 1000 x g for 2 minutes. Supernatants were removed except for the 50 µl above the cell pellet. 200 µl of mirVana Lysis/Binding Buffer from the mirVana miRNA Isolation Kit (Ambion) was added to each tube. Total RNA was purified according to the kit's protocols for total RNA purification (not the miRNA purification protocol). Typical total RNA yields from this protocol starting with 300 ml of *Oxytricha trifallax* culture were 100 µg.

^{32}P labeling of total RNA

3 µg of total RNA were treated with 30 units of calf intestine alkaline phosphatase (New England Biolabs) in a 100 µl reaction mixture using the

manufacturer's protocol. The reaction mixture was acid phenol:CHCl3 extracted, Na-Acetate added to 0.3 M and ethanol precipitated with 4 volumes of ethanol and 1 µl glycoblue (Ambion) as carrier. Precipitate pellets were washed with 70% ethanol, dried and then resuspended in 10 µl of deionized H_2O. ^{32}P labeling was done using T4 polynucleotide kinase (NEB) and gamma-^{32}P-ATP in 10 µl reaction mixtures. An equal volume of formamide dyes were added to the reaction mixtures and then heated to 95°C. Labeled RNA samples were separated on 40 cm long 15% acrylamide (19∶1) urea gels in TBE buffer. ^{32}P labeled Decade marker (Ambion) was used for a 10 nt ladder size reference. Gel images were recorded using a Typhoon PhosphorImager (GE Health Care) and ImageQuant software.

Beta elimination assay

The beta elimination assay was used to test for modifications at the 3′ end of RNAs. ^{32}P labeled total RNA was prepared and reaction mixtures were raised to 95°C for 2 minutes after the kinase reaction in order to inactivate the T4 polynucleotide kinase enzyme. In addition, 250 pmoles of a 21 nt synthesized RNA with the *C. elegans* lin-4 microRNA sequence and 250 pmoles of a 23 nt synthesized RNA with the *C. elegans* mir-90 sequence were ^{32}P labeled at their 5′ ends in 10 µl reaction mixtures to serve as positive controls for beta elimination, as these synthetic RNAs have a 2′ and 3′ OH group on their 3′ terminal ribose. After labeling and heat inactivation, 1.5 µg of unlabeled total *Oxytricha* RNA was added to each tube so that each control RNA was treated in the beta elimination assay in the same complex

RNA background as the labeled total RNA. Beta elimination reaction conditions were essentially those described by Horwich et al. (Horwich et al., 2007). 5 µl of heat inactivated ^{32}P kinase reaction mixture were added to 100 µl of 25 mM Na-meta-periodate in 60 mM Borax/60 mM Boric acid buffer pH 8.6. Tubes were incubated at room temperature for 30 minutes. Then, 13 µl of 1 M NaOH was added to each tube in order to raise the pH from 8.6 to 9.5. Reaction mixtures were then incubated for 90 minutes at 45°C to allow the beta elimination reaction to occur. RNA was recovered after acid phenol:CHCl3 extraction and ethanol precipitation, and dried pellets were raised in 10 µl of formamide loading dye. Products were separated on 15% polyacrylamide urea sequencing gels and visualized by autoradiography.

Small RNA cDNA Library Preparation

Small RNAs had linkers ligated to them and bar-coded cDNAs were prepared using the TruSeq Small RNA Sample Prep Kit (Illumina) following the manufacturer's instructions. Library preparation began with 1 µg of total RNA from mating cells in which small RNAs are abundant, or 2 µg of total RNA from vegetative cells in which small RNAs (especially the 27 nt species) are rare. Individual libraries were analyzed on a BioAnalyzer (Agilent) for the presence of linkered cDNA at the appropriate size (135–165 bp) and 11 bar-coded libraries were pooled into one sample by mixing 2.0 ng of the 135–165 bp peak from each sample as determined by the BioAnalyzer. The 135–165 bp peak of the pooled cDNAs was purified from the mixed sample using the Pippin Prep DNA Size Selection System (Sage Science), and confirmed by BioAnalyzer.

Illumina Sequencing

Sequencing of the pooled libraries was performed in one lane of the Illumina HiSeq2000 Sequencer at the UCSC Genome Technology Center. 100 bp paired-end reads of the libraries were obtained. After indexing and trimming of linker sequences, those reads of at least 16 nt in length that had 100% identity in the two directions (77% of the total) were further analyzed.

Non-coding RNA filter

We extracted *Oxytricha trifallax* non-coding RNA sequences from the list of non-coding RNAs published by Jung et al. (Jung et al., 2011). This included rRNAs, tRNAs, snRNAs, telomerase RNA and other non-coding RNA species. We also included a 40 base telomeric repeat sequence in this filter $(GGGGTTTT)_5$ to filter out any RNAs derived from telomeres. This non-coding RNA filter was used as a first pass filter of the libraries prior to alignment to the macronuclear genome.

Preparation of a macronuclear reference genome

In order to be able to understand where the 27 nt small RNAs originate, an assembly of macronuclear sequences was needed. While the macronuclear genome sequence of *Oxytricha trifallax* is currently incomplete, an extensive assembly of whole genome shotgun sequencing contigs, WGS2.1.1, was released in the supplemental material of Jung et al. (Jung et al., 2011). From that data we extracted the sequences of 10,137 full-length telomere-to-telomere nanochromosome sequences. We concatenated these nanochromosome sequences together into one file with 50 Ns inserted in between each full-length nanochromosome. This file is referred to as chr1 and contains only full-length nanochromosomes. An additional 46,417 contigs of incomplete nanochromosome sequence were extracted from WGS2.1.1 and these were also concatenated with 50 Ns inserted in between each contig. This collection of partial nanochromosome sequences is referred to as chr2. We noted that there is a small amount of gene duplication on chr1 (~5%) and that there is a good deal of sequence duplication within the incomplete contigs on chr2 and between chr1 and chr2. The 70 kb *Oxytricha trifallax* mitochondrial genome sequence (Swart et al., 2012) was used to analyze small RNAs (referred to as chrM). chr1 and chr2 were also used as reference genome sequences in a build of a minimal UCSC Genome Browser (Kent et al., 2002, Fujita et al., 2011) on a local Linux computer, which allowed for visualization of Illumina sequencing reads through Bam mapping (Li et al., 2009) and BLAT functionality (Kent, 2002).

Analysis Pipeline

Bowtie (Langmead et al., 2009) was used to first align the sequencing reads (trimmed and selected for equality of the two ends as noted above) from each library to the non-coding RNA filter described above. The reads that were not caught by this filter were aligned to the concatenated full-length nanochromosomes sequences (chr1) and to chrM. The remaining unmapped reads were aligned to the concatenated partial nanochromosomes (chr2). More stringent criteria were used for the mapping to the chromosomes (Bowtie parameters: -n 1 –e 60) than were used for the filtering of ncRNA reads (Bowtie parameters: -n 3 –e 150). For reads that mapped to multiple locations, a single, randomly selected, mapping was retained for the subsequent statistical analysis. For visualization of libraries on the genome browser, all mappings of multi-hit reads were retained. For the comparison of micronuclear and macronuclear origins of the RNA, the reads that were not caught by the non-coding RNA filter were mapped to the "micro/macro" reference described below, with all mappings retained.

Alignment of 27 nt small RNAs to micronuclear/macronuclear sequence pairs

In order to determine whether the 27 nt small RNAs were macronuclear or micronuclear in origin, we compared 26–28 nt sequence reads that passed through the non-coding RNA filter for the ability to align to a micronuclear version of a gene as compared to its macronuclear counterpart. To do this, we used the micronuclear and macronuclear versions as Bowtie filters for the sequence reads, and identified and counted the number of sequence reads that aligned to each set in the pair. This allowed us to create a Venn diagram (Table 2) of small RNAs that aligned to both the micronuclear and macronuclear versions, only the micronuclear version or only the macronuclear version. Gene pairs tested in this assay were trimmed so that all macronuclear sequence was contained within the micronuclear clone, and the regions of the micronuclear clone outside of the macronuclear sequences were removed. Essentially, the only differences between the micronuclear and macronuclear sequences were the presence/absence of IESs and the scrambled nature of the MDS order for some of the genes tested. Sequences used for this experiment were as follows: DNA Polymerase Alpha - micronuclear GenBank accession DQ525914.1 nucleotides 1837-9001 and macronuclear GenBank accession U59426.1 nucleotides 20-4665; Actin-I - micronuclear GenBank accession U19288.1 nucleotides 1-2115 and macronuclear GenBank accession HQ432909.1 nucleotides 1-1503; Zinc Finger gene - micronuclear GenBank accession FJ346576.1 nucleotides 1-2128 and macronuclear GenBank accession FJ346577.1 nucleotides 90-2084; L29Cyclo - micronuclear GenBank accession DQ081723.1 nucleotides 1-1811 and macronuclear GenBank accession DQ081724.1 nucleotides 21-1616; TEBPAlpha - micronuclear GenBank accession EU047939.1 nucleotides 1-2787 and macronuclear GenBank accession EU047938.2 nucleotides 21-2147, TEBPBeta - micronuclear GenBank accession EU047941.2 nucleotides 1-1753 and macronuclear GenBank accession EU047940.2 nucleotides 1–1250.

Availability of sequence data

Raw sequence files for all 11 Illumina sequencing libraries, and files listing the sequences of 26–28 nt RNAs that survived the non-coding RNA filter from each of the 7 mating libraries, have been deposited in GEO - Accession Number GSE37390.

Results

Oxytricha trifallax produces a class of 27 nt small RNAs during macronuclear development

Small RNAs known as scanRNAs play an important role in the DNA rearrangements that occur during the development of the macronucleus in *Tetrahymena* (Chalker and Yao, 2011, Kataoka and Mochizuki, 2011). We asked whether small RNA species are present in vegetative stichotrichous ciliates and in those undergoing macronuclear development. To do this, total RNA was isolated from *Oxytricha trifallax* vegetative cells or cultures of cells in which complementary mating types had been mixed together under mating conditions. Total RNA was treated with calf intestine alkaline phosphatase and then ^{32}P 5′ end labeled using gamma-^{32}P-ATP and T4 polynucleotide kinase. Labeled RNAs were separated on a 15% acrylamide sequencing gel and visualized with a PhosphorImager. As seen in Figure 1, a prominent band at 27 nt appears 24 hours after mixing of the mating competent strains. This band is not detectable 5, 6 or 12 hours after mixing of the mating competent strains (data not shown), but by 24 hours it is quite prominent. At later stages of macronuclear development, the relative intensity of this band fades. Three other bands are faint but detectable in all RNA isolations. These RNAs are 25, 22 and 21 nt in size. As these are not induced by mating, these are not a major focus of this paper. We did a control experiment to confirm that the labeled RNAs were from *Oxytricha* (and not a contaminant from the algal food source) by labeling an equivalent amount of total RNA from the *Chlorogonium elongatum* food source (data not shown). Equivalent amounts of total Chlorogonium RNA, labeled and run in parallel lanes, did not yield detectable bands at these sizes. We also did a mock mating with the ALXC9 strain, starving it and moving it into Pringsheim buffer for 24 hours, to test whether these 27 nt RNAs are induced by starvation as opposed to mating. This treatment, equivalent to the mating procedure but done with only one strain that does not self mate, did not lead to production of 27 nt small RNAs (Figure 1, lane 12).

Mating-specific small RNAs have 2′ and 3′ OH groups at their 3′ ends

Having demonstrated that *Oxytricha* produce a 27 nt species of small RNAs during mating, we asked whether these are similar to the 28–29 nt long scanRNAs that are produced during mating by the distantly related ciliate *Tetrahymena*. One distinguishing feature of scanRNAs is that they have a 2′O-methyl group added to the 3′ terminal ribose of the RNA by a *Tetrahymena* homolog of the Hen1p methylase, and this activity is essential for scanRNA stability and function (Kurth and Mochizuki, 2009). We performed a beta elimination assay to determine if there are modifications present at the 3′ end of the 27 nt *Oxytricha* mating-specific RNA species. If free 2′ OH and 3′ OH groups are present at the 3′ terminal ribose of RNA, then the ribose can be oxidized by periodate and subsequently removed during incubation at higher pH and temperature, resulting in an increase in mobility on a sequencing gel (Yang et al., 2007). Figure 2 shows the results of this assay. The 27 nt mating-specific RNA species is clearly modified by the beta elimination assay (lanes 4 through 8), in a way that is similar to the control RNAs (lanes 13–16). The beta elimination reaction results in the removal of the 3′ nucleotide, and leads to RNAs ending in a 3′ cyclic phosphate; the increased charge per length ratio contributed by the 3′ cyclic phosphate leads to these modified RNAs appearing to run 2 bases faster than their untreated controls. This result suggests that the 27 nt RNA has different properties than the 28–29 nt scanRNAs induced by mating in *Tetrahymena*. The 25 nt RNA species is sensitive to beta elimination as well (lanes 2 and 3, and 10 and 11), indicating that it too lacks modifications of the terminal ribose.

Small RNA Sequencing and Analysis

In order to better understand the small RNA species observed in Figure 1, we performed high-throughput sequencing of the small RNA in these samples. Libraries of small RNAs were prepared for sequencing using the TruSeq small RNA Sample Prep Kit (Illumina) starting with 1–2 micrograms of total RNA from preparations shown in lanes 1–11 of Figure 1. The cDNA library was prepared by the manufacturer's standard protocol, which required that small RNAs possess a 3′OH group and a 5′ monophosphate in order to be ligated to the adapters. After linker attachment and amplification, equal amounts (2 ng) of cDNA representing small RNAs of 15–45 nt in length (as determined by BioAnalyzer (Agilent) analysis) were pooled, and the proper sized cDNAs selected and sequenced in one lane of an Illumina HiSeq2000 Sequencer at the UCSC Genome Technology Center using a 100 base paired-end read protocol. Figure 3 shows a flowchart of the sequence analysis for each library and Table 1 shows our initial analysis of the sequences in each library. Taking advantage of the fact that our lane was on a slide that underwent bidirectional sequencing, and that the lengths of the RNAs under investigation (<45 nt) were shorter than the sequencing length (100 bp), we only processed sequences that were identical in both directions. We used the data from a careful analysis of non-coding RNAs encoded in the *Oxytricha trifallax* macronucleus (Jung et al., 2011) to assemble a filter for non-coding RNAs. The percentage of bi-directional identical reads in each library that derived from non-coding RNA fragments varied between 13% and 43%. Ciliates possess two genomes, a micronuclear genome and a macronuclear genome. While very limited micronuclear sequence is available for *Oxytricha trifallax*, an extensive but incomplete assembly of macronuclear sequence data called WGS2.1.1 is available (Jung et al., 2011). From those sequence data, we extracted the sequences of 10,137 full-length telomere-to-telomere nanochromosome sequences. Given the short nature of nanochromosomes, we decided to concatenate these sequences together into one longer file with 50 Ns inserted between each full-length nanochromosome. An additional 46,417 contigs of incomplete nanochromosome sequence were extracted from WGS2.1.1 and these were also concatenated with 50 Ns inserted between each contig. We also compared the small RNA reads to the published 70 kb mitochondrial genome sequence. The percentage of sequence reads that survive the non-coding RNA filter and map to these sequences are also indicated in Table 1. An initial analysis of the mappings to the mitochondrial genome identified a denser mapping of short RNA reads to the mitochondrial genome relative to the macronuclear genome. However, our initial cursory analysis found that the majority of these mappings were fragments of mitochondrial tRNAs; many classes of mitochondrial non-coding RNAs were not filtered by the non-coding RNA filter step (those non-coding RNAs were encoded by the macronuclear genome). Given the poor mitochondrial non-coding RNA filtering, and the finding that the abundant 27 nt reads from RNAs derived from mating cells were insubstantial in their mapping to the mitochondrial genome (976 of the distinct 27 nt reads matched mitochondrial genome vs. 2.7 million distinct 27 nt reads matching the macronuclear sequence assembly in a 24 hour mating library), we decided not to pursue an analysis of the remaining small RNAs that matched the mitochondrial genome.

Analysis of size classes of small RNAs

For many small RNAs studied to date, different functional classes are characterized by their length. In order to look at this further, we created histograms of size distributions for RNAs that survived the non-coding RNA filter step (Figure 4). We generated length histograms from distinct sequence reads in each library (only one read was plotted if that exact sequence and length occurred for multiple reads in the library). We further analyzed the size distribution for the ability of reads to be mapped to the macronuclear sequence assembly. Several conclusions can be drawn from viewing these histograms. The 20 nt, 21 nt and 22 nt species predominate in vegetative cells while the 27 nt species is the predominant size in RNAs sequenced from mated cells. One interesting observation from looking at these histograms is the lack of any peak in any of the libraries for a 25 nt RNA species, which our ^{32}P-labeling results indicate are present in all samples, both vegetative and mating (Figure 1). The cDNAs in the library were prepared by a protocol that required a 3′OH group and a 5′PO_4 on the small RNAs for adapter ligation. The 25 nt species is not modified at its 3′ end (see lanes 2 vs. 3 and 10 vs. 11 in Figure 2). This RNA cannot be labeled by T4 polynucleotide kinase without first treating with calf intestine alkaline phosphatase (data not shown) suggesting that it has phosphate groups at its 5′ end. We suggest that the inability to recover these 25 nt RNAs in the library may be due to the presence of multiple phosphates at the 5′ end, which has been observed for *C. elegans* secondary silencing RNAs (Pak and Fire, 2007, Sijen et al. 2007).

Analysis of the histograms indicates distinct properties of the 20 nt and 21 nt classes vs. the 22 nt class of small RNA that predominate in the vegetative libraries. For distinct reads in the four vegetative libraries sequenced, 75.5±2.5% of the 20 mers and 70.0±2.2% of the 21 nt reads map to the macronuclear sequence assembly. Given that the macronuclear sequence assembly is incomplete, this indicates a strong likelihood that the 20 mers and 21 mers are macronuclear in origin. In contrast, only 46.6±5.1% of the distinct 22 nt reads from the vegetative libraries map to the macronuclear sequence assembly. This suggests the possibility that a substantial number of the 22 nt RNAs may arise from micronuclear sequences.

We performed an analysis to look at the 5′ nucleotide identity of these classes of small RNAs. Several classes of small RNAs in *C. elegans*, such as 21U and 26G, are distinguished by a distinct 5′ nucleotide and length (Adl and Berger, 2000). In an analysis of the 20, 21 and 22 nt RNAs from one of the vegetative libraries (Figures 5A, 5B and 5C), and the 27 nt RNAs from a 24 hour mating library (Figure 5D), we found that there is a strong bias towards U as the nucleotide at the 5′ end of each of these RNA classes. For the 20 nt species, 74.2% of the reads start with U, for the 21 nt species, 78.0% of the reads start with U, for the 22 nt species, 77.7% of the reads start with U and for the 27 nt species, 96.9% of the reads start with U. No other striking biases are observed at other positions, with the exception that over 50% of the 22 mers end in U. This preponderance of a 3′ terminal U is unique among the size classes. For the 27 mers, we asked whether the more abundant members of the group, comprising 17981 distinct sequences found between 10 and 100 times in the mat24_05 library, had a different nucleotide frequency at certain positions relative to the total pool of 2,919,225 distinct 27 mers in the same library (Figures 5D vs. 5E). With the exception of an increase in the frequency of U in the first nucleotide position from 96.9% in the total 27 mer pool to 99.4% in the abundant 27 mers, and a slight decrease in U frequency for nucleotides 2 through 9, there were no obvious sequence bias differences between these sets.

The 27 nt RNA species, whose production was induced by mating (Figure 1) is, as expected, the most prominent size class in all of the libraries made from mated cells. This 27 nt species can also be detected at a low level in ALXC2 vegetative cells but not in ALXC9. The ALXC2 strain undergoes a low level of self-mating under vegetative growth conditions while the ALXC9 strain does not. Therefore we conclude that production of the 27 nt RNA species is induced by mating. We find that between 63% and 67% of the distinct 27 nt reads in the mating libraries map to the macronuclear genome assembly. Given that this is an incomplete macronuclear genome assembly, and given that the macronucleus contains only 5% of the sequence complexity of the micronucleus, this implies that the majority of 27 nt RNAs are derived from macronuclear sequence. Given that the production of these RNAs peaks 24 hours after mixing of mating competent cells, which would correspond to an early polytene chromosome stage of macronuclear development in stichotrichs (Adl and Berger, 2000, Postberg et al., 2008), two hypotheses could explain their origin. These RNAs could originate from the parental macronucleus before it becomes degraded. Alternatively, these RNAs could be derived from the developing macronuclear anlage, having been transcribed specifically from the subset of DNA sequences that are destined to be retained in the new macronucleus.

27 nt RNAs originate from the parental macronucleus

In order to distinguish whether the 27 nt mating-specific RNAs originate from the parental macronucleus or the developing macronucleus, we aligned the 26–28 nt RNAs that survived the non-coding RNA filter against the micronuclear and macronuclear sequences for 6 different genes. We were limited in this analysis to six genes, because these are the only genes with complete micronuclear and macronuclear sequence pairs of the same gene alleles (there is very little micronuclear sequence available in public databases). We used the micronuclear and macronuclear sequences as targets for mapping and counted the number of sequence hits to each gene of either nuclear origin. Micronuclear and macronuclear gene sequences compared in this study were trimmed so that sequence length differences for each pair are only due to the presence of IESs and pointer sequence duplication in the micronuclear version. Table 2 shows the alignment statistics for all distinct 26–28 nt RNAs from the seven mating libraries. We treated this as a Venn diagram, determining the number of distinct 26–28 nt reads in the libraries that align to both the micronuclear and macronuclear versions, and the number that align to only one of the two versions. Alignment of a sequence only to the micronuclear version of a gene could be the result of alignment to an IES or an IES/MDS junction, which are unique to the micronucleus. Alignment to only the macronuclear sequence could be a result of an alignment that spans an MDS junction by extending past both sides of the "pointer" sequence. For all six genes tested, the majority of 26–28 nt reads align with both sequences. The number of reads that align with only the macronuclear version far exceeds the number of reads that align to only the micronuclear version for all six genes (see last four columns of Table 2). For example, the micronuclear version of the alpha telomere binding protein gene (TEBPAlpha) has 17 scrambled MDSs (Mitcham et al., 1992). 996 reads aligned to both the micronuclear and macronuclear versions of this gene. 252 reads aligned only to the macronuclear version of the gene while only 4 reads aligned only to the micronuclear version of the gene. Since for each MDS/MDS junction, the micronuclear version of the gene contains more unique sites for alignment than the macronuclear version (the micronuclear version has the entire IES plus the two MDS/IES junctions as unique sequence while the macronuclear version only has the MDS/MDS junction as unique sequence), the bias towards macronuclear-specific reads is even more striking. The macronuclear genome only contains 5% of the complexity of the micronuclear genome but its nanochromosomes are amplified relative to the micronuclear genome. This could account for the relative paucity of micronuclear-specific reads even if the micronuclear genome is also transcribed into 27 mers. We controlled for that possibility by only mapping distinct sequences, as opposed to total sequence reads, to avoid multiple counting of sequences that derive from amplified nanochromosomes in the macronucleus. Taken together, the data in Table 2 indicate that the majority of the 26–28 nt RNA species in mating cells originate from the parental macronucleus. The small number of micronuclear-specific reads for scrambled genes in Table 2 suggest that there may be a potential for some micronuclear production of 27 nt RNAs. Alternatively, since all of the micronuclear-matching sequences came from the three scrambled genes, there may be some partially unscrambled nanochromosomes in the parental macronucleus that generate these small number of micronuclear-specific 27 nt RNAs.

Visualization of small RNA coverage

In order to visualize the coverage of 26–28 nt RNAs on the macronuclear genome, we employed a minimal build of the UCSC Genome Browser (Kent et al., 2002) on an Ubuntu Linux computer. We made an *Oxytricha trifallax* genome consisting of two chromosomes, chr1 and chr2, which we call "oxytri1". chr1 corresponds to the concatenated complete nanochromosomes and chr2 corresponds to the concatenated partially-assembled nanochromosomes that were used in our Bowtie macronuclear alignments (Table 1, Figures 3 and and4).4). Using mapping of the reads in Bam format (Li et al., 2009), we are able to visualize alignments of 26–28 nt small RNA sequencing reads to their (possibly multiple) alignment sites on the assembled contigs.

Figure 6 shows a screen shot from the *Oxytricha trifallax* macronuclear genome browser with the alignment of 26–28 nt sequencing reads from 7 different mating libraries to the alpha telomere binding protein macronuclear gene. All the libraries have 27 nt sequences that originate from both strands of the nanochromosome. There appears to be a non-random distribution of reads, with pileups at specific places that are maintained in all the libraries and which are specific to each strand. In order to give a sense for coverage density relative to total reads in each library, the total number of 26–28 nt reads from each library that mapped to the macronuclear genome is indicated at the left of the figure.

To test whether there is evidence to suggest that any of the small RNA sequences that appear in multiple copies in the different libraries decrease or increase in number at different rates relative to the other sequences during macronuclear development, we analyzed the data using DESeq (Anders and Huber, 2010). For all distinct 27 mers whose sequence occurred at least 10 times in *any* of the seven libraries, we counted the number of occurrences of that sequence in each of the libraries. We compared the expression of these 27 mers between the early mating libraries (24–30hrs - mat24_05, mat24_06 and mat30_07) and the later mating libraries (55–72hrs - mat55_10 and mat72_11). DESeq determined an unadjusted p-value for the difference in relative expression of each of the 27 mers between these two sets of libraries. Figure 7 shows a histogram of these p-values. For 52.6% of the 27 mers examined, the unadjusted p-value for these two expression conditions is between 0.99 and 1.0, indicating that the relative expression of these 27 mers does not change during macronuclear development. A small fraction of abundant 27 mers did show significant changes (even after adjusting for multiple hypothesis testing) during macronuclear development, and these outliers will be interesting for further study. Based on our ^{32}P labeling of total RNA during a developmental time course (Figure 1), we know that the total amount of the 27 mer pool peaks 24 hours from the start of mating and decreases as macronuclear development proceeds. We hypothesize that for the majority of mating-specific 27 mers, their abundance decreases at a uniform rate as macronuclear development proceeds. This is in contrast to the scanRNAs of *Tetrahymena* which undergo a population filtering step during macronuclear development (Chalker and Yao, 2011, Nowacki et al., 2011).

From the small RNA sequence alignment to alpha telomere binding protein in Figure 6, it is clear that sequence reads are found that correspond to each of the 17 MDS sequences whose location is indicated at the top of the figure. As seen in Figure 8B, some of these reads cross the MDS/MDS junctions and include sequences on both sides of the pointers as predicted by the analysis of the data in Table 2. An interesting question is whether the 27 nt small RNAs that correspond to the plus strand of the open reading frame are produced from processed mRNA or perhaps the pre-mRNA or some other precursor specific to small RNA production in which the introns are not removed by splicing. Compared to other eukaryotes, introns are relatively rare in *Oxytricha trifallax* (Cavalcanti et al., 2004). One intron is present in the alpha telomere binding protein pre-mRNA (Figure 8C). For this intron and others not shown here, there is evidence of plus strand 27 nt RNAs derived from the intron, and from the intron/exon junctions. This indicates that a least a subset of the 27 nt RNAs from this nanochromosome are derived from RNAs in which introns are not removed.

The 27 nt RNAs arise from both strands and are not uniformly distributed

Figure 9 shows genome browser screen shots of 26–28 nt RNA coverage of 4 additional nanochromosomes. These nanochromosomes all have micronuclear gene sequence available, and the locations of the MDSs are indicated at the top of each panel. This figure supports the conclusion that the 27 nt RNA species is produced from both strands of the nanochromosomes and that they do not have a uniform distribution across those strands. In order to determine how well-correlated the production of 27 nt RNAs are from both strands of the macronuclear genome, we plotted the number of 26–28 nt reads derived from each strand for each nanochromosome that had 10 or more 26–28 nt RNAs align to it. In Figure 10A we plotted the total number of reads from the mat24_06 library that align to each strand of each nanochromosome (these read alignment counts include multiple identical reads of the same sequence - left side of the figure) and we plotted the number of distinct reads that align to each strand (only one occurrence of a sequence that occurs multiple times in a library was used - right side of the figure). Pearson correlations (R) of 0.91 and 0.94 are obtained for the two plots, indicating a strong correlation between the numbers of small RNAs matching the two different strands of each nanochromosome. The red lines in each plot represent predicted positions of two standard deviations from the mean; in a model where a read is equally likely to arise from either strand, 95% of points should fall between these lines. The blue dotted lines represent the measured region in which 95% of the points lie. As can be seen, plotting distinct read coverage allows for an almost perfect fit to the theoretical curve, while plotting the total reads for each strand allows for more variation than expected. This difference may be related to the fact that we generally do not see a uniform distribution of read density along the nanochromosomes, and we observe many piles of identical reads that align to either strand. To quantify this effect, we calculated the coefficient of variation (standard deviation divided by mean) of the counts of 26–28 nt reads across the positions of each complete nanochromosome in chr1. The distribution of this statistic over all the nanochromosomes has a peak at about 6.0 or greater (depending on the library). Figure 10B shows a graph of coefficient of variance for 26–28 nt RNAs in the mat24_06 library. The peak at 6.0 (standard deviation of the coverage of 26–28 nt RNAs at any position on a nanochromosome is 6 times greater than the mean coverage at any position on that nanochromosome), is consistent with a highly non-uniform distribution of reads across the nanochromosome. This effect is not due to insufficient coverage, as when we calculate the coefficient of variation measurement for the top 10% of nanochromosomes with the highest 26–28 nt RNA coverage, we still obtain a coefficient of variation for the mat24_06 library of 5.80 (data not shown). From the plots in Figure 10, we conclude that there is a strong correlation between the two strands of the nanochromosomes in terms of the number of 26–28 nt sequence reads derived from each strand and that the coverage of reads on each strand is highly non-uniform.

The 27 nt RNA class has lower coverage in the telomere-proximal region

In addition to the non-uniformity in read coverage across entire nanochromosomes, we have observed that small telomere-proximal regions of the nanochromosomes are strongly depleted for the 27 nt small RNAs, aligned to either strand, relative to the rest of the nanochromosome. To quantify this, we determined the average 26–28 nt coverage density of each of 10,002 complete nanochromosomes from chr1 that were at least 500 bp in length. For each of these nanochromosomes, we measured the density of 5′ ends of 26–28 nt small RNAs at each of the proximal 500 nt positions from the 5′ or 3′ end of the nanochromosome, for both the plus strand or minus strand-aligning reads. These distributions were plotted relative to a plot of uniform distribution based on the average coverage over the 500 positions at the appropriate end of the nanochromosome (Figure 11). The telomere repeat in the nanochromosome sequences in the chr1 collection is on average 20 nt, and is indicated by a dashed vertical line in the figure. As small RNA sequences made up of the telomeric repeat would have been filtered out as part of the initial non-coding RNA filter step, we expect to have no coverage in this region. It should be noted that only 237 reads out of 10,570,000 total sequences applied to the non-coding RNA filter in the mat24_06 library matched the telomeric repeat, so the non-coding RNA filter step cannot explain the paucity of coverage of small RNAs at the telomere and subtelomeric regions. Since the counts of the 5′ ends of the RNAs were plotted, a shaded region 27 nt long was added to the plot for the plus strand reads measured from the 3′ end of the nanochromosome to account for the 27 nt read length. When these data are plotted, it is clear that there is a zone of 30 nt proximal to the telomere that has over 8-fold lower density of 27 nt RNA coverage relative to a uniform distribution. We did identical plots of the minus strand-aligning 26–28 nt RNAs for mat24_06, and found that the graph obtained is a mirror-image of the one shown in Figure 11 (data not shown), again indicating that the 30 bp proximal to the telomere on the nanochromosome have over 8-fold lower small RNA coverage than the rest of the nanochromosome. The WGS2.1.1 full-length nanochromosome sequence collection has no information regarding which strand contains an open reading frame, but given our results here and the fact that both strands of nanochromosomes are transcribed to generate the 27 mers (Figure 10A), there seems to be no correlation between the coding strand and the lack of 27 nt RNA coverage near the telomeres.

Discussion

In ciliates, there is a clear role for the parental macronucleus to provide epigenetic information to control the complex DNA elimination and rearrangements involved in the development of the new macronucleus. This epigenetic information is carried in the form of RNA (Nowacki et al., 2011). We explored the production of small RNAs in *Oxytricha trifallax* during mating and subsequent macronuclear development. We have identified an RNA species of 27 nt in length whose production is induced upon mating. Next generation sequencing and analysis of these RNAs demonstrates that they are macronuclear in origin. We have decided to name these "27macRNAs", because of their length and their origin in the parental macronucleus.

The potential biogenesis of the 27macRNAs is interesting. They are transcribed from both strands of the nanochromosomes, indicating a requirement for an RNA polymerase activity that recognizes a common feature of both strands of the nanochromosomes. A good candidate for such a common feature is the telomeres. We had previously identified an RNA polymerizing activity from vegetative macronuclei of *Oxytricha nova* that could use telomeric extensions and their internal double-stranded sequences as templates (Zahler and Prescott, 1989). At the time, we proposed that this activity was responsible for creating the RNA primers for DNA replication at the telomeres (a DNA primase), but this activity also fits the requirements of an enzyme that can transcribe both strands of the nanochromosomes from the telomeres.

We propose that the 27macRNAs are processed from long double-stranded RNAs by a dicer-like activity, similar to the micronuclear-derived scanRNAs in Tetrahymena (Mochizuki and Gorovsky, 2005, Malone et al., 2005). Our observation of the presence of a 5′ monophosphate and 3′ OH group on the 27macRNAs is consistent with dicer products. It has been reported in *Oxytricha trifallax* that there is bidirectional transcription of long RNAs from the nanochromosomes 6 hours after mating strains are mixed together. These have been hypothesized to guide macronuclear rearrangements (Nowacki et al., 2008). The timing of production of these bi-directional long RNAs occurs prior to the appearance of 27macRNAs. We hypothesize that these may serve as precursors for 27macRNA production by a dicer family enzyme. In *Drosophila*, some classes of piRNAs are found matching both genomic strands and associate with distinct PIWI proteins for each strand (Siomi et al., 2011). The ping-pong model for production of these small RNAs from both strands holds that piRNAs processed from one strand direct cleavage of complementary piRNAs from the other strand between nucleotides 10 and 11 of the cleaving piRNA. This cleavage results in half of the piRNAs having an A at the tenth nt position, which is complementary to a U at the 5′ end of the partner piRNA that directed its cleavage (Brennecke et al., 2007). We see no evidence for an A bias in the 27 mers at position 10 in all distinct 27macRNAs (Figure 5D) nor even in abundant 27macRNAs (Figure 5E), nor any evidence of 10 base overlap of 27macRNAs on opposite strands, so it is unlikely that the ping-pong model can explain bidirectional production of piRNAs.

Our observation that the subtelomeric regions of the nanochromosomes have much lower coverage of 27macRNAs from either strand relative to the rest of the nanochromosome has several possible explanations. In one theory, RNA polymerase would generate a long RNA from each strand of a nanochromosome, with subsequent pairing of the complementary strands to form a long double-stranded RNA molecule. An inability of the RNA polymerase activity that makes these longer 27mac precursors to begin transcription at the very beginning of the nanochromosomes, or an inability of this enzyme to copy RNA to the very 3′ end of the nanochromosomes, would lead to single-stranded regions at the ends of the double-stranded RNA that result from pairing of the two long complementary RNA strands. These single-stranded regions would not be substrates for processing by the double-stranded endonuclease activity of dicer. Whatever the reason for low representation of 27macRNAs in subtelomeric regions, the lack of coverage of 27macRNAs in this region rules out the possibility that these RNAs play a direct role in guiding the precise *de novo* addition of telomeres to the nanochromosome ends in the developing macronucleus. However, their absence from the region may be important, allowing for the enzymes that cleave and add telomeres to function at the sites of precise telomere addition.

Epigenetic information from the parental macronucleus, sent through an RNA intermediate, has been hypothesized to control three different aspects of macronuclear development in stichotrichous ciliates. These are (1) the joining of MDSs and importantly the unscrambling of the micronuclear genome (Nowacki et al., 2008), (2) the control of macronuclear nanochromosome copy number (Heyse et al., 2010, Nowacki et al., 2010), and (3) the control of chromatin formation in the developing macronucleus which may play a role in guiding DNA elimination (Postberg et al., 2008). The 27macRNAs may be involved in several of these processes. Long RNAs made bi-directionally from the nanochromosomes in the parental macronucleus have been observed and have been demonstrated experimentally to function in guiding unscrambling of the macronuclear genome (Nowacki et al., 2008). These long RNAs may possibly serve as precursors to the 27macRNAs. While 27macRNAs have been observed that cross MDS/MDS junctions, it is unclear whether the population that span MDS junctions are present in sufficient quantities, and contain sufficient information, to serve as guides for MDS joining and unscrambling. We see a wide range in depth of coverage of the different nanochromosomes by the 27macRNAs (Figure 10A), and it is possible that the copy number of these nanochromosomes could be reflected in the number of 27macRNAs produced from each nanochromosome. If that is the case, then the 27macRNAs could be the RNA species that provide information to the developing macronucleus on the copy number of individual nanochromosomes in the parental macronucleus (Heyse et al., 2010, Nowacki et al., 2010).

Another potential function of 27macRNAs is to direct chromatin modifications in the developing macronucleus. A thorough analysis of chromatin modifications present in the developing macronucleus in the stichotrichous ciliate *Stylonichia* has been performed (Postberg et al., 2008). Postberg et al. found that during macronuclear development, sequences to be retained in the mature macronucleus are associated with permissive histone chromatin modifications, while sequences to be eliminated are associated with repressive histone modifications. They also identified a Piwi-family protein, Piwi/mdp1, whose accumulation in different nuclear types changes during the course of macronuclear development, and which is associated with chromatin in the macronuclear anlage (Postberg, et al., 2008). Piwi proteins bind small RNAs and can affect heterochromatin (Chalker and Yao, 2011, Lin and Yin, 2008), so a satisfying hypothesis is that 27macRNAs function to target the macronuclear-destined regions in the anlagen for protection from DNA elimination by altering their chromatin state. Given that nucleosome core protects ~146 bp of DNA, this potential role for 27macRNAs would exclude a role for them in guiding MDS joining as the IESs to be eliminated are smaller than the amount of DNA associated with the nucleosome.

Alternative models for 27macRNA function can be proposed in which this class of RNAs do an important job during macronuclear development but do not provide epigenetic information. They may function to change the chromatin state of the parental macronucleus in which they are made, helping to down-regulate gene expression from the parental macronucleus. They may act as siRNAs, targeting mRNAs from the parental macronucleus for destruction. They may function to promote DNA elimination during destruction of the parental macronucleus. Further studies to identify the protein binding partners of the 27macRNAs and to determine the subcellular localization of these RNAs during macronuclear development will be required to shed light on a functional role for the 27macRNAs.

Acknowledgements

We are indebted to John Paul Donohue for his help in establishing an *Oxytricha* minimal UCSC Genome Browser, without which much of this analysis would not have been possible. We would like to thank Robert Hammersmith for generously providing *Oxytricha trifallax* strains JRB310 and JRB510 and for advice on their maintenance. Many thanks to Jake Shandling and Cameron Ferguson for their help in growth and mating of *Oxytricha trifallax* cultures. Thanks to Shahid Khan and Nader Pourmand and the UCSC Genome Technology Center for reagents, help and advice in library preparation and for size selection of linkered cDNAs prior to sequencing. Thanks to Hiram Clawson for his helpful advice on the UCSC Genome Browser. Thanks to our UCSC RNA Center colleagues Manny Ares, Jeremy Sanford, Todd Lowe and Melissa Jurica for helpful discussions. This paper is dedicated to the memory of a great scientist and mentor, David M. Prescott, 1926–2011.

Figure 1: Mating in *Oxytricha trifallax* leads to production of a class of 27 nt RNAs. Total RNA was purified from vegetative *Oxytricha trifallax* (lanes 1–4) or at various time points after mixing together complementary mating strains (lanes 5–11). In addition, total RNA was purified from strain ALXC9 treated under identical conditions as a mating for 24 hrs (lane 12 - Mock 24hr). Total RNA was phosphatase treated followed by 5′ end labeling with ^{32}P, separated on a 15% polyacrylamide denaturing gel and visualized using a PhosphorImager. Sizes from Decade RNA 10 nt Ladder (Ambion) are indicated at left. Positions of small RNAs of interest are indicated at right. Lanes 1–11 directly correspond to RNA preparations used to prepare libraries for Illumina sequencing listed in the same order in Table 1.

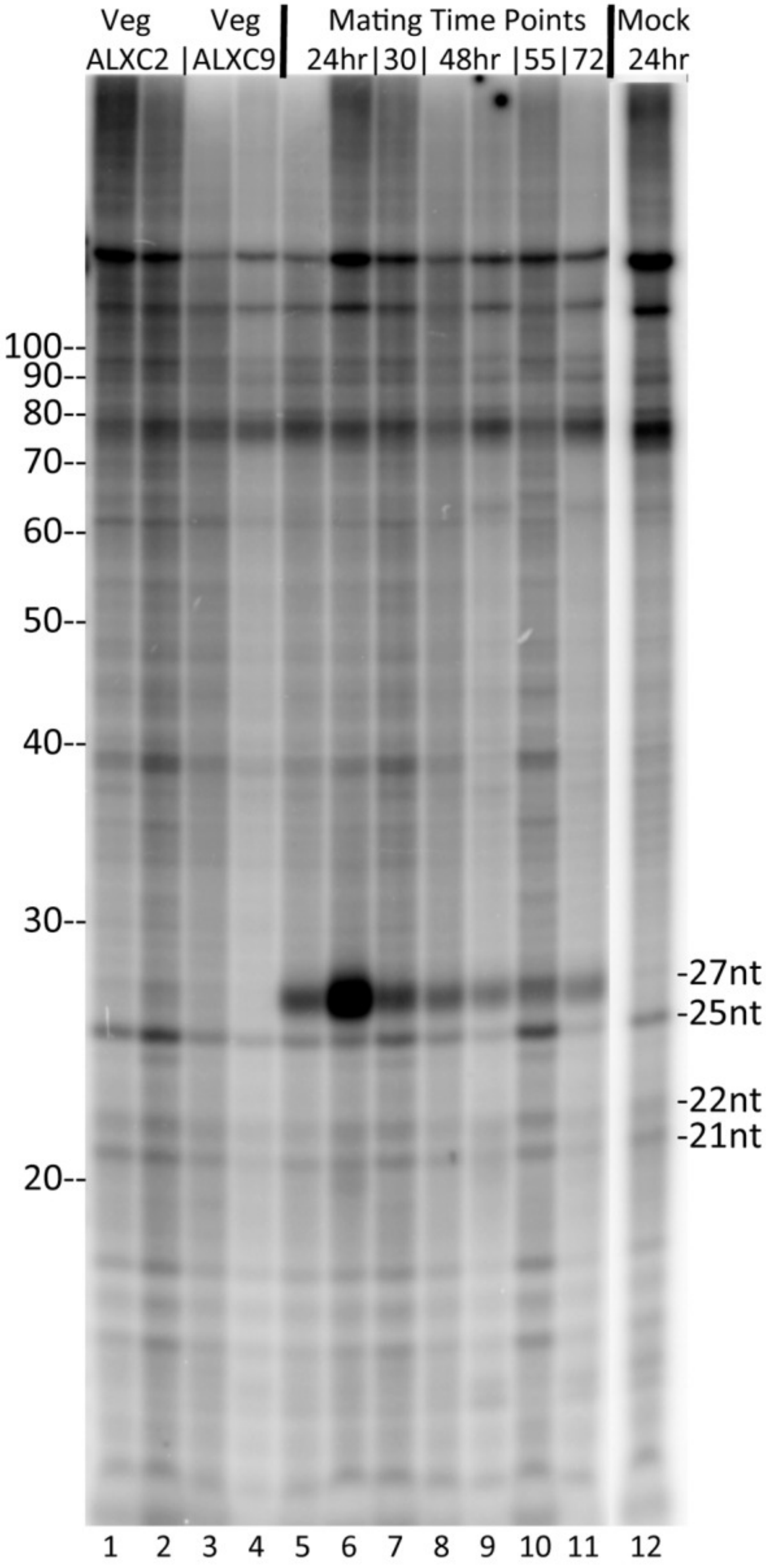

Veg
ALXC2
Veg
ALXC9
Mating Time Points
24hr
30
48hr
55
72
Mock
24hr
100--
90--
80--
70--
60--
50--
40--
30--
20--
-27nt
-25nt
-22nt
-21nt
1
2
3
4
5
6
7
8
9
10
11
12

Figure 2: The mating-specific 27 nt RNAs in *Oxytricha trifallax* are not modified at their 3′ ends. RNAs were tested with the beta elimination assay in order to determine if there is a modification at the 3′ end. RNAs tested are ^{32}P 5′ end labeled total *Oxytricha trifallax* RNA (lanes 2–11) or 5′ end labeled synthetic positive control RNAs containing the sequence of the *C. elegans lin-4* (21 nt) or *mir-90* (23 nt) microRNAs which were subsequently mixed with 1.5 µg of unlabeled *Oxytricha* total RNA prior to beta elimination (lanes 13–16). Control untreated samples (-) or samples subjected to the beta elimination reaction (+) are indicated. RNAs were isolated from vegetative ALXC2 (lanes 2 and 3) or from a mating between ALXC2 and ALXC9 that were harvested at the indicated timepoint after mixing (lanes 4–11). Beta elimination will remove the terminal ribose if both a free 2′ OH and 3′ OH are present, and will leave a 3′ cyclic phosphate. This loss of a nucleotide and the presence of the extra phosphate group will result in the treated RNAs running almost two bases faster than their untreated counterparts.

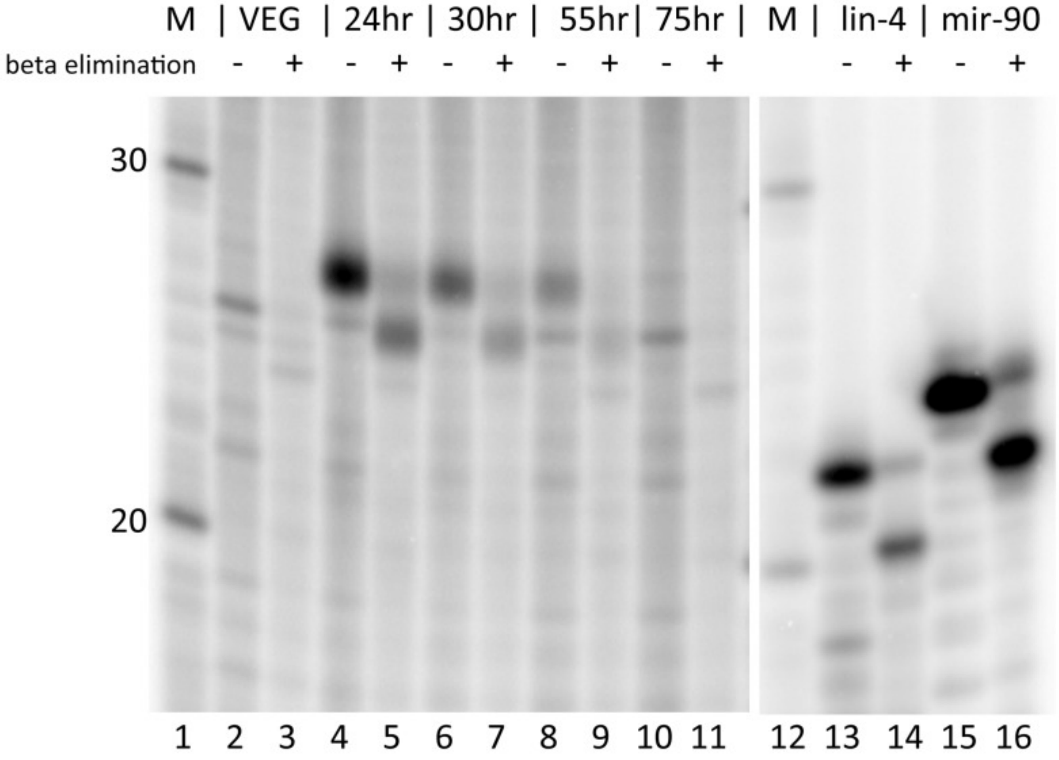

M | VEG | 24hr | 30hr | 55hr| 75hr | M | lin-4 | mir-90
beta elimination - + - + - + - + - + - + - +
30
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Figure 3: Flowchart of sequence analysis.

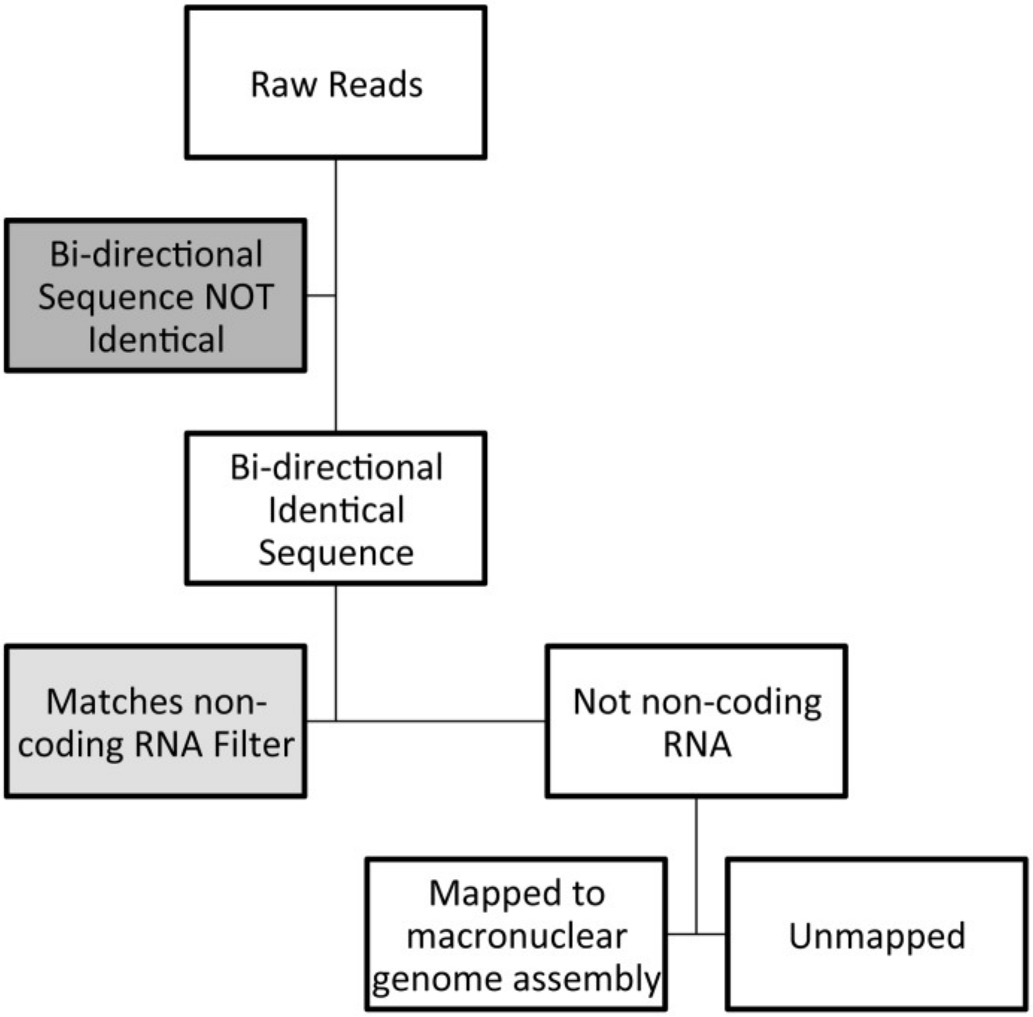

Raw Reads
Bi-directional Sequence NOT Identical
Bi-directional Identical Sequence
Matches non-coding RNA Filter
Not non-coding RNA
Mapped to macronuclear genome assembly
Unmapped

Figure 4: Vegetative small RNA libraries contain mostly 20–22 nt RNAs and mating libraries contain mostly 27 nt RNAs. The length distributions of sequencing reads in the small RNA range (18–30 nt) are plotted for five of the 11 libraries (comprising one replicate from each cell type or timepoint). The histograms are for distinct reads; only one occurrence of a sequence that appears more than once in a library is counted. The five representative libraries shown can be cross-referenced to table 1; VegALXC2 - veg02_01, VegALXC9 - veg09_04, Mat24hr - mat24_06, Mat48hr - mat48_08, Mat72hr - mat72_11. Black bars represent reads that mapped to the macronuclear sequence assembly. White portions of the bars represent reads that did not map to the macronuclear assembly. The vegetative strain VegALXC2 is capable of self-mating, leading to a minor peak at 27 nt.

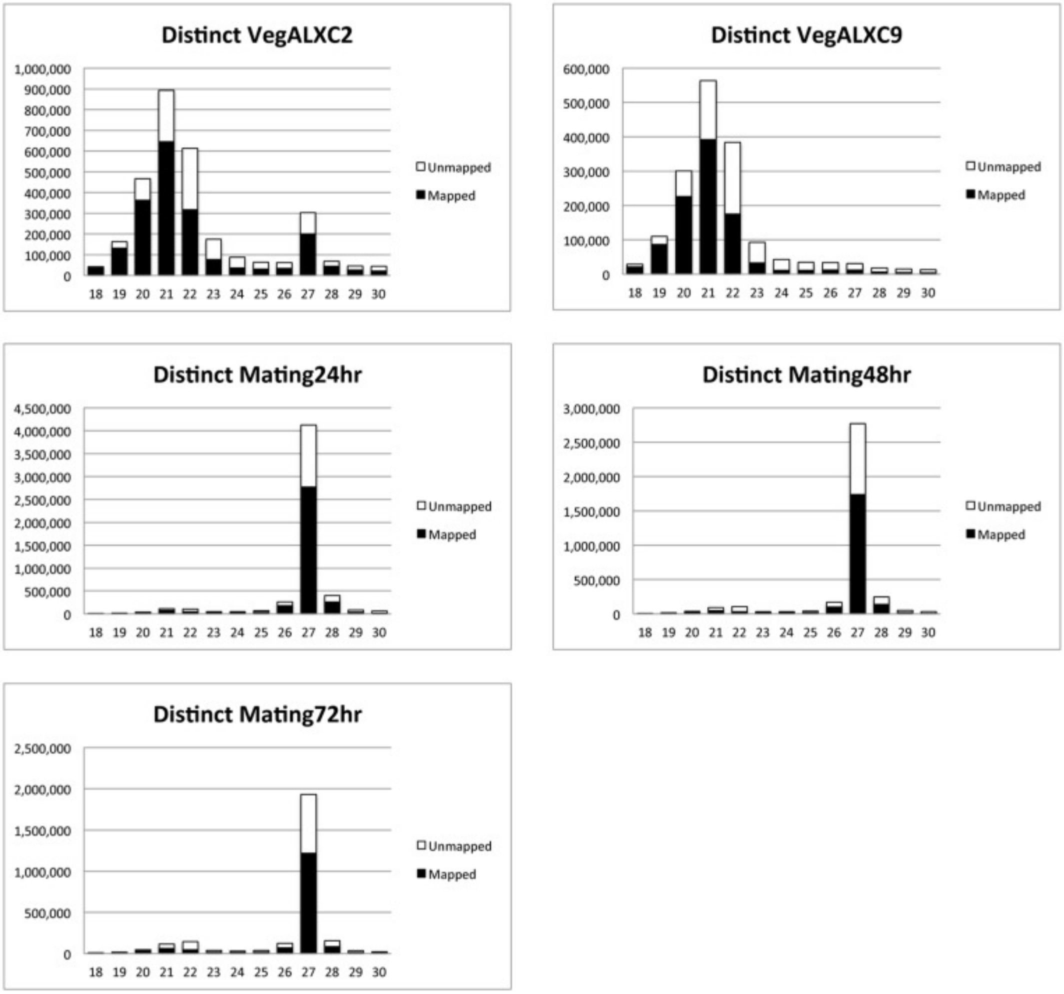


Figure 5: Nucleotide position bias in the different classes of sequenced small RNAs. These charts show the nucleotide frequency at each position for different small RNA size classes. The distinct sequences in the veg09_03 library and the mat24_05 library were filtered against non-coding RNAs. The distinct RNAs that made it through the filter were selected by size and the nucleotide composition of each position for the indicated size class was determined. **A.** Distinct 20 mers from the veg09_03 library (222,272 sequences). **B.** Distinct 21 mers from the veg09_03 library (416,2227 sequences). **C.** Distinct 22 mers from the veg09_03 library (327,422 sequences). **D.** Distinct 27 mers from the mat24_05 library (2,919,225 sequences). **E.** Abundant 27 mers from mat24_05 library comprising distinct sequences that were detected 10 or more times in the library (17,981 sequences).

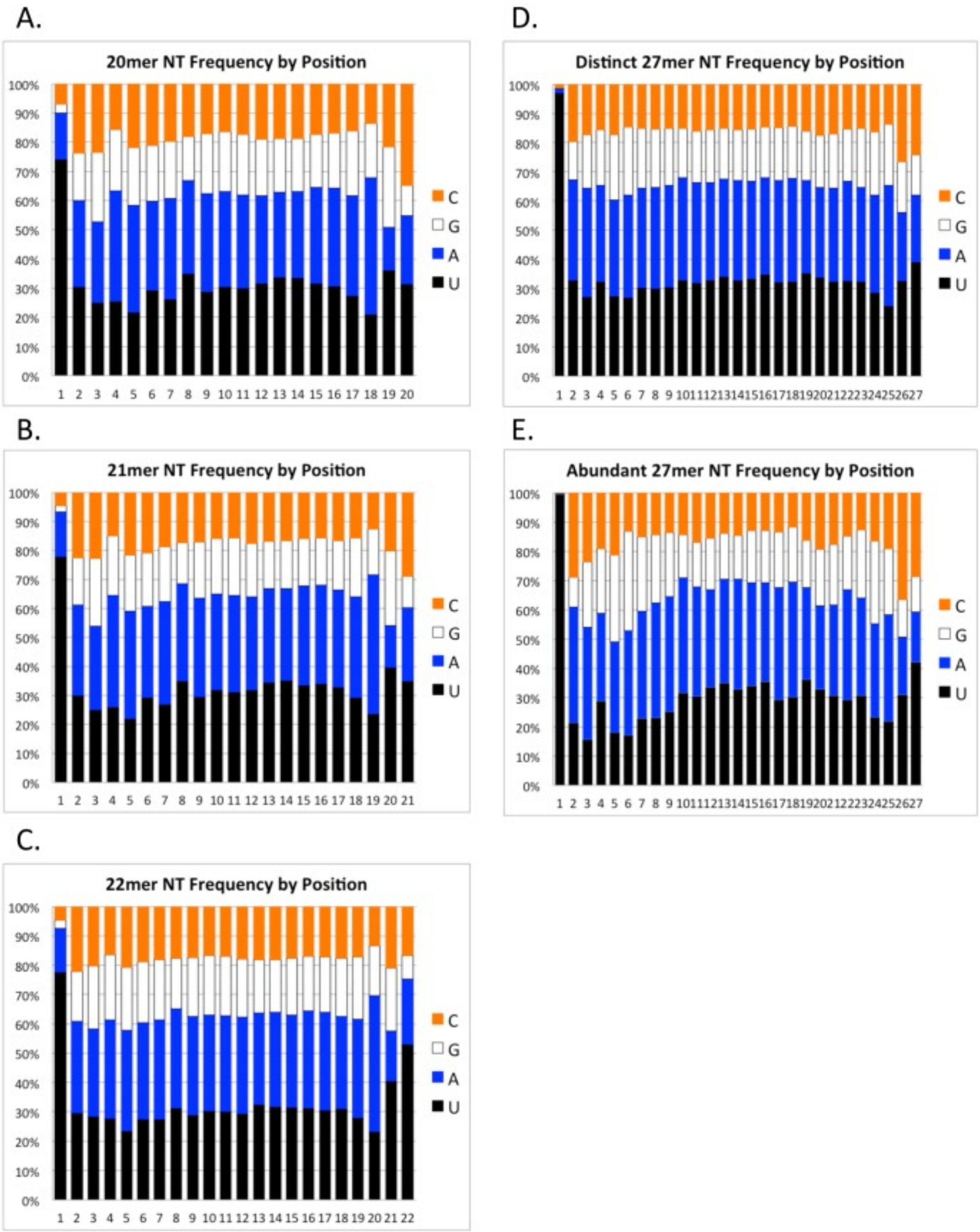

A.
20mer NT Frequency by Position
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
C
G
A
U
B.
21mer NT Frequency by Position
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
C.
22mer NT Frequency by Position
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
D.
Distinct 27mer NT Frequency by Position
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
E.
Abundant 27mer NT Frequency by Position
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27

Figure 6: 27 nt RNA reads from different mating libraries align to both strands of a nanochromosome. This screen shot from the *Oxytricha trifallax* macronuclear genome build of the UCSC Genome Browser shows the macronuclear nanochromosome corresponding to the alpha telomere binding protein gene indicated as a red bar (Nano42874.1) with the telomere sequences at the beginning and end of the nanochromosome noted as small black boxes in the track above that. The top track shows the locations of groups of MDS sequences or individual MDSs, as aligned using BLAT of the micronuclear sequence against the macronuclear genome. The alignments of 26–28 nt small RNAs from 7 different mating small RNA libraries (mat24_05, mat24_06, mat30_07, mat48_08, mat48_09, mat55_10 and mat72_11 - see Table 1) are shown with alignments to the plus strand of the nanochromosome indicated in blue and alignments to the minus strand of the nanochromosome indicated in orange. Numbers at the left indicate the total number of 26, 27 and 28 nt RNA sequences from each of the seven libraries that mapped to the macronuclear genome.

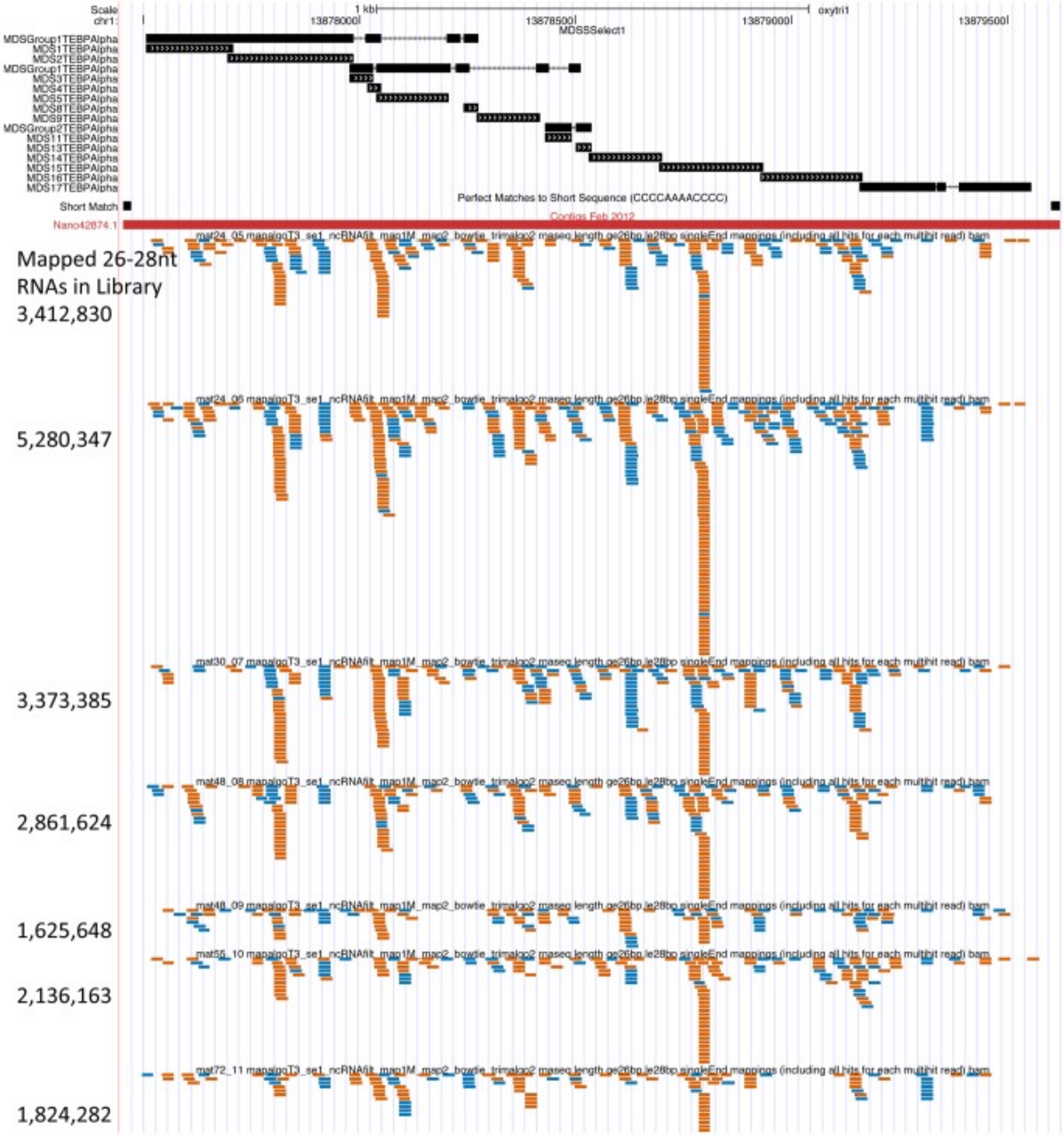

Scale
1 kb
oxytri1
chr1:
13878000
13878500
13879000
13879500
MDSSelect1
MDSGroup1TEBPAlpha
MDS1TEBPAlpha
MDS2TEBPAlpha
MDSGroup1TEBPAlpha
MDS3TEBPAlpha
MDS4TEBPAlpha
MDS5TEBPAlpha
MDS8TEBPAlpha
MDS9TEBPAlpha
MDSGroup2TEBPAlpha
MDS11TEBPAlpha
MDS13TEBPAlpha
MDS14TEBPAlpha
MDS15TEBPAlpha
MDS16TEBPAlpha
MDS17TEBPAlpha
Perfect Matches to Short Sequence (CCCCAAAACCCC)
Short Match
Contigs Feb 2012
Nano42874.1
mat24_05 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp le28bp singleEnd mappings (including all hits for each multihit read) bam
Mapped 26-28nt RNAs in Library
3,412,830
mat24_06 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp le28bp singleEnd mappings (including all hits for each multihit read) bam
5,280,347
mat30_07 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp le28bp singleEnd mappings (including all hits for each multihit read) bam
3,373,385
mat48_08 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp le28bp singleEnd mappings (including all hits for each multihit read) bam
2,861,624
mat48_09 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp le28bp singleEnd mappings (including all hits for each multihit read) bam
1,625,648
mat55_10 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp le28bp singleEnd mappings (including all hits for each multihit read) bam
2,136,163
mat72_11 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp le28bp singleEnd mappings (including all hits for each multihit read) bam
1,824,282

Figure 7: Histogram of unadjusted p-values for changes in relative expression level for 27 mers between early and late mating libraries. DESeq (Anders and Huber, 2010) was used to compare relative expression levels of distinct 27 mers whose sequence was found 10 or more times in any of the mating libraries. The relative expression of each of 57,140 abundant 27 mers between early mating time points (mat24_05, mat24_06 and mat30_07) and later mating time points (mat55_10 and mat72_11) was compared. Unadjusted p-values for changes in expression for each of the 27 mers were plotted on the histogram.

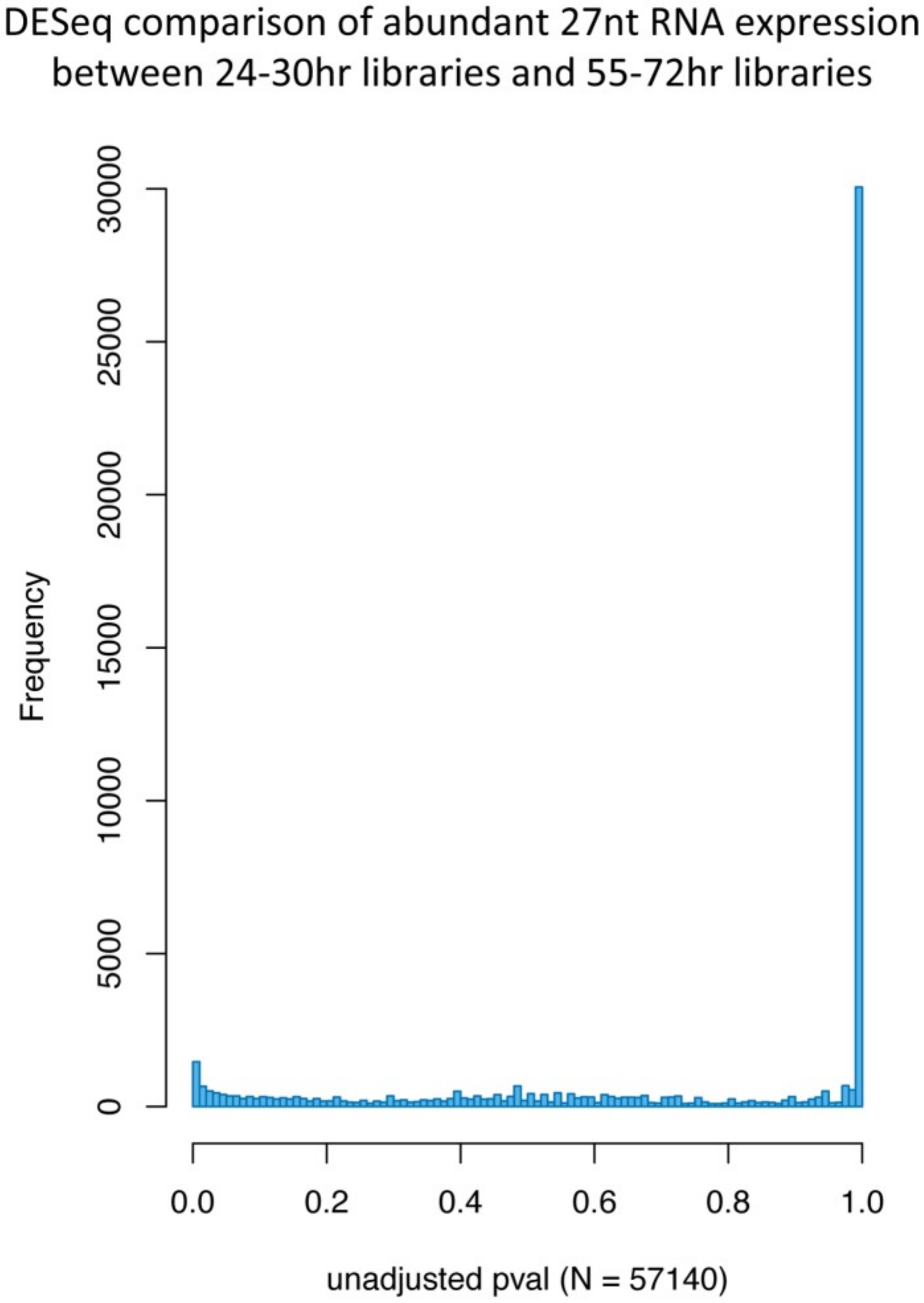

DESeq comparison of abundant 27nt RNA expression
between 24-30hr libraries and 55-72hr libraries
Frequency
0
5000
10000
15000
20000
25000
30000
0.0
0.2
0.4
0.6
0.8
1.0
unadjusted pval (N = 57140)

Figure 8: 27 nt small RNAs are macronuclear in origin and do not require intron removal for their generation. A. Schematic diagram of internally eliminated sequence (IES) removal and macronuclear destined sequence (MDS) joining during macronuclear development. Pointer sequences are direct repeats found in the micronucleus at MDS borders. Only one copy of the pointer is found in the macronucleus. **B.** *Oxytricha trifallax* Macronuclear Genome Browser screen shot of a region of the alpha telomere binding protein gene showing the junctions and overlapping pointer sequences of MDSs 9, 10 and 11. Only the mat24_06 library 26–28 nt small RNA track is shown. Note that the 26–28 nt small RNAs from the mat24_06 library from both strands overlap the MDS junction and pointer sequences. This is consistent with these small RNAs having originated from the mature parental macronucleus. **C.** Screen shot of a region of the alpha telomere binding protein gene containing its intron. 26–28 nt small RNAs from all seven mating libraries are shown. Note that sense strand small RNAs are found within the intron, indicating that they could be processed from introns. Also, small RNAs in the sense strand are found crossing the downstream intron/exon border, indicating that these are made from an RNA that did not undergo intron processing.

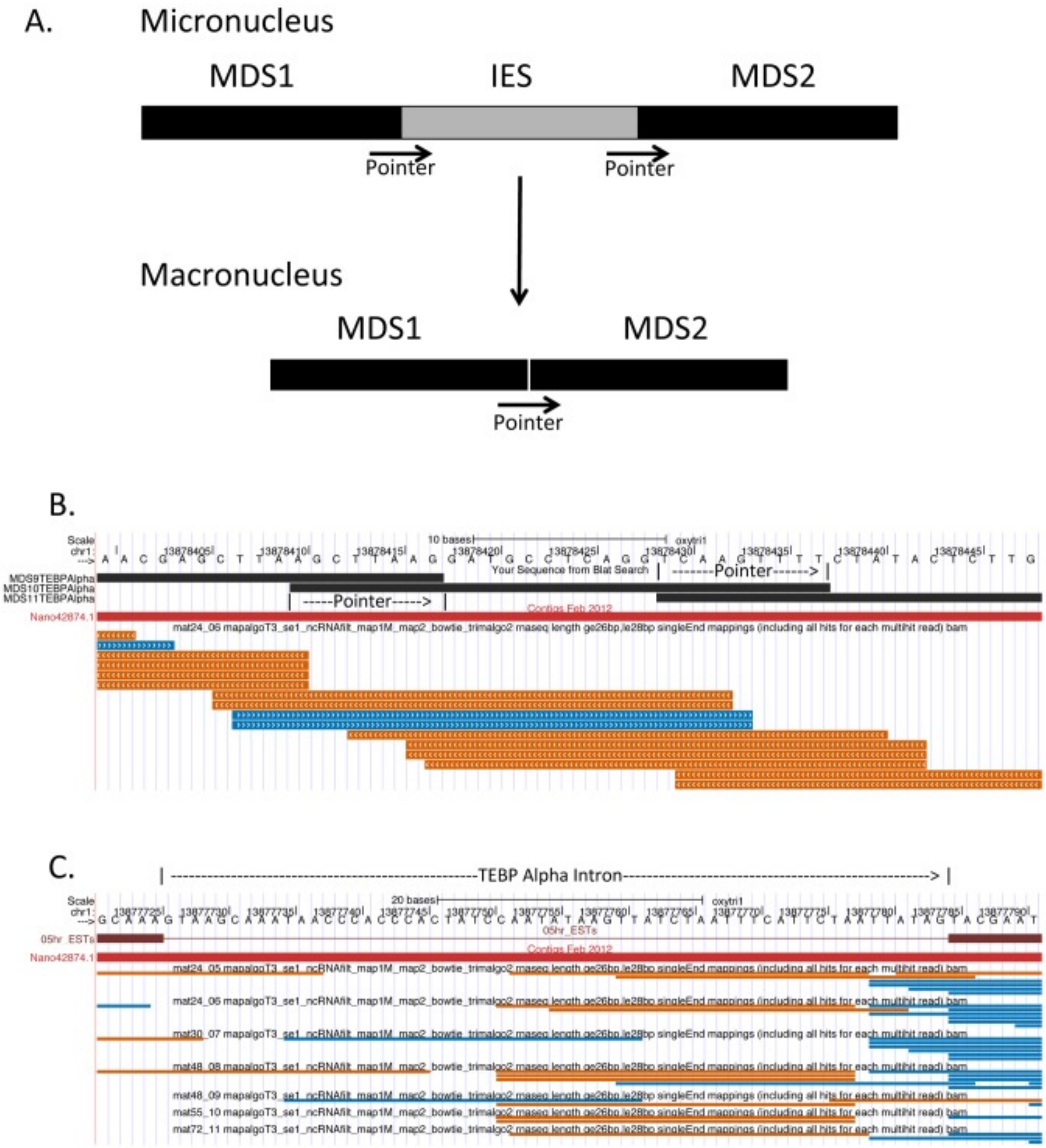

A.
Micronucleus
MDS1
IES
MDS2
Pointer
Pointer
Macronucleus
MDS1
MDS2
Pointer
B.
Scale
chr1:
10 bases
oxytri1
13878405
13878410
13878415
13878420
13878425
13878430
13878435
13878440
13878445
A A C G A G C T T A A G C T T A A G G A T G C C T C A G G T C A A G T T T T C T A T A C T C T T G
Your Sequence from Blat Search
--------Pointer------->
MDS9TEBPAlpha
MDS10TEBPAlpha
MDS11TEBPAlpha
-----Pointer----->
Contigs Feb 2012
Nano42874.1
mat24_06 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
C.
---TEBP Alpha Intron--->
Scale
chr1:
20 bases
oxytri1
13877725
13877730
13877735
13877740
13877745
13877750
13877755
13877760
13877765
13877770
13877775
13877780
13877785
13877790
G C A A A G T A A G C A A A T A A C C C A C C C C A C T A T C C A A T A T A A G T T A T C T A A T T T C A T T C T A A T T A T A G T A C G A A T
05hr_ESTs
05hr_ESTs
Contigs Feb 2012
Nano42874.1
mat24_05 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
mat24_06 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
mat30_07 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
mat48_08 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
mat48_09 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
mat55_10 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
mat72_11 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam

Figure 9: Screen shots of four different macronuclear genes from the *Oxytricha* macronuclear genome browser. In each panel, black boxes on top are a BLAT alignment of individual MDS sequences to show MDS junction location. Short match below that shows telomeric repeat locations. The red bar below that shows the nanochromosome contig extent. Blue and orange bars indicate the alignment location of 26–28 nt long small RNAs from the mat24_06 library. Plus strand alignments are in blue and minus strand alignments are in orange. **A.** DNA Polymerase Alpha. Note that this nanochromosome is incompletely assembled and is spread across three partially assembled contigs on chr2. **B.** Beta Telomere Binding Protein. **C.** Actin-I. **D.** CCCH Zinc Finger protein.

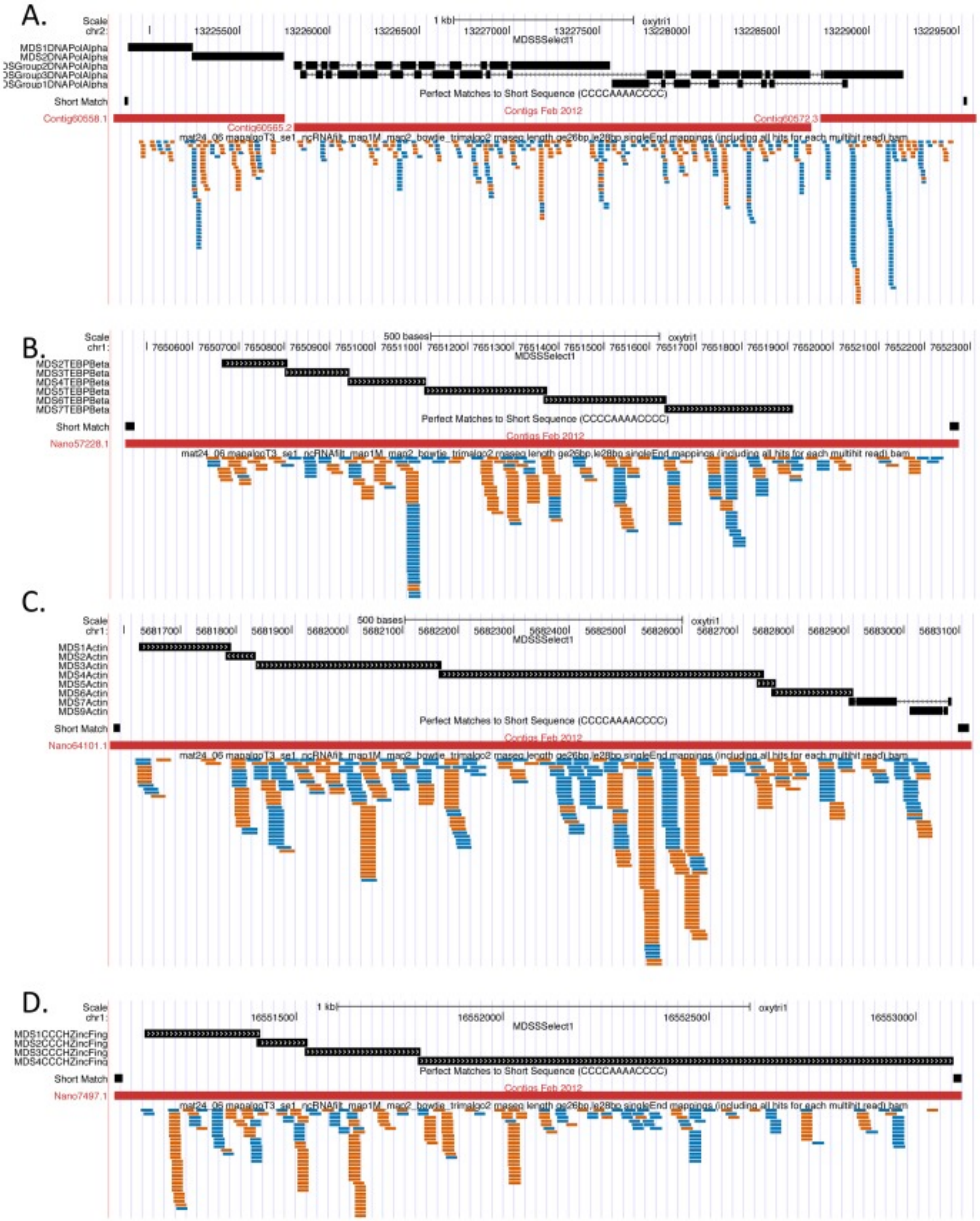
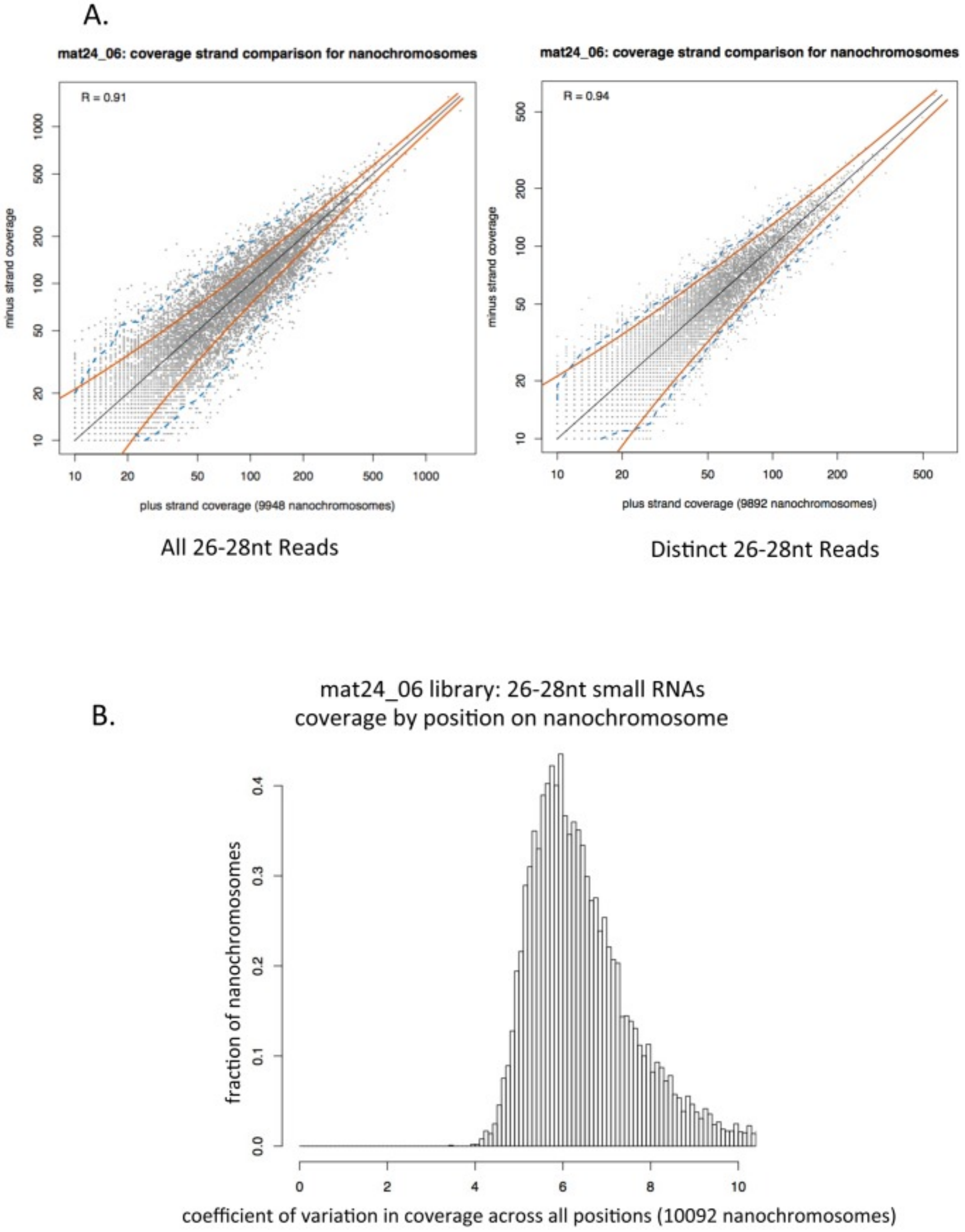

A.
Scale
chr2:
1 kb
oxytri1
13225500
13226000
13226500
13227000
13227500
13228000
13228500
13229000
13229500
MDSSelect1
MDS1DNAPolAlpha
MDS2DNAPolAlpha
)SGroup2DNAPolAlpha
)SGroup3DNAPolAlpha
)SGroup1DNAPolAlpha
Perfect Matches to Short Sequence (CCCCAAAACCCC)
Short Match
Contigs Feb 2012
Contig60558.1
Contig60565.2
Contig60572.3
mat24_06 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
B.
Scale
chr1:
500 bases
oxytri1
7650600
7650700
7650800
7650900
7651000
7651100
7651200
7651300
7651400
7651500
7651600
7651700
7651800
7651900
7652000
7652100
7652200
7652300
MDSSelect1
MDS2TEBPBeta
MDS3TEBPBeta
MDS4TEBPBeta
MDS5TEBPBeta
MDS6TEBPBeta
MDS7TEBPBeta
Perfect Matches to Short Sequence (CCCCAAAACCCC)
Short Match
Contigs Feb 2012
Nano57228.1
mat24_06 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
C.
Scale
chr1:
500 bases
oxytri1
5681700
5681800
5681900
5682000
5682100
5682200
5682300
5682400
5682500
5682600
5682700
5682800
5682900
5683000
5683100
MDSSelect1
MDS1Actin
MDS2Actin
MDS3Actin
MDS4Actin
MDS5Actin
MDS6Actin
MDS7Actin
MDS9Actin
Perfect Matches to Short Sequence (CCCCAAAACCCC)
Short Match
Contigs Feb 2012
Nano64101.1
mat24_06 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam
D.
Scale
chr1:
1 kb
oxytri1
16551500
16552000
16552500
16553000
MDSSelect1
MDS1CCCHZincFing
MDS2CCCHZincFing
MDS3CCCHZincFing
MDS4CCCHZincFing
Perfect Matches to Short Sequence (CCCCAAAACCCC)
Short Match
Contigs Feb 2012
Nano7497.1
mat24_06 mapalgoT3_se1_ncRNAfilt_map1M_map2_bowtie_trimalgo2 maseq length ge26bp,le28bp singleEnd mappings (including all hits for each multihit read) bam

Figure 10: Analysis of the distribution of 27 nt RNAs on macronuclear nanochromosomes. A. 27 nt RNA small RNAs are produced in equal numbers from both strands of the nanochromosome. For each full length nanochromosome with at least ten 26–28 nt small RNAs aligning to each strand, the total number of 26–28 nt small RNAs that map to each strand was plotted (left graph) or the number of distinct 26–28 nt small RNAs that map to each strand was plotted (right graph). Pearson R correlation values of 0.91 for all reads and 0.94 for distinct reads were obtained. This indicates a strong correlation of small RNA production from one strand of the nanochromosome with production of small RNAs from the other strand of the nanochromosome. Red curved lines represent 2 standard deviations from the mean; 95% of points would be expected to fall within these regions if there is a one-to-one correlation between the number of 26–28 nt small RNAs aligning to each strand of a nanochromosome (thin black line along the main diagonal). The blue dotted line indicates the actual lines within which 95% of the data points fall. **B.** There is a non-uniform distribution in the positioning of small RNAs on the nanochromosomes. For each position in each complete nanochromosome, the number of 26–28 nt small RNAs that start at that position in the mat24_06 sequencing library were determined. Then we determined the mean and standard deviation of coverage density on the nanochromosome. The coefficient of variation (standard deviation of coverage divided by the mean coverage) is plotted for each nanochromosome in the histogram. The peak of coefficient of variation at ~6.0 indicates that the standard deviation is 6.0 times greater than the mean. This is highly indicative of a non-uniform distribution of small RNA coverage on the nanochromosomes.



All 26-28nt Reads

Distinct 26-28nt Reads

mat24_06 library: 26-28nt small RNAs
coverage by position on nanochromosome

Figure 11: The first 30 nt of the nanochromosomes proximal to the telomeres have >8-fold lower coverage of 26–28 nt small RNAs relative to the rest of the nanochromosome. Graphic at the top shows a nanochromosome (rectangle) with 20 nt telomere sequence in gray shadow. 27 nt small RNAs for the plus strand of the nanochromosome are indicated as arrows above the nanochromosome. The average density of the location of 5′ ends of plus strand-aligning 26–28 nt RNAs was determined over 500 positions (in 20 bins of 25 bp each, shown by tick marks) relative to the 5′ (right side, positive values) and 3′ (left side, negative values) ends of 10,002 complete nanochromosomes. These densities were plotted relative to a uniform distribution of the aligned 26–28 nt RNAs over the same sets of 500 positions on the same nanochromosomes. A dotted line at 20 nt from either end is included on both sides of the zero point to indicate the average telomere length of 20 nt on the nanochromosome sequences to which these data were plotted. The shaded gray area, 27 nt wide, is included on the plot from the nanochromosome 3′ end because the plus strand-aligning reads would have 27 nt of sequence between their plotted 5′ end and the 3′ end of the nanochromosome. When minus strand-aligning reads were analyzed by this same method, a mirror image plot was obtained that is otherwise identical to the graph for plus end reads (not shown).

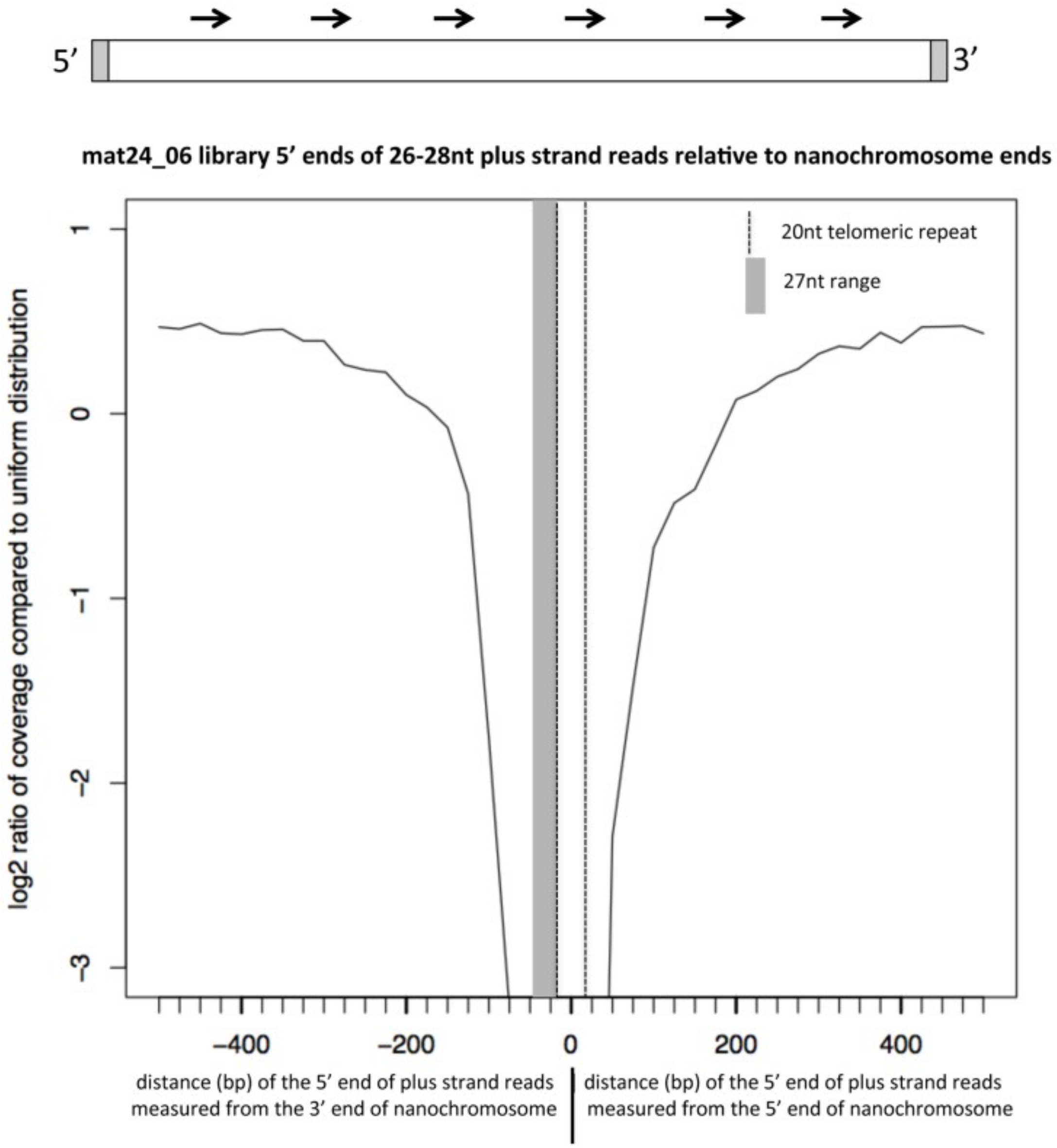

5′
3′
mat24_06 library 5′ ends of 26-28nt plus strand reads relative to nanochromosome ends
log2 ratio of coverage compared to uniform distribution
1
0
−1
−2
−3
20nt telomeric repeat
27nt range
−400
−200
0
200
400
distance (bp) of the 5′ end of plus strand reads measured from the 3′ end of nanochromosome
distance (bp) of the 5′ end of plus strand reads measured from the 5′ end of nanochromosome

Table 1: *Oxytricha trifallax* small RNA sequencing mapping statistics. Small RNA libraries were prepared from 11 different biological samples indicated under "Library name". veg02 is vegetative ALXC2 and veg09 is from vegetative ALXC9. mat## libraries were made from matings between ALXC2 with ALXC9, and the time in hours after mixing of the strains that the RNA for the library was isolated is indicated (mat24 RNA was extracted 24 hours after mixing mating strains). "Raw Reads" are the number of bidirectionally sequenced reads for each library (M=million). "Bidirectional Identical Reads" are the number of reads that were identical in both directions. These were used as high confidence sequences for further mapping. The bidirectional identical reads were filtered through a collection of *Oxytricha trifallax* non-coding RNAs (ncRNAs). "Matches to ncRNA Filter" are the number of sequences that match ncRNAs, and ncRNA% are the percentage of bidirectional identical reads that match the ncRNA filter. "Not ncRNA" are reads that did not match the ncRNA filter, and "Not ncRNA%" are the percentage of Bidirectional Identical Reads that are not a match to ncRNA. Bowtie was used to align the Not ncRNA reads to the macronuclear genome chr1 and chr2, as well as the 70 kb mitochondrial genome (chrM). "No macronuclear match" is the number of reads that did not match to the macronuclear sequences. "Mapped to Macronucleus" indicates the number of reads that mapped to the macronuclear sequences, and "Mapped to Macronucleus %" shows the fraction of "not ncRNA" reads that are macronuclear in origin. ">=10 Map Sites" are the number of reads that map to 10 or more places in the macronuclear sequence assembly, "<10 Multi Map Sites" are the number of reads that map to between 2 and 9 places in the macronuclear sequence assembly, and "Unique Map Sites" are the number of reads that map to only one position in the macronuclear sequence assembly.

Library Name	Raw Reads	Bi- directional Identical Reads	Matches to ncRNA Filter	ncRNA%	Not ncRNA	Not ncRNA %	No macro-nuclear match	Mapped to Macro-nucleus	Mapped to Macro-nucleus%	>=10 Map Sites	<10 Multi Map Sites	Unique Map Sites
veg02_01	12.64M	9.68M	3.62M	37.50%	6.05M	62.50%	2.68M	3.38M	55.80%	0.17M	1.55M	1.66M
veg02_02	11.02M	8.37M	1.59M	18.90%	6.79M	81.10%	3.59M	3.19M	47.10%	0.05M	1.40M	1.75M
veg09_03	8.71M	6.42M	0.84M	13.00%	5.58M	87.00%	4.30M	1.28M	23.00%	0.02M	0.58M	0.68M
veg09_04	8.28M	6.28M	1.64M	26.10%	4.64M	73.90%	3.05M	1.59M	34.20%	0.04M	0.73M	0.83M
mat24_05	9.16M	7.34M	1.13M	15.40%	6.21M	84.60%	2.25M	3.96M	63.70%	0.13M	1.42M	2.41M
mat24_06	13.20M	10.57M	1.55M	14.60%	9.02M	85.40%	3.11M	5.91M	65.50%	0.16M	2.11M	3.64M
mat30_07	12.39M	9.66M	2.90M	30.10%	6.75M	69.90%	2.53M	4.22M	62.50%	0.37M	1.46M	2.39M
mat48_08	9.59M	7.56M	1.90M	25.20%	5.66M	74.80%	2.31M	3.35M	59.10%	0.16M	1.18M	2.00M
mat48_09	7.32M	5.67M	1.58M	27.90%	4.09M	72.10%	1.77M	2.32M	56.60%	0.11M	0.88M	1.33M
mat55_10	10.74M	8.26M	3.59M	43.40%	4.67M	56.60%	1.86M	2.80M	60.10%	0.28M	0.95M	1.57M
mat72_11	8.43M	6.63M	2.49M	37.60%	4.14M	62.40%	1.79M	2.34M	56.80%	0.15M	0.82M	1.37M

Table 2: Venn Diagram analysis of 26–28 nt small RNA alignment to macronuclear/micronuclear sequence pairs - evidence for a macronuclear origin for the mating-specific small RNAs. Macronuclear sequences and the micronuclear sequences from which they were derived were used as targets for Bowtie mapping. Micro and macronuclear sequence pairs were trimmed so that all macronuclear sequence was contained within the micronuclear clone and regions of the micronuclear clone outside of the macronuclear gene were trimmed. The result of this is that sequence length differences between the micronuclear and macronuclear versions of the gene are due to IESs and repeated pointers in the micronuclear sequence. The gene name, length of the micronuclear and macronuclear sequences used in the filter, the number of MDSs, and whether the order of the MDSs is scrambled in the micronucleus are indicated. The number of distinct 26–28 nt small RNA sequences from each of the mating-derived small RNA libraries that aligned to each sequence in the pair were determined and then added together ("MicroCount" and "MacroCount"). "UnionCount" is the sum of the distinct sequences in each library found in the union of MicroCount and MacroCount. "XsectCount" is the number of distinct 26–28 nt sequences that aligned to both the macronuclear and micronuclear targets for that gene, and "XsectPct" is the percentage of UnionCount found in XsectCount. "OnlyMicroCount" is the number of distinct sequences in the seven libraries that only align to the micronuclear gene sequence and "OnlyMicroPct" is the percentage of sequences in UnionCount that are found in OnlyMicroCount. "OnlyMacroCount" is the number of distinct sequences in the seven libraries that only align to the macronuclear gene sequence and "OnlyMacroPct" is the percentage of sequences in UnionCount that are found in OnlyMacroCount.

Gene	Micro Length	Macro Length	MDSs	Scrambled?	Micro Count	Macro Count	Union Count	Xsect Count	Xsect Pct	Only Micro Count	Only Micro Pct	Only Macro Count	Only Macro Pct
Actin	2115	1503	10	Yes	718	1056	1078	696	64.56%	22	2.04%	360	33.40%
Zinc Finger	2128	1994	4	No	641	681	681	641	94.13%	0	0.00%	40	5.87%
DNAPolAlpha	7165	4645	47	Yes	1326	1771	1826	1271	69.61%	55	3.01%	500	27.38%
L29Cyclo	1811	1596	3	No	500	503	503	500	99.40%	0	0.00%	3	0.60%
TEBPAlpha	2787	2127	17	Yes	744	992	996	740	74.30%	4	0.40%	252	25.30%
TEBPBeta	1753	1250	7	No	612	651	651	612	94.01%	0	0.00%	39	5.99%

References

Adl SM, Berger JD (2000) Timing of life cycle morphogenesis in synchronous samples of *Sterkiella histriomuscorum*. II. The sexual pathway. *J Eukaryot Microbiol* 47: 443–449.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11: R106.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, et al. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128: 1089–1103.

Cavalcanti AR, Stover NA, Orecchia L, Doak TG, Landweber LF (2004) Coding properties of *Oxytricha trifallax* (*Sterkiella histriomuscorum*) macronuclear chromosomes: analysis of a pilot genome project. *Chromosoma* 113: 69–76.

Chalker DL, Yao MC (2001) Nongenic, bidirectional transcription precedes and may promote developmental DNA deletion in *Tetrahymena thermophila*. *Genes Dev* 15: 1287–1298.

Chalker DL, Yao MC (2011) DNA elimination in ciliates: transposon domestication and genome surveillance. *Annu Rev Genet* 45: 227–246.

Chang WJ, Stover NA, Addis VM, Landweber LF (2004) A micronuclear locus containing three protein-coding genes remains linked during macronuclear development in the spirotrichous ciliate *Holosticha*. *Protist* 155: 245–255.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 39: D876–882.

Greslin AF, Prescott DM, Oka Y, Loukin SH, Chappell JC (1989) Reordering of nine exons is necessary to form a functional actin gene in *Oxytricha nova*. *Proc Natl Acad Sci U S A* 86: 6264–6268.

Heyse G, Jonsson F, Chang WJ, Lipps HJ (2010) RNA-dependent control of gene amplification. *Proc Natl Acad Sci U S A* 107: 22134–22139.

Hoffman DC, Prescott DM (1996) The germline gene encoding DNA polymerase alpha in the hypotrichous ciliate *Oxytricha nova* is extremely scrambled. *Nucleic Acids Res* 24: 3337–3340.

Horwich MD, Li C, Matranga C, Vagin V, Farley G, et al. (2007) The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol* 17: 1265–1272.

Jahn CL, Klobutcher LA (2002) Genome remodeling in ciliated protozoa. *Annu Rev Microbiol* 56: 489–520.

Jung S, Swart EC, Minx PJ, Magrini V, Mardis ER, et al. (2011) Exploiting *Oxytricha trifallax* nanochromosomes to screen for non-coding RNA genes. *Nucleic Acids Res* 39: 7529–7547.

Kataoka K, Mochizuki K (2011) Programmed DNA elimination in *Tetrahymena*: a small RNA-mediated genome surveillance mechanism. *Adv Exp Med Biol* 722: 156–173.

Kent WJ (2002) BLAT–the BLAST-like alignment tool. *Genome Res* 12: 656–664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. *Genome Res* 12: 996–1006.

Kurth HM, Mochizuki K (2009) 2′-O-methylation stabilizes Piwi-associated small RNAs and ensures DNA elimination in *Tetrahymena. RNA* 15: 675–685.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25.

Lauth MR, Spear BB, Heumann J, Prescott DM (1976) DNA of ciliated protozoa: DNA sequence diminution during macronuclear development of *Oxytricha. Cell* 7: 67–74.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Lin H, Yin H (2008) A novel epigenetic mechanism in *Drosophila* somatic cells mediated by Piwi and piRNAs. *Cold Spring Harb Symp Quant Biol* 73: 273–281.

Liu Y, Taverna SD, Muratore TL, Shabanowitz J, Hunt DF, et al. (2007) RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in *Tetrahymena. Genes Dev* 21: 1530–1545.

Malone CD, Anderson AM, Motl JA, Rexer CH, Chalker DL (2005) Germ line transcripts are processed by a Dicer-like protein that is essential for developmentally programmed genome rearrangements of *Tetrahymena thermophila. Mol Cell Biol* 25: 9151–9164.

Mitcham JL, Lynn AJ, Prescott DM (1992) Analysis of a scrambled gene: the gene encoding alpha-telomere-binding protein in *Oxytricha nova. Genes Dev* 6: 788–800.

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) Conjugation-specific small RNAs in *Tetrahymena* have predicted properties of scan (scn) RNAs involved in genome rearrangement. *Genes Dev* 18: 2068–2073.

Mochizuki K, Gorovsky MA (2005) A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev* 19: 77–89.

Nowacki M, Haye JE, Fang W, Vijayan V, Landweber LF (2010) RNA-mediated epigenetic regulation of DNA copy number. *Proc Natl Acad Sci U S A* 107: 22140–22144.

Nowacki M, Landweber LF (2009) Epigenetic inheritance in ciliates. *Curr Opin Microbiol* 12: 638–643.

Nowacki M, Shetty K, Landweber LF (2011) RNA-Mediated Epigenetic Programming of Genome Rearrangements. *Annu Rev Genomics Hum Genet* 12: 367–38.

Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, et al. (2008) RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature* 451: 153–158.

Pak J, Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in *C. elegans. Science* 315: 241–244.

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: 3.

Prescott DM (2000) Genome gymnastics: unique modes of DNA evolution and processing in ciliates. *Nat Rev Genet* 1: 191–198.

Prescott DM, DuBois ML (1996) Internal eliminated segments (IESs) of *Oxytrichidae. J Eukaryot Microbiol* 43: 432–441.

Prescott DM, Prescott JD, Prescott RM (2002) Coding properties of macronuclear

DNA molecules in *Sterkiella nova (Oxytricha nova). Protist* 153: 71–77.

Sijen T, Steiner FA, Thijssen KL, Plasterk RH (2007) Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science* 315: 244–247.

Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 12: 246–258.

Swanton MT, Heumann JM, Prescott DM (1980) Gene-sized DNA molecules of the macronuclei in three species of hypotrichs: size distributions and absence of nicks. DNA of ciliated protozoa. VIII. *Chromosoma* 77: 217–227.

Swart EC, Nowacki M, Shum J, Stiles H, Higgins BP, et al. (2012) The *Oxytricha trifallax* mitochondrial genome. *Genome Biol Evol* 4: 136–154.

Taverna SD, Coyne RS, Allis CD (2002) Methylation of histone h3 at lysine 9 targets programmed DNA elimination in *Tetrahymena. Cell* 110: 701–711.

Williams K, Doak TG, Herrick G (1993) Developmental precise excision *of Oxytricha trifallax* telomere-bearing elements and formation of circles closed by a copy of the flanking target duplication. *EMBO J* 12: 4593–4601.

Yang Z, Vilkaitis G, Yu B, Klimasauskas S (2007) Chen× (2007) Approaches for studying microRNA and small interfering RNA methylation in vitro and in vivo. *Methods Enzymol* 427: 139–154.

Yao MC, Chao JL (2005) RNA-guided DNA deletion in *Tetrahymena*: an RNAi-based mechanism for programmed genome rearrangements. *Annu Rev Genet* 39: 537–559.

Zahler AM, Prescott DM (1989) DNA primase and the replication of the telomeres in *Oxytricha nova. Nucleic Acids Res* 17: 6299–6317.

Zoller SD, Hammersmith RL, Swart EC, Higgins BP, Doak TG, et al.. (2012) Characterization and Taxonomic Validity of the Ciliate *Oxytricha trifallax* (Class Spirotrichea) Based on Multiple Gene Sequences: Limitations in Identifying Genera Solely by Morphology. *Protist*.

CHAPTER 3

mRNA Expression Profiles During Macronuclear Development in the Ciliate *Oxytricha trifallax*

Introduction

Specialization of the germline and soma is a hallmark of multicellular organisms. Germline cells transmit genetic and epigenetic information to offspring, while somatic cells carry out day to day functions of the organism. While details vary, germline cells have a distinct genome organization from somatic cells and are capable of self-renewal and also differentiation, which includes meiosis and cellular and nuclear reorganization, for sexual propagation. In some lineages somatic cells influence epigenetic modifications in the germline, which can be passed on to subsequent generations. The mechanisms by which germline genomes remain quiescent and are then restructured during differentiation as well as the degree to which and mechanisms by which epigenetic information is transmitted from soma to germline to offspring are still largely uncharted, in part because in most species germline cells are rare and difficult to propagate and differentiate *ex vivo*.

Ciliates are diverse, abundant, extremely successful unicellular eukaryotes that display a special case of germline-soma specialization vis-à-vis nuclear dimorphism: a germline nucleus (micronucleus) used for propagation of genetic information, and a somatic nucleus (macronucleus) used for cell growth (Prescott, 1994). When starved cells of different mating types pair, they exchange haploid micronuclei following meiosis, perform one to several rounds of micronuclear mitosis, and then develop a new macronucleus from one of the newly formed micronuclei. This differentiation program is associated with reorganization of the genome and in some cases extraordinary genome rearrangements. In all ciliate lineages studied genome rearrangements are epigenetically determined by communication of DNA content between macronuclei and micronuclei via RNA intermediates (Nowacki et al., 2011). Several factors intrinsic to germline specialization were first characterized in ciliates, such as telomerase and histone acetyl transferase (Greider and Blackburn, 1987; Zahler and Prescott, 1988), as well as seminal studies on specialized histones, PIWI and HP1 proteins (Aeschlimann et al., 2014; Bannon et al., 1984; Bouhouche et al., 2011; Fetzer et al., 2002; Forcob et al., 2014; Huang et al., 1999; Jacobs et al., 2001; Jahn et al., 1997; Lipps et al., 1974; Mochizuki et al., 2002; Mochizuki and Gorovsky, 2004). Thus, ciliates provide relatively simple and facile systems to study principles of germline-soma specialization, germline differentiation and RNA-mediated epigenetic memory.

Stichotrichous ciliates are a special case in which nuclear duality led to the evolution of two extraordinary and distinct genomes. While micronuclear DNA is organized on long chromosomes similar to other eukaryotes, genes are interrupted by multiple, short, noncoding DNA sequences called internally eliminated sequences (IESs) that interrupt gene pieces, called macronuclear destined sequences (MDSs). During macronuclear development IESs are precisely eliminated and MDSs are recombined to form a functional gene (Prescott, 1994). MDS recombination in some genes has resulted in the formation of scrambled genes, with some genes scrambled into more than 100 MDSs, sometimes dispersed over multiple loci. There are >200,000 IESs and >3,000 scrambled genes in *Oxytricha trifallax*, the only stichotrich whose micronuclear genome has been sequenced (Chen et al., 2014).

The DNA molecules in the macronucleus of stichotrichous ciliates are the smallest known in nature, on average approximately 2kb. Each DNA molecules, present at 100-100,000 copies per macronucleus, typically contain a single coding sequence along with regulatory information and short telomeres (Aeschlimann et al., 2014; Swart et al., 2013). At the onset of macronuclear development micronuclear chromosomes undergo polytenization. Subsequently, macronuclear-destined sequences (MDSs) are unscrambled and spliced, IESs, non-genic DNA and transposable elements are removed and degraded, and gene-sized molecules are excised from chromosomes. These "nanochromosomes" are then replicated dozens of times and telomeres are added *de novo* to form a mature, functional macronucleus (Figure 1) (For general reviews of macronuclear development, see (Adl and Berger, 2000; Chalker and Yao, 2011; Jahn and Klobutcher, 2002; Prescott, 2000)). During this process there are millions of precise DNA splicing and ligation events resulting in a streamlined somatic genome of 20-fold reduced complexity. The complexity and sheer magnitude of DNA splicing and processing in stichotrichs dwarf those in the better characterized and distantly related ciliates *Tetrahymena* and *Paramecium*.

While these phenomena were first described over thirty years ago, we are now just beginning to characterize the molecular mechanisms underlying this extraordinary genome transformation (Greslin et al., 1989; Klobutcher et al., 1984). A key question surrounds how specific DNA sequences are precisely recombined, retained and amplified while others are excised and eliminated. For instance, the junctions of MDSs and IESs contain short "pointer" sequences that are likely involved, but inadequate, for proper MDS splicing, as their sequence can occur multiple times within the gene (Prescott and DuBois, 1996). Therefore, David Prescott proposed that a template DNA or RNA from the parental macronucleus must guide MDS splicing (Prescott et al., 2003). Long dsRNA transcripts, corresponding to entire macronuclear DNA molecules, are produced early during macronuclear development and are suggested to act as the proposed "templates" (Nowacki et al., 2008). Injection of synthetic long dsRNAs with altered MDS arrangements led to production of correspondingly altered macronuclear DNA molecules, not only in the injected cells, but in offspring as well, suggesting epigenetic inheritance through these RNA templates. In addition, 27 nt small RNAs mapping to both strands of macronuclear DNA molecules, called 27macRNAs, are produced en masse during early macronuclear development (Fang et al., 2012; Zahler et al., 2012). These 27macRNAs are associated with a PIWI homolog called Otiwi1, and are also referred to as piRNAs. This class of small RNAs specify which segments of micronuclear DNA will be protected from degradation during macronuclear development (Fang et al., 2012), perhaps by specifying DNA methylation of MDSs. The relationship between 27mer piRNAs and the long dsRNA "templates" involved in MDS rearrangements remains unknown.

Correspondingly little is known about the protein machinery involved in genome conversion in strichotrichs. Electron microscopy studies show dramatic reorganization of chromatin and nuclear architecture during the developmental program. Analyses of single genes suggest extensive chromatin changes occur during macronuclear development and that chromatin marks distinguish DNA regions with different fates (Bracht et al., 2012; Bulic et al., 2013; Postberg et al., 2008; Prescott, 1994). Not surprisingly strichotrichs encode a large array of histone proteins, several of which are expressed exclusively during macronuclear development (Aeschlimann et al., 2014; Bannon et al., 1984; Forcob et al., 2014; Jahn et al., 1997). The first study to identify mRNAs differentially expressed during macronculear development, in *Stylonychia lemnae*, utilized subtractive cDNA hybridization and cloning (Fetzer et al., 2002). This work identified a PIWI protein, a protein containing the largely uncharacterized Alba nucleic acid binding domain, a novel Kelch domain protein, and several other well-conserved DNA and RNA binding proteins. Another study suggested that the transposase encoded within a transposon family that is precisely excised during macronuclear development, called TBE transposons, is the enzyme responsible for producing dsDNA breaks for IES excision and MDS ligation. Indeed "domesticated" transposases are implicated in IES excision in other ciliate lineages (Baudry et al., 2009). It was recently reported that a paralog of RNA polymerase II second largest subunit is expressed exclusively during macronuclear development, and this factor, RBP2b, binds dsRNA templates first in the parental macronucleus and then in the developing macronucleus, suggesting a role in DNA rearrangements (Khurana et al., 2014). Electron microscopy studies from the 1970's showed that, coincident with the mass reduction in DNA content, proteinaceous "vesicles" transect the polytene chromosomes; potentially these "vesicles" contain the protein machinery involved in excision of macronuclear destined DNA molecules and/or degradation of the rest of the genome (Kloetzel, 1970; Murti, 1973, 1976; Prescott et al., 1973).

As an important step towards gaining a system-level understanding of the developmental program in strichotrichs and identifying the molecular machinery involved, we characterized the mRNA expression program during macronuclear development in *Oxytricha trifallax* via high throughput sequencing. We identified hundreds of mRNAs preferentially expressed at specific times during macronuclear development. We find that a disproportionate number of these mRNAs encode proteins that are involved in DNA and RNA functions. Many mRNAs preferentially expressed during macronuclear development have paralogs that are not differentially expressed during development, suggesting that gene duplication and functional specialization were a key source of evolutionary innovation. While many mRNAs preferentially expressed during macronuclear development encode proteins with no recognizable homolog, hundreds encode proteins that are well-conserved, including a disproportionate number with links to germline function or development in multicellular lineages. These analyses have thus identified scores of factors with "ancient" roles in sexual propagation, germline specialization and differentiation, in addition to dozens of novel factors with striking expression patterns, some of which likely evolved to carry out these extraordinary DNA manipulations. Understanding the molecular mechanisms underlying stichotrichs' extraordinary genetic apparatus and genome transformation will broadly inform our understanding of molecular and evolutionary biology and may also identify new means to manipulate genomes.

Materials and Methods

Vegetative growth of *Oxytricha trifallax*

Oxytricha trifallax strains ALXC2 and ALXC9 (Zahler et al., 2012) were grown vegetatively in Pyrex dishes in inorganic salts media (Chang et al., 2004) using the food source *Chlorogonium elongatum* (UTEX collection strain B203). Typical daily feedings consisted of 10 mL of washed algae per 300 mL Pyrex dish of *Oxytricha trifallax*, depending on culture density.

Mating of *Oxytricha trifallax*

Mating competent strains ALXC2 and ALXC9 were grown vegetatively to high density, fed lightly the day before a mating and allowed to completely exhaust their food supply. Cells were then cotton filtered to remove any residual algae and were concentrated on 10μM Nitex membranes into Pringhsheim salts media (Fang et al., 2012). Each individual mating strain was counted and cells were then mixed at equal numbers to a total concentration of 1,500 cells per mL in Pyrex dishes. These mating cells were then fed 1 mL of unwashed *Klebsiella pneumonia* stationary phase culture as a food source. Aggregates of 10-30 cells were observed by 2 hours after mixing mating types, with the first mating pairs visible by 4 hours post mixing. Typical mating efficiency when mixing ALXC2 and ALXC9 strains is ~70%, with visible anlage present by 48 hours.

Total RNA isolation

Pyrex dishes of mating cells were cotton filtered to remove cellular debri and concentrated onto 10 μM Nitex and transferred into microcentrifuge tubes. Cells were gently pelleted for 2 minutes at 500 x g in a microcentrifuge and supernatants were removed leaving 50 μL of pelleted cells. 200 μL of mirVana Lysis/Binding Buffer from the mirVana miRNA Isolation Kit (Ambion) was added to each tube. Total RNA was purified using the kit's protocol for total RNA purification. Total RNA yields from a single 300mL Pyrex dish of cells (~450,000 cells) were typically between 100 and 300μg.

mRNA cDNA Library preparation

mRNA cDNA libraries were prepared using the TruSeq RNA Sample Preparation v2 Kit (Illumina) following the manufacturer's LT protocol. Library preparation started with 3μg of total RNA from each of the control and mating cell timepoints. Poly-A containing mRNA molecules were selected for using poly-T oligo attached beads with two rounds of purification. Individual libraries were analyzed on a Bioanalyzer (Agilent) for cDNA quality and bar-coded libraries were pooled into one sample for sequencing.

Illumina Sequencing

Sequencing of the pooled libraries was performed in one lane of the Illumina HiSeq 2000 Sequencer at the UCSC Genome Technology Center. 100bp paired-end reads of the libraries were obtained.

Phylogenetic Tree Generation

Multiple sequence alignments for all RPB1 and SPT5 proteins were performed using Clustal-Omega 1.1.0 Multiple Alignment on the Mobyle@Pasteur portal (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::clustalO-multialign)(Sievers et al., 2011). These Clustal-Omega multiple sequence alignments were then run through PhyML 20130219 on the Mobyle@Pasteur portal (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::phyml) to create maximum likelihood phylogenetic trees (Guindon and Gascuel, 2003). 100 bootstrap sets were used for all of the multiple sequence alignments analyzed. The maximum likelihood Newick Tree Format files that included the bootstrap values produced using PhyML were then uploaded into the tree viewer PhyloWidget (www.phylowidget.org) to create tree images (Jordan and Piel, 2008). The tree images incorporate branch lengths, but bootstrap values were not included in the tree images for the sake of visualization.

Alignment and Post-alignment Processing

Raw sequencing results in the form of FASTQ files were used as input for alignment of the sequencing data to the *Oxytricha trifallax* macronuclear genome reference RNA database (Swart et al., 2013). The reference RNA GTF and macronuclear genome sequence files were retrieved from the *Oxytricha* genome website (oxy.ciliate.org). Alignments were performed with Tophat2, which produced

BAM alignment files (Kim et al., 2013). Tophat BAM files were used directly as input into Cuffdiff2 to generate normalized FPKM values (Trapnell et al., 2013).

Sequencing data analyses

Gene-centric normalized FPKM values were filtered as follows: mRNAs in which FPKM was not $\geq$ 3 in at least one sample were removed, remaining mRNAs were $\log_2$ (FPKM +1) transformed, and each mRNA was then normalized by subtracting the mean $\log_2$ (FPKM+1) value of 0 hour time points. mRNAs preferentially expressed during macronuclear development were defined as those whose average relative mRNA expression in one of the developmental time points was at least 3-fold greater than average 0 hour and vegetative cells. Normalized relative expression of these 1162 mRNAs was input for WGCNA (Langfelder and Horvath, 2008), which, under default settings with signed correlations, produced six modules. One module was removed because mRNAs were highly expressed in one of the two samples from vegetative cells. Modules were ordered by temporal expression and mRNA similarity within modules was determined with Cluster 3.0 using average-linkage centered Pearson correlation and results were visualized with Java Treeview (Saldanha, 2004).

***Oxytricha* protein domain and Gene Ontology analyses**

Protein domains and GO annotations were retrieved from the *Oxytricha* genome website (oxy.ciliate.org). The p-values of enrichment of protein domains and GO terms in specific gene sets were determined using the hypergeometric density distribution function and corrected for multiple hypothesis testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

Identification of Orthogroups

Orthogroups were defined using Orthofinder with default settings (Emms and Kelly, 2015). *Oxytricha, Stylonychia lemnae, Tetrahymena thermophila* and *Paramecium tetraurelia* protein fasta files were obtained from their respective genome websites. Fasta files for *Saccharomyces cerevisiae, Arabidopsis thaliana, Drosophila melanogaster, Caenorhabditis elegans* and human were obtained from the InParanoid website (inparanoid.sbc.su.se). Gene annotations were obtained from the genome sites for *Tetrahymena* (www.ciliate.org), *Paramecium* (www.paramecium.cgm.cnrs-gif.fr) and *Saccharomyces* (http://www.yeastgenome.org) and from Uniprot for *Arabidopsis thaliana, Drosophila melanogaster, Caenorhabditis elegans* and human (www.uniprot.org).

Curated Gene lists

Genes associated with RNA functions (“RNA”), DNA synthesis and repair (“DNA”), RNA polymerase II transcription (“transcription”) and chromatin (“chromatin”) were identified as follows:

RNA

Retrieved genes annotated as RBPs in humans and common RNA binding domains (Gerstberger et al., 2014). Retrieved genes associated with RNA metabolism in *Saccharomyces cerevisiae* (Hogan et al., 2008). Manually added specific genes based on homology to genes linked to RNA functions in literature. A compendium of *Oxytricha* genes linked to RNA biology was generated by first identifying orthogroups in which members of the orthogroup were annotated as RBPs in human or yeast. To this set we added genes with canonical RNA binding domains and whose GO or protein domain annotation included “RNA”. This set was manually filtered to remove genes that did not appear to directly be associated with RNA functions.

DNA

Retrieved genes associated with “DNA repair” and “DNA replication” in humans from Reactome database (www.reactome.org). Retrieved genes associated with “DNA repair”, “DNA replication” and “DNA recombination” in yeast from SGD. A compendium of *Oxytricha* genes was generated by first identifying orthogroups in which members of the orthogroup were in the lists above. To this set we added genes whose GO or protein domain annotation included “DNA”. This set was manually filtered to remove genes that did not appear to directly be associated with DNA replication, repair or recombination.

Transcription

We retrieved genes associated with “RNA polymerase II Transcription” in humans from the Reactome database (www.reactome.org). We retrieved genes associated with “core RNA polymerase II recruiting transcription factor activity”, “RNA polymerase II core promoter sequence-specific DNA binding”, “RNA polymerase II core promoter sequence-specific DNA binding transcription factor activity involved in preinitiation complex assembly”, “DNA-directed RNA polymerase activity” in yeast from the SGD (http://www.yeastgenome.org). To this set we added genes whose GO or protein domain annotation included “transcription”. This set was manually filtered to remove genes that did not appear to directly be associated with core RNA polymerase II transcription.

Chromatin

We retrieved genes associated with “chromatin organization” in humans from the Reactome database. We retrieved genes associated with “chromatin modification” and “chromatin remodeling” in yeast from the SGD. Histone genes in *Stylonychia* were retrieved from the genome sequence (Aeschlimann et al., 2014) and via BLAST searches against the *Oxytricha* genome. We added genes whose GO or protein domain annotation included “chromatin” and genes with bromo or chromo domains. This compiled gene set was then manually filtered.

Gene Naming Conventions

Gene naming largely focused on orthogroups whose members were annotated to one of four groups described above and/or for which at least one member was preferentially expressed during macronuclear development. Emphasis was placed on naming genes with orthogroup members in one or more of the non-stichotrich species. Generally, *Oxytricha* genes within an orthogroup were ordered by absolute expression level in vegetative cells (based on average FPKM) and names were guided by names associated with orthogroup members with emphasis on human and yeast members. The member of the orthogroup with highest absolute expression in vegetative cells were assigned the name GENEa, then second highest GENEb and so on.

Results

mRNA expression profiles during macronuclear development in *Oxytricha trifallax*

In order to identify mRNAs preferentially expressed during macronuclear development we performed high-throughput sequencing of poly-A selected RNAs isolated from various developmental timepoints (Figure 1). Biological replicate sequencing data were obtained from seven timepoints including vegetatively growing cells as well as 0, 6, 12, 24, 48 and 72 hours post-mixing of cells of complementary mating types. Each sample produced an average of 22 million paired-end reads. Raw sequencing reads were mapped to the published macronuclear genome reference transcriptome (Swart et al., 2013) using Tophat2 and normalized expression data in the form of FPKM values were obtained with Cuffdiff2. We obtained quality measurements (FPKM of 3 in at least one experiment) from 17055 of 24885 annotated mRNAs.

We used several approaches to define gene function. Protein domain and associated GO terms were extracted from genome database. Gene names and homology to characterized proteins in other eukaryotic lineages, including other published ciliate genomes, was limited to handfuls of genes. Therefore, we used Orthofinder (Emms and Kelly, 2015) to infer orthogroups (set of genes that are descended from a single gene in the last common ancestor of all the species being considered) among *Oxytricha trifallax*, the other published stichotrichous ciliate *Stylonychia lemnae*, the distantly related oligophors *Tetrahymena thermophila* and

Paramecium tetraurelia, as well as "model" organisms *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Drosophila melanogaster* and *Caenorhabditis elegans,* and human. Based on annotation information for orthogroup genes, protein domain information and previously published results, we annotated and named ~2300 *Oxytricha* genes, including 402 genes preferentially expressed during macronuclear development as defined below. As we expected many genes involved in macronuclear development are linked to RNA and DNA related processes. We also manually curated gene sets linked to RNA-binding ("RNA"), DNA synthesis and repair ("DNA"), "Chromatin" and "Transcription".

In order to study mRNA expression as a function of macronuclear development, FPKM values were $\log_2 +1$ transformed and the average value from the 0 hour time point was subtracted from each sample. mRNAs preferentially expressed during macronuclear development were defined as those in which average fold change for one developmental time-point was at least 3-fold greater than 0 hour and vegetative cells, resulting in 1162 mRNAs. We grouped the 1162 mRNAs according to their relative expression during macronuclear development using weighted correlation network analysis (WGCNA) (Langfelder and Horvath, 2008), which identified six expression modules (Figure 2A), corresponding to different temporal patterns (Figure 2B). One of the modules, including 58 mRNAs, was omitted because the mRNAs were abundantly expressed in vegetative cells and thus not "macronuclear development" specific.

We utilized curation information to gleam broad themes among genes preferentially expressed during macronuclear development. Relative to all mRNAs for which we obtained quality measurements there was a striking enrichment for genes with protein domains linked to DNA and RNA metabolism (Figure 2C). For instance, there were 105 orthogroups containing 614 genes annotated as “DNA binding” ($p < 1e-16$), 19 orthogroups containing 84 genes annotated as “DNA replication” associated ($p < 1e-4$), 24 orthogroups containing 155 genes annotated to be involved in “DNA repair” ($p = 0.001$), and 7 orthogroups containing 8 genes annotated as “mRNA catabolic process” related ($p < 1e-7$). Among protein domains, RRM, dsRBD, HMG box, PHD-finger, Core histone, PARP, CHROMO, Alba DEAD/DEAH box helicase, OB-fold and MULE transposase were widely utilized. Using our manually curated lists, 107 of 987 RNA genes, 150 of 398 DNA genes, 61 of 365 Chromatin genes, and 25 of 113 Transcription genes were preferentially expressed during macronuclear development (Table 1). There was a modest change in proportion of RNA/transcription genes earlier relative to DNA/chromatin, consistent with production of template RNA guides and PIWI-associated 27macRNAs preceding the macronuclear DNA rearrangements. Individual modules are described in more detail below.

Module 1

Module 1 includes 110 mRNAs whose expression peaks at 6 to 12 hours into macronuclear development. During this time, mating cells have begun to conjugate and meiosis of the maternal micronucleus begins to occur. Many membrane protein-associated genes are upregulated which may play roles in mating and conjugation, including a TRP ion channel, a Zip zinc transporter, a Pitr-6 homolog involved in phosphate transport and also a nuclear pore protein associated with protein import to the nucleus. Several meiosis linked mRNAs are also upregulated at this time, including Hop2 and Dmc1, known to play roles in homologous chromosome pairing as well as meiotic recombination (Bishop et al., 1992; Brown and Bishop, 2015; Petukhova et al., 2003). Not surprisingly DNA replication factors (MCM6 and MCM8, SNF2 DNA helicase), multiple cyclins and cyclin dependent kinases (CDK3 and another unknown CDK) also peak, likely corresponding to the initiation of micronuclear DNA replication. RNA metabolism proteins make up the majority of mRNAs upregulated in Module 1 and include PIWI homologs *Otiwi4 and Otiwi7*, ALBA family nucleic acid binding proteins, an RNA-dependent RNA polymerase, Dicer2, a DEAD/DEAH box helicase and multiple RNA recognition domain (RRM) containing proteins, including a dsRNA binding protein. Interestingly there are also dynein and myosin motor protein subunits upregulated at this time, which may correspond to the shuttling of nuclei around during conjugation and meiosis.

Module 2

Module 2 includes 360 mRNAs whose expression peaks at 12 to 24 hours into macronuclear development. During this time frame, meiosis of the parental micronucleus is ongoing/completed and the early stages of polytenization of micronuclear chromosomes in the anlagen begins to occur. By 12 hours after mixing complementary mating types, all ciliates that are going to mate will have formed conjugant pairs. Little is known about what actually constitutes different mating types in *Oxytricha*, but interestingly a homolog of *prgU*, a sex pheromone gene expressed in *Enterococcus faecalis* bacteria is highly upregulated (Hirt et al., 2005). Many meiotic genes continue to be highly expressed in Module 2, including REC8, part of the cohesin complex responsible for binding sister chromatids together and MSH4 and MSH5, shown to play roles in mediating homologous recombination (Pochart et al., 1997; Ponticelli and Smith, 1989). In addition, two *Bub1* homologs, serving roles in the spindle assembly checkpoint and to recruit REC8, and CENPO, a central component of the kinetochore, are coexpressed (Basu et al., 1998; Yamaguchi et al., 2003). Upon completion of the micronuclear meiotic divisions, haploid micronuclei are exchanged between mating cells. Four Kinesin homologs are upregulated, including Kinesin-1, which has known roles in driving nuclear distribution in muscle cells (Wilson and Holzbaur, 2015). DNA replication factors and DNA damage genes make up a large portion of the mRNAs we have identified in this module. DNA replication factor CTD1 along with MCM2 and MCM9, distinct MCMs from Module 1, play a role in replication at this stage. DNA double-stranded break repair genes are widely utilized and include *Mnd1, Mre11, Brca2* and a V(D)J DBS repair gene. In addition, six distinct Poly ADP ribose polymerases (PARPs) peak at this time, also likely to play roles in DNA repair. Timing of the expression of these DNA replication factors/repair genes coincides with the initiation of polytene chromosome formation in the developing macronucleus. SMC1 and SMC3 both components of the cohesin protein complex, and SMC4, a component of the condensin protein complex family, are highly expressed alongside RCC1, which is also involved in chromosome condensation (Nishimoto et al., 1978; Strunnikov and Jessberger, 1999). *Top2*, a gene known to actively promote chromosome pairing in *Drosophila* is also expressed (Williams et al., 2007). RNA metabolism genes definitely make up the majority of homologs identified as differentially expressed during this time period. RNA Polymerase II subunits RPB2, RPB4 and RPB11, along with several transcription factors including TFIIS, SPT4 and SPT5 are upregulated. These are likely to play a role in the transcription of 27macRNAs or their biological precursors as well as the long double-stranded guide RNAs involved in MDS unscrambling, both of which peak in expression at 24 hours post-mixing. Additional evidence for this includes the spike in expression of the PIWI homolog *Otiwi1*, shown to associate with 27macRNAs (Fang et al., 2012), and the homolog of *Ema1*, known to play a role in the interaction between *Twi1p* and chromatin in the distantly related ciliate *Tetrahymena* (Aronica et al., 2008). Several other protein domains involved in nucleic acid interactions are also broadly utilized, including RRMs, Zinc/RING fingers, HMG boxes and chromodomains.

Module 3

Module 3 includes 62 mRNAs whose expression peaks at 24 to 48 hours into macronuclear development. By this time in macronuclear development, chromosome polytenization is complete and excision of internally eliminated sequences (IESs), transposable elements and non-genic DNA begins to occur. In addition, this module corresponds to the approximate timing of the initiation of rearrangement and reordering of macronuclear destined sequences (MDSs). Many of the genes upregulated in Module 3 play roles in protein modification, including acetyltransferases and glucosyltransferases. In addition, multiple proteases and protease inhibitors spike in expression and presumably play roles in both the protection and turnover of proteins involved in macronuclear development processes. Four putative chromosome scaffold proteins (undetermined chromosome scaffolds 7, 8, 104, 105) are also upregulated; these may play a role in the transient maintenance of the polytene state of chromosomes in the developing macronucleus. This module contains a homolog of *NOWA1*, which is implicated in elimination of transposable elements and a subset of IESs in *Paramecium* through nuclear crosstalk between the maternal and developing macronucleus (Nowacki et al., 2005).

Module 4

Module 4 includes 491 mRNAs, whose expression peaks at 48 to 72 hours into macronuclear development, with an emphasis on the earlier timepoint. During this time excision of internally eliminated sequences (IESs), transposable elements and non-genic DNA will have occurred, along with the rearrangement and reordering of macronuclear destined sequences (MDSs). This timing also coincides with the initiation of bulk DNA elimination and the initiation of the last rounds of nanochromosome replication. Multiple nucleases are upregulated, including members of the exonuclease 1 family and the exonuclease III family, as well as two distinct Flap endonucleases, most likely involved in bulk DNA elimination happening during this time. A CAF1 family RNA exonuclease homolog of PARN-1, shown to play a role in piRNA end trimming in *C. elegans*, also peaks during this time and may be involved in sRNA biogenesis (Tang et al., 2016). DNA polymerase II, DNA polymerase III and DNA polymerase IV subunits also spike in expression, along with various replication factors such as Replication Protein A (RPA) homologs, PCNA, MCM8 and Dna2, a DNA replication helicase, presumably playing roles in the replication of nanochromosomes. Not surprisingly, Module 4 also contains dozens of DNA repair enzymes from numerous families, including MutS and MutL domain-containing DNA mismatch repair proteins (MSH6, MLH1, PMS1/PMS2) and members of the Rad50 and Rad51 protein families. Consistent with when DNAs are cut and telomeres are added *de novo*, two Pot1-like alpha telomere-binding proteins are upregulated, playing roles in the protection of the newly added telomeres (Baumann and Cech, 2001). Histone variants (H2A, H2B and H3), histone methyltransferases, and proteins involved in chromosome organization and condensation, including condensin complex subunits and SMC2, which has been shown to be required for MIC and MAC division in *Tetrahymena*, also make up a large portion of the genes we see differentially expressed in this module (Cervantes et al., 2006). RRM-containing proteins are also widely utilized as well the PIWI homologs *Otiwi3* and *Otiwi11*, performing unknown functions at this stage in macronuclear development

Module 5

Module 5 includes 81 mRNAs, whose expression peaks at 48 and 72 hours into macronuclear development, with emphasis on the latter timepoint. During this time, bulk DNA elimination in the developing macronucleus has been completed and nanochromosome replication is taking place to complete macronuclear maturation. Multiple DNA Replication Protein A homologs are upregulated along with the DNA mismatch repair protein MSH6 and other MutS domain containing proteins involved in DNA repair. Proteins playing roles in nucleosome assembly and chromatin remodeling such as SNF2-type helicases, histones H2A and H3, and HMG-box containing proteins also appear, most likely contributing to the last rounds of replication and packaging of the newly formed macronuclear nanochromosomes. During this stage in macronuclear development, bulk DNA elimination is completed and we find chromodomain-containing homologs of LHP1, involved in chromatin organization, and Pdd1, involved in DNA elimination from the developing macronuclear genome in *Tetrahymena* highly expressed (Coyne et al., 1999; Libault et al., 2005). RNA recognition motif-containing proteins, including some KH domain-like and a dsRNA binding protein are upregulated, alongside PIWI homolog *Otiwi2*; the roles of PIWIs at this stage of macronuclear development remains unclear. The Zinc/RING finger protein domain is also widely utilized in Module 5.

Gene Duplication and Specialization

Where do these genes come from? In the process of analyzing the data and curating genes, we noticed that many mRNAs whose expression was specific to macronuclear development appeared to have paralogs either not preferentially expressed during macronuclear development or expressed at a different time during macronuclear development. Paralagous genes with divergent mRNA expression is often an indication/suggestion of functional specialization post gene duplication, and could be one mechanism to evolve the machinery necessary to carry out the macronuclear development program (Semon and Wolfe, 2007). Indeed, many of the handfuls of previously characterized macronuclear development-specific genes are part of evolutionarily conserved families, with multiple members in *Oxytricha* that have divergent expression patterns. For instance, *Oxytricha* encodes eleven PIWI proteins, six of which are macronuclear development specific, but with different temporal patterns and absolute expression levels (Table 2). Another example entails *Rbp2*, which encodes the second largest subunit of RNA polymerase II core complex, as is commonly found as a singleton in eukaryotes. However, in *Oxytricha* there is a constitutively expressed paralog and one strikingly specific to macronuclear development (Figure 3) (Khurana et al., 2014).

In order to determine if gene duplication and specialization of an isoform for macronuclear development is widespread, we systematically identified paralogous gene sets for the orthogroup dataset (set of genes that are descended from a single gene in the last common ancestor of all the species being considered) in which at least one member was macronuclear development-specific. Restricting our analyses to orthogroups with fewer than 20 members in *Oxytricha*, there were a total of 2169 orthogroups that met this criterion, of which 233 groups (367 mRNAs) had at least one member that was macronuclear development-specific; of these 183 had at least one member that was not macronuclear development specific and for the other 50, all members were macronuclear development-specific. Nearly half of these 233 orthogroups included members associated with the RNA, DNA, chromatin or transcription gene sets (Figure 4A) (49 RNA orthogroups, 46 DNA orthogroups, 22 chromatin orthogroups and 18 transciption orthgroups). These orthogroups were often phylogenetically conserved with over half (131) having a presumptive human ortholog (Figure 4C). Thus, ~30% of macronuclear development specific genes are members of phylogenetically conserved paralogous gene families, many of which encode proteins linked to RNA and DNA biology. Below we highlight a specific case:

RNA polymerase II transcription

RNA polymerase II transcribes mRNA precursors and is nearly universally composed of 12 core subunits in eukaryotes, named RPB 1-12. In most species RPBs are singletons and there is a single core RNA polymerase II complex with a suite of other factors that regulate initiation, elongation, termination and processing. However, in *Oxytricha*, there are two paralogs of RBP1, 2, 4, 7 and 10 and three paralogs of 11; in each case one of the paralogs is preferentially expressed during macronuclear development (Figure 5), and in several cases mRNA levels of the macronuclear development-specific paralog are undetectable in vegetative cells, but are dramatically induced during macronuclear development and rise to levels that rival or exceed that of the constitutively expressed paralog. In addition to these core components of RNA polymerase II, a number of factors that assist in initiation, elongation, termination as well as co-transcriptional capping, splicing and polyadenylation have multiple paralogs in *Oxytricha* with at least one preferentially expressed during mac development. These include elongation factors SPT5, SPT4, TFIIS, TFIIF, ELF1 and SSRP1, initiation factors TBP1 and IWS1, and processing factors SUB1, SEN1, RTT103, CBP20, U2AF1, SC35 and SUB2 (Figure 6). As with the core components, macronuclear development-specific accessory proteins are often not detected in vegetative cells and rise to levels on par with constitutive paralog at one or more stages during macronuclear development. As shown in Figures 5 and 6, the expression patterns of these mRNAs during macronuclear development fall into two main groups. There is one subset whose expression peaks at 12-24 hours, but often remains highly expressed throughout macronuclear development and there is a second group that has variable expression at early time points and peaks at 48 and 72 hours. These results suggest there are potentially at least two RNA polymerase II-like complexes with specific roles during mac development in *Oxytricha*.

The best characterized example of duplication and functional specialization of RNA polymerase II subunits comes from *Arabidopsis thaliana* and other flowering plants, which contain two additional nuclear multisubunit RNA polymerases, RNA Polymerase IV (Pol IV) and RNA Polymerase V (Pol V). These plant-specific RNA polymerases have non-redundant roles in RNA-mediated gene-silencing pathways, specifically in RNA-directed DNA methylation (RdDM) (Tucker et al., 2010) (Reviewed in (Haag and Pikaard, 2011; Matzke et al., 2015). Pol IV is responsible for transcribing long RNAs, copied into dsRNAs by an RNA-dependent RNA Polymerase (RDR2). After these double-stranded substrates are cleaved by

Dicer-like enzyme DCL3, they associate with the Argonaute family protein AGO4 to form a RISC-complex. Pol V produces nascent long non-coding RNA (lncRNA) transcripts from specified regions of the genome which base-pair with the AGO4-associated siRNAs and result in *de novo* cytosine methylation of the corresponding DNA template. This often leads to gene silencing through repressive histone modifications (Haag and Pikaard, 2011; Kanno et al., 2010; Matzke and Mosher, 2014). Proteomic analyses have revealed that *Arabidopsis* Pol IV and Pol V have a 12-subunit composition like Pol II. In fact, half of the subunits of Pols II, IV, and V are encoded by the same genes. The remaining Pol IV- or Pol V-specific subunit genes arose through duplication and sub-functionalization of ancestral Pol II subunit genes. Unique paralogs of the largest subunit of Pol II (NRPB1) make up the catalytic core of the polymerases and are unique to either the Pol IV or Pol V complex, being referred to as NRPD1 and NRPE1 respectively. While the NRPB1 C-terminal domain (CTD) contains heptapeptide repeats, the CTDs of both NRPD1 and NRPD1 lack this signature, likely facilitating their alternative functions. The NRPE1 CTD is extended by approximately 300 amino acids and is shown to associate with AGO4 through WG/GW repeats, called the Argonaute "hook", to direct DNA methylation (Li et al., 2006). Interestingly, although the macronuclear development-specific *Oxytricha* paralog of RPB1, RPB1b, lacks these repeats, it does have an extended CTD that may facilitate binding to Otiwi1, the closest homolog to plant AGO4, known to play roles in MDS retention during macronuclear development (Fang et al., 2012).

In addition to core RPB components, an SPT5 paralog called SPT5L, is an essential component of these specialized plant RNA polymerase IV and V complexes. SPT5L associates with Pol V and plays roles in Pol V-mediated transcriptional silencing, binding regions in chromatin that will ultimately be transcriptionally silenced, potentially recruiting AGO4 and associated siRNAs to these loci (Rowley et al., 2011). SPT5L differs from SPT5 homologs in other species in that it has a large carboxyl-terminal extension, containing dozens of WG/GW repeats that allow interaction with AGO4 (Bies-Etheve et al., 2009). Upon investigation, we find that *Oxytricha* SPT5b lacks an extended C-terminus as well as the canonical WG/GW-containing Argonaute hook domain present in SPT5L, but does in fact have a ~300 amino acid extension at the N-terminal domain (NTD). Further biochemical studies are necessary to uncover the unique function of the NTD of *Oxytricha* SPT5b.

The presence of RPB paralogs that are differentially expressed during distinct stages of macronuclear development, along with the expression patterns of paralogous *Oxytricha* RNA polymerase II-associated accessory factors also found in *Arabidpopsis* and other flowering plants such as SPT5, indicates that a non-canonical, specialized RNA pol II-like complex exists during macronuclear development and may play unique roles in the biogenesis and function of template guide RNAs and/or PIWI-associated 27macRNAs involved in genome rearrangements.

Discussion

Stichotrichous ciliates, such as *Oxytricha trifallax*, have evolved two remarkable genomes that push the limits of our understanding of DNA manipulation, epigenetic inheritance and evolution's potential. Here, we have characterized the mRNA expression program throughout macronuclear development in *Oxytricha trifallax* and find that hundreds of genes are preferentially expressed during macronuclear development, clustering into five regulatory sets or modules with distinct functional themes. A disproportionate number of genes encode well-conserved proteins involved in DNA and RNA metabolism across species. Several protein domains are widely utilized, such as RRM, dsRBD, HMG box, PHD-finger, Core histone, PARP, CHROMO, Alba, DEAD/DEAH box helicase and Replication Protein A (RPA). A large fraction of these genes have paralogs that are not differentially expressed during macronuclear development, suggesting that genome duplications and functional specialization of homologous genes was a key source of evolutionary innovation. As discussed previously, many components of the RNA polymerase II machinery have paralogs specifically expressed during macronuclear development, including many RNA polymerase II accessory factor proteins, potentially playing unique roles in the biogenesis and function of guide RNA templates and PIWI-associated 27macRNAs involved in DNA elimination/retention and MDS unscrambling in the developing macronucleus, perhaps similar to the roles of Pol IV and Pol V in plants. Other interesting families with macronuclear development-specific paralogs include histones, RPAs, PARPs, RNA/DNA helicases and DNA replication, damage and recombination factors.

While most organisms contain only a single homolog of each of the proteins making up the Replication Protein A (RPA) complex, the *Oxytricha* genome codes for substantially more (Table 3). RPA is a heterotrimeric protein complex, widely conserved across eukaryotes. RPA is involved in many aspects of DNA metabolism, binding with high affinity to single-stranded DNA (ssDNA) (Fairman and Stillman, 1988; Wold and Kelly, 1988). RPA functions during DNA replication to prevent reannealing of DNA strands or formation of secondary structure (Fairman and Stillman, 1988; Wold and Kelly, 1988). RPA protects against degradation of the ssDNA strands by nucleases and is also capable of directly unwinding double-stranded DNA (dsDNA) complexes (Georgaki et al., 1992; Treuner et al., 1996). RPA is involved in DNA repair by binding damaged DNA and playing roles in nucleotide excision (Coverley et al., 1992; Coverley et al., 1991), as well as contributing to homologous recombination processes (Heyer et al., 1990; Moore et al., 1991), DNA strand exchange (Sung and Robberson, 1995) and double-strand break repair (Firmenich et al., 1995; Smith and Rothstein, 1995). Interestingly, hsRPA has also been shown to directly interact and bind the RNA Polymerase II complex, increasing the activation of transcription (Maldonado et al., 1996). While *Oxytricha* constitutively express six RPA homologs, most likely performing canonical everyday functions, eight RPAs are macronuclear development specific, all peaking among 48-72 hours post-mixing of mating types. We hypothesize that the differentially expressed RPAs, stemming from genome duplications, may play specialized roles in the last steps of macronuclear development, including protection against DNA damage during the last rounds of nanochromosome replication, as their expression peaks during this time frame. Considering, the complexity and magnitude of the large-scale genome rearrangements occurring throughout macronuclear development, it is not difficult to imagine the necessity for additional RPAs to aid in these processes.

Recently, evidence has emerged to implicate “R Loops” in the disruption of genome stability and also in gene expression regulation (reviewed in (Skourti-Stathaki and Proudfoot, 2014). R loops are formed when a nascent RNA hybridizes with a DNA template, leaving one ssDNA molecule displaced and susceptible to DNA damage. R loops form naturally during transcription and have also been shown to exist over the promoter sequences of non-transcribed genes as a method to regulate their expression, bound by a ssDNA-binding protein that stabilizes their structure (Sun et al., 2013). Surely, a tremendous number of R loops must be formed during the mass transcription of the macronuclear genome to create the guide template RNAs and the PIWI-associated 27macRNAs involved in genome rearrangements in *Oxytricha*. Presumably, there must also be R loops formed when the guide template RNAs and/or 27macRNAs pair with complementary DNA to aid in the retention and unscrambling of MDSs, which may present a major threat to genome integrity. In addition to its traditional function in termination of transcription of short non-coding RNA genes (Martin-Tumasz and Brow, 2015), the RNA/DNA helicase Sen1, has been shown to actively remove R loops, helping to protect against DNA damage

(Mischo et al., 2011). In addition to constitutively expressed *Sen1*, *Oxytricha* also has a paralog called *Sen1b* that peaks in expression at 12-24 hours post-mixing, during the early stages of macronuclear development corresponding to the peak in transcription of guide templates and 27macRNAs. We hypothesize that Sen1b plays an important role in preserving genome integrity by resolving R loops during transcription of these RNAs and perhaps takes part in the termination of these transcripts. Sen1b may also be involved in mediating genome rearrangements, based on the fact that Sen1 also interacts with genes involved in homologous recombination in yeast. Further work is needed to elucidate the unique functions of *Oxytricha* Sen1b.

Here we have shown that close to 30% of macronuclear development specific genes are members of phylogenetically conserved paralogous gene families, many of which encode conserved proteins linked to RNA and DNA biology across species and thus have helped to identify a core set of "ancestral" factors involved in the preservation of genome integrity. We also find that many human homologs of macronuclear development specific genes are expressed exclusively in germline and stem cells, although additional and more thorough analyses are necessary to elucidate how extensive and widespread this is. This work provides an enticing glimpse into the *Oxytricha* macronuclear development program and helps to refine future experiments aimed at dissecting the underlying molecular mechanisms behind these processes both in ciliates and higher level eukaryotes.

Acknowledgements

First, I would like to thank Daniel Hogan and Sol Katzman for their help with our bioinformatics analyses, including alignment of expression libraries and performing module analysis, using either WGCNA or DEseq, respectively. I'd like to thank Micaela Neeb for Adobe Illustrator tips and tricks and for help putting together Figure 1. I'd also like to thank colleague Michael Doody for his help comparing *Oxytricha* SPT5 paralogs to those in yeast. Finally, I'd like to thank my mentor Alan Zahler, as well as Jeremy Sanford and Melissa Jurica, who provided helpful discussions throughout the work performed.

Figure 1: ***Oxytricha*** **sexual life cycle. 1.** Two vegetative *Oxytricha* cells of different mating types (represented by the difference in nuclei colors). **2.** Under starvation conditions the two ciliates fuse and begin to conjugate. **3.** The parental micronucleus in each cell undergoes meiosis. **4-6.** Three of the newly formed haploid micronuclei will break down while the one remaining will undergo mitosis. One mitosis-derived haploid micronucleus is exchanged between the mating cells. **7.** The newly acquired micronucleus fuses with the remaining maternal micronucleus to become diploid. **8.** The newly formed diploid micronucleus undergoes mitosis. **9.** One newly formed micronucleus develops into a new macronucleus while the maternal macronucleus is broken down and degraded. **10.** Two genetically identical exconjugant *Oxytricha* cells. Inside the circle is a graph showing the general timing of macronuclear development events in hours and the corresponding DNA content of the developing macronucleus (anlagen). Outside of the circle are the alternate MIC and MAC versions of two hypothetical genes, one scrambled and one unscrambled (MDSs in orange, IESs and nongenic DNA in blue and telomeres in black).

Oxytricha **Sexual Life Cycle**

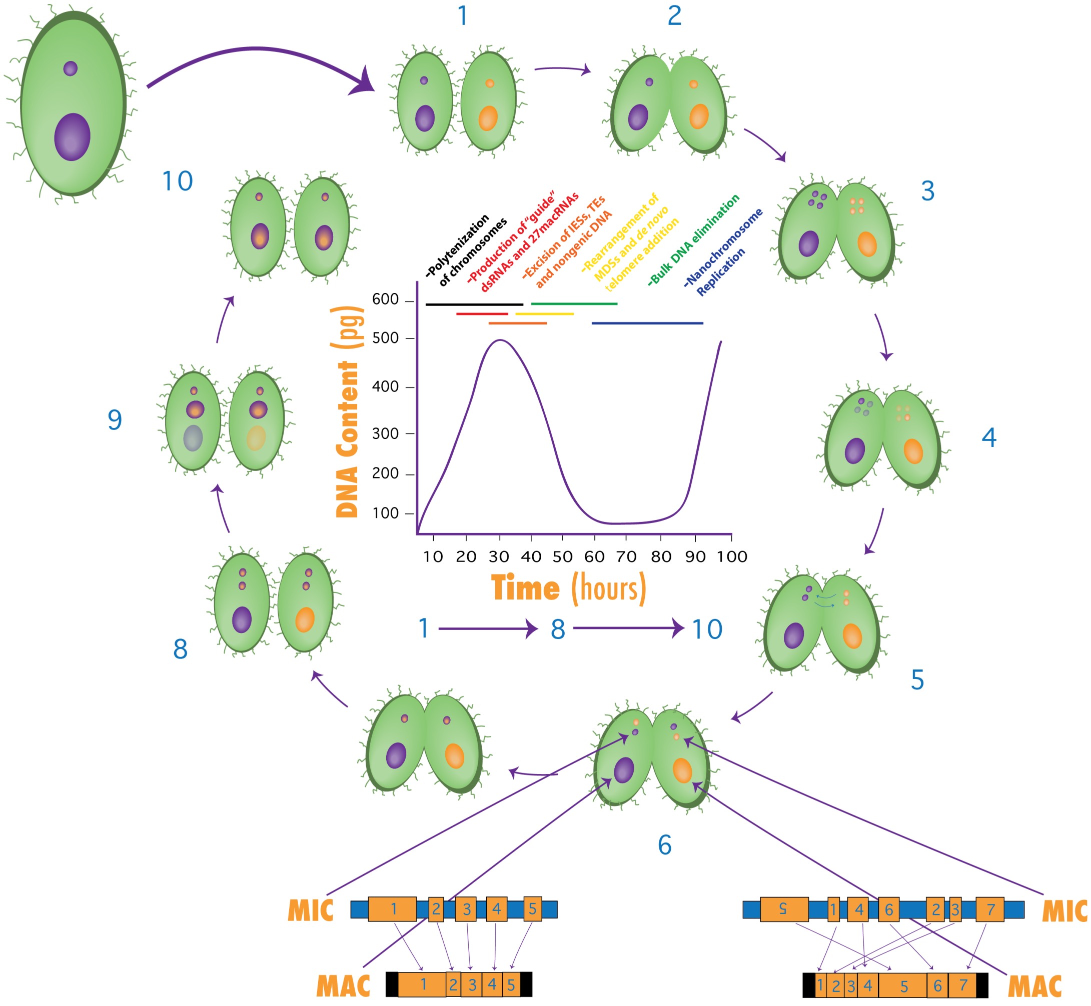

1
2
3
4
5
6
7
8
9
10
-Polytenization of chromosomes
-Production of "guide" dsRNAs and 27macRNAs
-Excision of IESs, TEs and nongenic DNA
-Rearrangement of MDSs and de novo telomere addition
-Bulk DNA elimination
-Nanochromosome Replication
600
500
400
300
200
100
DNA Content (pg)
10 20 30 40 50 60 70 80 90 100
Time (hours)
1 → 8 → 10
MIC
1 2 3 4 5
MAC
1 2 3 4 5
MIC
5 1 4 6 2 3 7
MAC
1 2 3 4 5 6 7

Figure 2: Overview of the mRNA expression program during macronuclear development in *Oxytricha trifallax*. A. Heat map representation of relative mRNA expression of 1104 mRNAs whose expression increases significantly during macronuclear development. mRNAs are grouped according to co-expression module (1-5) and within each module grouped by hierarchical clustering. Relative mRNA expression was normalized such that log2(FPKM +1) levels in 0 hour cells was zero on average. The color bar to the right of the figure indicates if the mRNA was in one of four manually curated gene sets linked to DNA and RNA biology. **B.** Bar plot representation of the average relative mRNA expression in each module at each time point. **C.** (Top) Enrichment of GO terms (rows) in mac dev specific mRNAs and modules 1-5. The significance of enrichment of the GO term is represented as a heat map (scale is below the figure) in which the color intensity corresponds to the negative log10 p-value, calculated using the hypergeometric density distribution function and corrected for multiple hypothesis testing using the Benjamini-Hochberg method. Only a subset of significantly enriched GO terms are shown. (Bottom) Same as above except for protein domains.

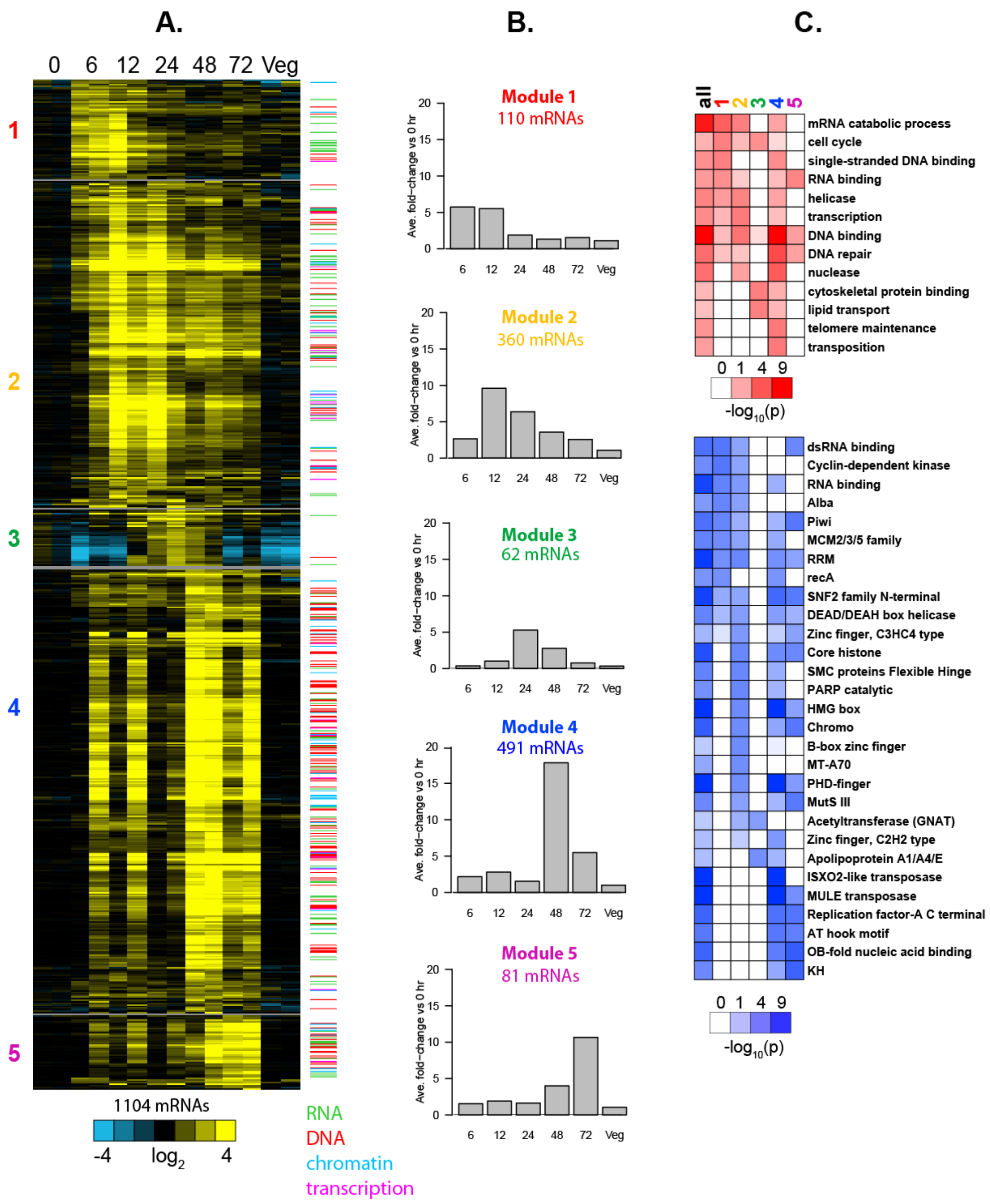

A.
0
6
12
24
48
72
Veg
1
2
3
4
5
1104 mRNAs
-4
$\log_2$
4
RNA
DNA
chromatin
transcription
B.
Module 1
110 mRNAs
Module 2
360 mRNAs
Module 3
62 mRNAs
Module 4
491 mRNAs
Module 5
81 mRNAs
Ave. fold−change vs 0 hr
0
5
10
15
20
6
12
24
48
72
Veg
C.
all
1
2
3
4
5
mRNA catabolic process
cell cycle
single-stranded DNA binding
RNA binding
helicase
transcription
DNA binding
DNA repair
nuclease
cytoskeletal protein binding
lipid transport
telomere maintenance
transposition
0 1 4 9
$-\log_{10}(p)$
dsRNA binding
Cyclin-dependent kinase
RNA binding
Alba
Piwi
MCM2/3/5 family
RRM
recA
SNF2 family N-terminal
DEAD/DEAH box helicase
Zinc finger, C3HC4 type
Core histone
SMC proteins Flexible Hinge
PARP catalytic
HMG box
Chromo
B-box zinc finger
MT-A70
PHD-finger
MutS III
Acetyltransferase (GNAT)
Zinc finger, C2H2 type
Apolipoprotein A1/A4/E
ISXO2-like transposase
MULE transposase
Replication factor-A C terminal
AT hook motif
OB-fold nucleic acid binding
KH
0 1 4 9
$-\log_{10}(p)$

Figure 3: Expression profiles of *Oxytricha* RPB2 paralogs throughout macronuclear development. *Oxytricha trifallax* Macronuclear Genome Browser screen shot of constitutively expressed RPB2a (Contig1141.1) and macronculear development specific RPB2b (Contig19200.0) coverage tracks.

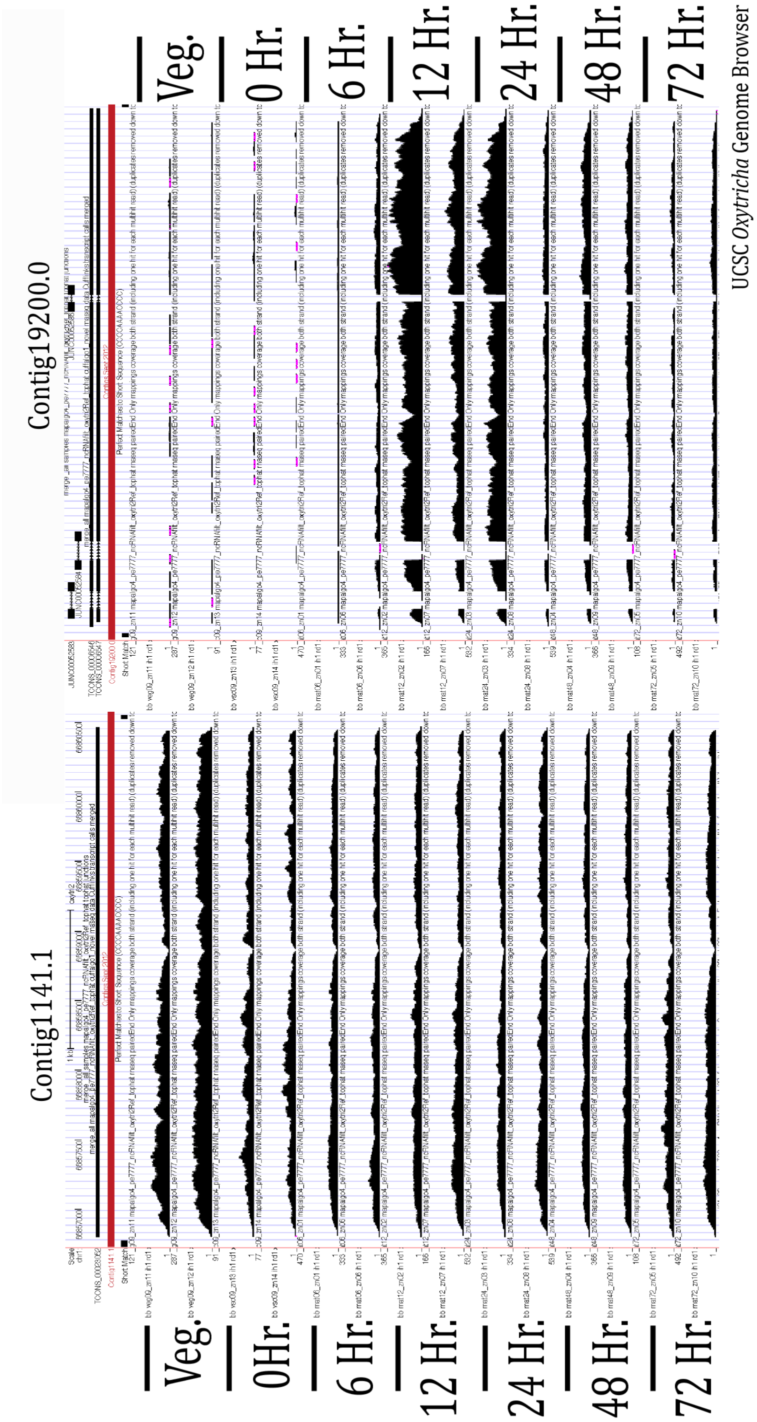


Figure 4: Extensive paralogous gene sets encoding proteins involved in RNA and DNA functions. A. Bar plot representation of the number of genes (black) and orthogroups (purple) from the macronuclear development paralogous gene sets annotated to one of four manually curated gene lists. **B.** Bar plot representation of the number of orthogroups in *Oxytricha* in which at least one gene member is preferentially expressed during macronuclear development. **C.** Bar plot representation of the number of orthogroups in other model organisms sharing at least one gene member with *Oxytricha*.

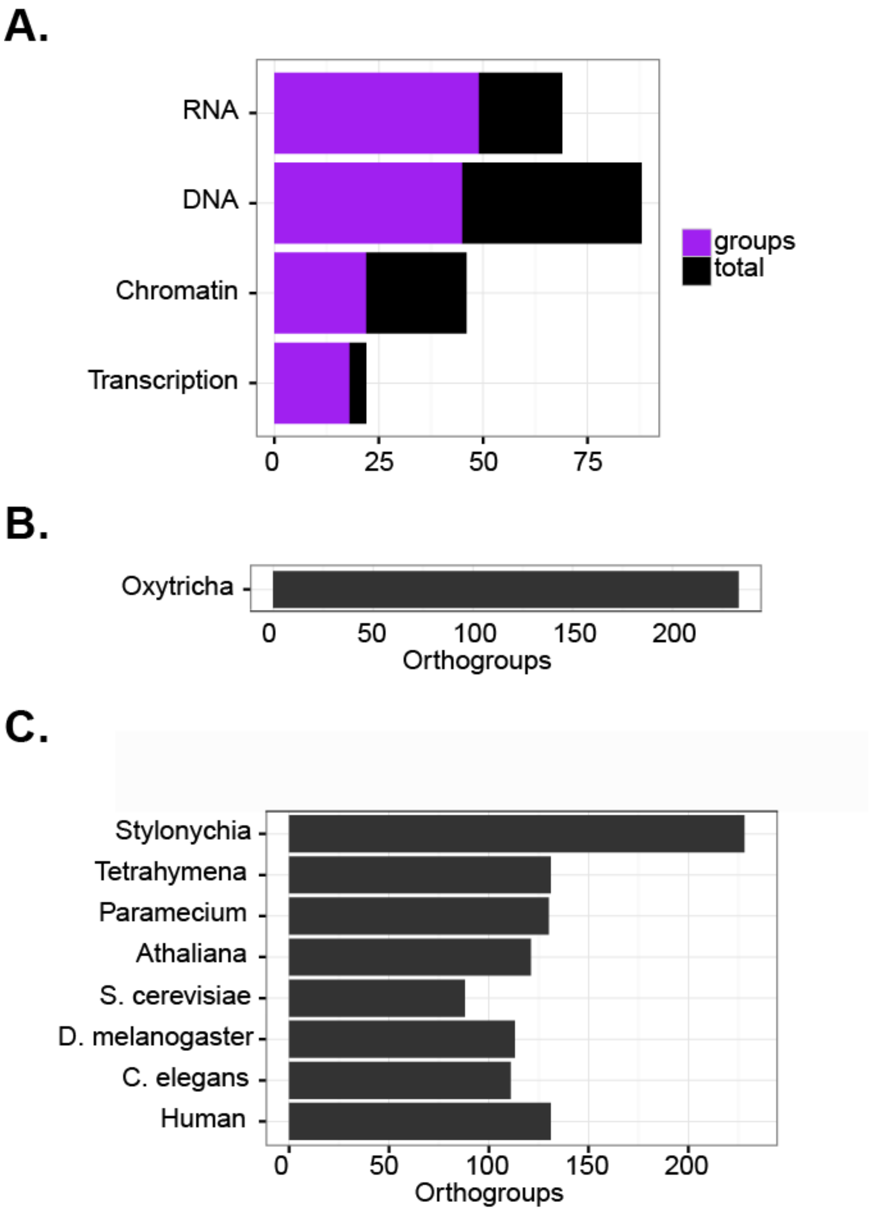


Figure 5: Absolute expression levels of *Oxytricha RPB* paralogs throughout macronuclear development. Bar plots showing absolute expression levels of the constitutively expressed (black bars) versus macronuclear development specific (blue and red bars) paralogs of *Oxytricha RPB*s. RPKM values were calculated for each gene in each sequencing library; the average RPKM value for each gene is shown.

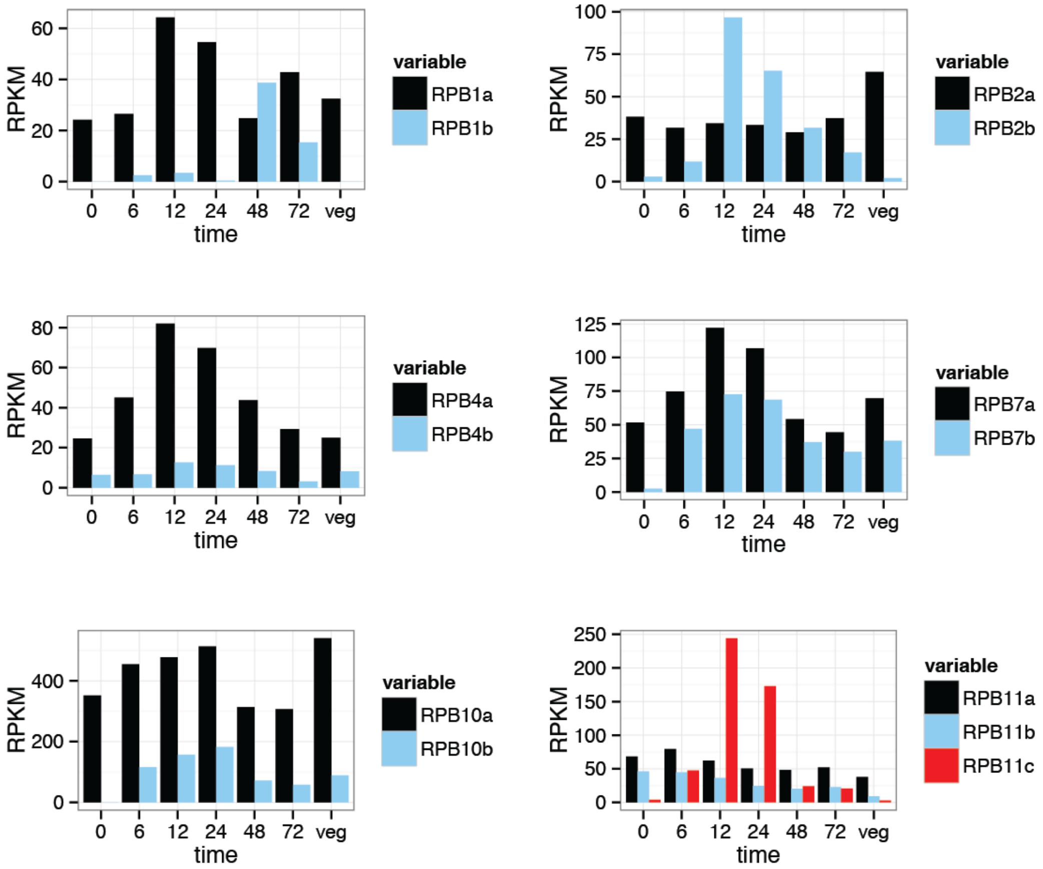


Figure 6: Absolute expression levels of *Oxytricha* RNA polymerase II-associated paralogs throughout macronuclear development. Bar plots showing absolute expression levels of the constitutively expressed (black bars) versus macronuclear development specific (blue and red bars) paralogs of *Oxytricha* RNA polymerase II-associated factors. RPKM values were calculated for each gene in each sequencing library; the average RPKM value for each gene is shown.

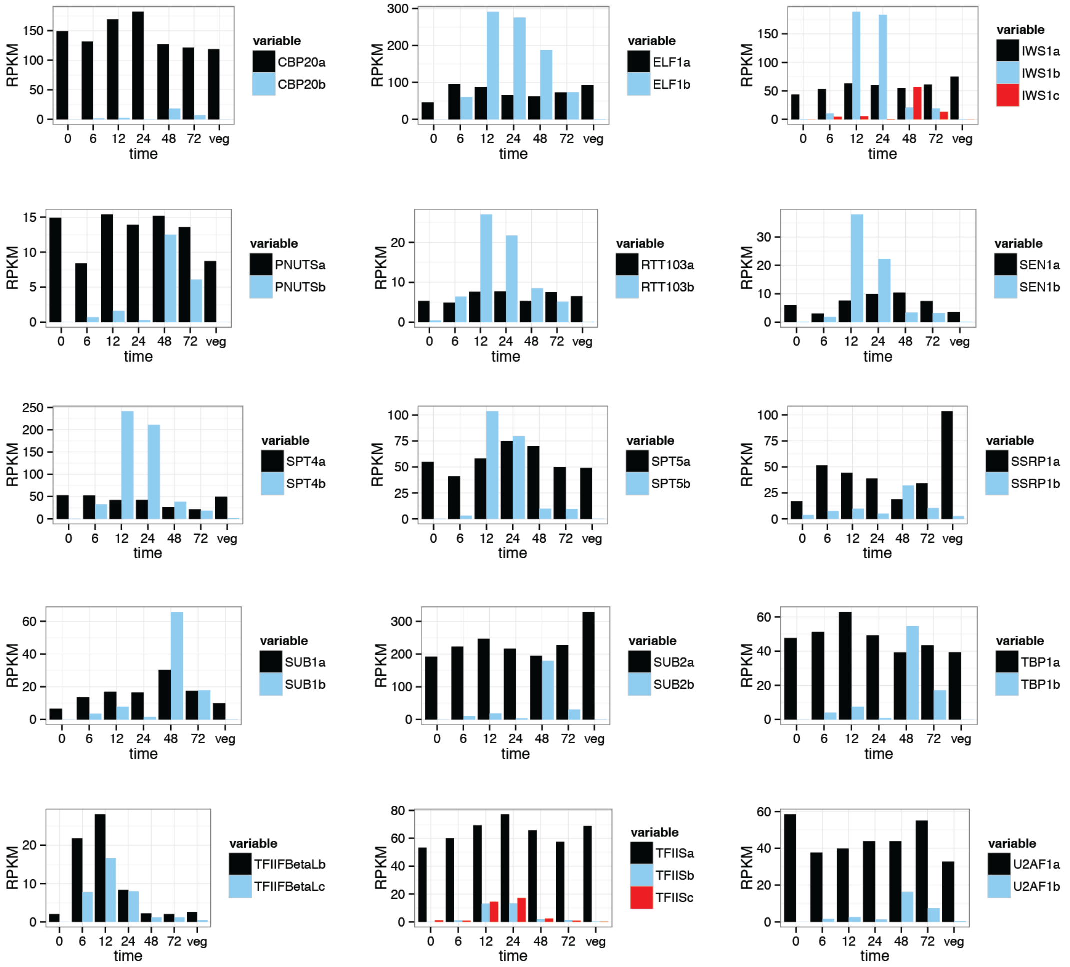


Table 1: Manually curated gene lists expression profiles. The expression patterns of macronuclear development specific RNA, DNA, chromatin and transcription associated genes is shown, along with their associated gene modules.

Gene Set	**Module 1**	**Module 2**	**Module 3**	**Module 4**	**Module 5**	**Total Genes**
RNA	**12**	**39**	**2**	**43**	**11**	**107**
DNA	**5**	**32**	**1**	**97**	**15**	**150**
Chromatin	**2**	**18**	**0**	**34**	**17**	**61**
Transcription	**1**	**13**	**0**	**10**	**1**	**25**

Table 2: *Oxytricha Otiwi* expression profiles. The expression profiles of *Oxytricha Otiwis* throughout macronuclear development is shown, along with their associated gene modules and contig reference numbers.

Otiwi	Contig Number	Expression Pattern	Module
Otiwi1	Contig16116.0	Early MAC Dev.	Module 2
Otiwi2	Contig13836.0	Late MAC Dev.	Module 5
Otiwi3	Contig276.0	Late MAC Dev.	Module 4
Otiwi4	Contig5427.0	Early MAC Dev.	Module 1
Otiwi5	Contig19536.0	Constitutive/Vegetative	N/A
Otiwi6	Contig21631.0	Constitutive/Vegetative	N/A
Otiwi7	Contig16885.0	Early MAC Dev.	Module 1
Otiwi8	Contig20685.0	Constitutive/Vegetative	N/A
Otiwi9	Contig22121.0	Constitutive/Vegetative	N/A
Otiwi10	Contig947.1	Constitutive/Vegetative	N/A
Otiwi11	Contig317.0	Late MAC Dev.	Module 4
Otiwi12	Contig6121.0	Constitutive/Vegetative	N/A
Otiwi13	Contig4847.0	Constitutive/Vegetative	N/A

Table 3: *Oxytricha RPA* expression profiles. The expression profiles of *Oxytricha RPAs* throughout macronuclear development is shown, along with their associated gene modules and contig reference numbers.

RPA	**Contig Number**	**Expression Pattern**	**Module**
RPA1a	Contig5713.0	Constitutive/Vegetative	N/A
RPA1b	Contig2420.0	Constitutive/Vegetative	N/A
RPA1c	Contig9380.0	Constitutive/Vegetative	N/A
RPA1d	Contig9697.0.1	Late MAC Dev.	Module 4
RPA1e	Contig13699.0	Late MAC Dev.	Module 5
RPA1f	Contig14779.0	Late MAC Dev.	Module 5
RPA1g	Contig15857.0.1	Late MAC Dev.	Module 4
RPA1h	Contig16302.0	Late MAC Dev.	Module 4
RPA2a	Contig22204.0	Constitutive/Vegetative	N/A
RPA2b	Contig1943.0	Late MAC Dev.	Module 4
RPA2c	Contig8654.0	Late MAC Dev.	Module 5
RPA2d	Contig11466.0	Late MAC Dev.	Module 5
RPA3	Contig713.1	Constitutive/Vegetative	N/A
RFA1	Contig10831.0	Constitutive/Vegetative	N/A

References

Adl, S.M., and Berger, J.D. (2000). Timing of life cycle morphogenesis in synchronous samples of Sterkiella histriomuscorum. II. The sexual pathway. The Journal of eukaryotic microbiology *47*, 443-449.

Aeschlimann, S.H., Jonsson, F., Postberg, J., Stover, N.A., Petera, R.L., Lipps, H.J., Nowacki, M., and Swart, E.C. (2014). The draft assembly of the radically organized Stylonychia lemnae macronuclear genome. Genome Biol Evol *6*, 1707-1723.

Aronica, L., Bednenko, J., Noto, T., DeSouza, L.V., Siu, K.W., Loidl, J., Pearlman, R.E., Gorovsky, M.A., and Mochizuki, K. (2008). Study of an RNA helicase implicates small RNA-noncoding RNA interactions in programmed DNA elimination in Tetrahymena. Genes Dev *22*, 2228-2241.

Bannon, G.A., Bowen, J.K., Yao, M.C., and Gorovsky, M.A. (1984). Tetrahymena H4 genes: structure, evolution and organization in macro- and micronuclei. Nucleic Acids Res *12*, 1961-1975.

Basu, J., Logarinho, E., Herrmann, S., Bousbaa, H., Li, Z., Chan, G.K., Yen, T.J., Sunkel, C.E., and Goldberg, M.L. (1998). Localization of the Drosophila checkpoint control protein Bub3 to the kinetochore requires Bub1 but not Zw10 or Rod. Chromosoma *107*, 376-385.

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., and Betermier, M. (2009). PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev *23*, 2478-2483.

Baumann, P., and Cech, T.R. (2001). Pot1, the putative telomere end-binding protein in fission yeast and humans. Science *292*, 1171-1175.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289-300.

Bies-Etheve, N., Pontier, D., Lahmy, S., Picart, C., Vega, D., Cooke, R., and Lagrange, T. (2009). RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. EMBO Rep *10*, 649-654.

Bishop, D.K., Park, D., Xu, L., and Kleckner, N. (1992). DMC1: a meiosis-specific yeast homolog of E. coli recA required for recombination, synaptonemal complex formation, and cell cycle progression. Cell *69*, 439-456.

Bouhouche, K., Gout, J.F., Kapusta, A., Betermier, M., and Meyer, E. (2011). Functional specialization of Piwi proteins in Paramecium tetraurelia from post-transcriptional gene silencing to genome remodelling. Nucleic Acids Res *39*, 4249-4264.

Bracht, J.R., Perlman, D.H., and Landweber, L.F. (2012). Cytosine methylation and hydroxymethylation mark DNA for elimination in Oxytricha trifallax. Genome Biol *13*, R99.

Brown, M.S., and Bishop, D.K. (2015). DNA strand exchange and RecA homologs in meiosis. Cold Spring Harb Perspect Biol *7*, a016659.

Bulic, A., Postberg, J., Fischer, A., Jonsson, F., Reuter, G., and Lipps, H.J. (2013). A permissive chromatin structure is adopted prior to site-specific DNA demethylation of developmentally expressed genes involved in macronuclear differentiation. Epigenetics Chromatin *6*, 5.

Cervantes, M.D., Coyne, R.S., Xi, X., and Yao, M.C. (2006). The condensin complex is essential for amitotic segregation of bulk chromosomes, but not nucleoli, in the ciliate Tetrahymena thermophila. Mol Cell Biol *26*, 4690-4700.

Chalker, D.L., and Yao, M.C. (2011). DNA elimination in ciliates: transposon domestication and genome surveillance. Annu Rev Genet *45*, 227-246.

Chang, W.J., Stover, N.A., Addis, V.M., and Landweber, L.F. (2004). A micronuclear locus containing three protein-coding genes remains linked during macronuclear development in the spirotrichous ciliate Holosticha. Protist *155*, 245-255.

Chen, X., Bracht, J.R., Goldman, A.D., Dolzhenko, E., Clay, D.M., Swart, E.C., Perlman, D.H., Doak, T.G., Stuart, A., Amemiya, C.T.*, et al.* (2014). The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell *158*, 1187-1198.

Coverley, D., Kenny, M.K., Lane, D.P., and Wood, R.D. (1992). A role for the human single-stranded DNA binding protein HSSB/RPA in an early stage of nucleotide excision repair. Nucleic Acids Res *20*, 3873-3880.

Coverley, D., Kenny, M.K., Munn, M., Rupp, W.D., Lane, D.P., and Wood, R.D. (1991). Requirement for the replication protein SSB in human DNA excision repair. Nature *349*, 538-541.

Coyne, R.S., Nikiforov, M.A., Smothers, J.F., Allis, C.D., and Yao, M.C. (1999). Parental expression of the chromodomain protein Pdd1p is required for completion of programmed DNA elimination and nuclear differentiation. Mol Cell *4*, 865-872.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol *16*, 157.

Fairman, M.P., and Stillman, B. (1988). Cellular factors required for multiple stages of SV40 DNA replication in vitro. EMBO J *7*, 1211-1218.

Fang, W., Wang, X., Bracht, J.R., Nowacki, M., and Landweber, L.F. (2012). Piwi-interacting RNAs protect DNA against loss during Oxytricha genome rearrangement. Cell *151*, 1243-1255.

Fetzer, C.P., Hogan, D.J., and Lipps, H.J. (2002). A PIWI homolog is one of the proteins expressed exclusively during macronuclear development in the ciliate Stylonychia lemnae. Nucleic Acids Res *30*, 4380-4386.

Firmenich, A.A., Elias-Arnanz, M., and Berg, P. (1995). A novel allele of Saccharomyces cerevisiae RFA1 that is deficient in recombination and repair and suppressible by RAD52. Mol Cell Biol *15*, 1620-1631.

Forcob, S., Bulic, A., Jonsson, F., Lipps, H.J., and Postberg, J. (2014). Differential expression of histone H3 genes and selective association of the variant H3.7 with a specific sequence class in Stylonychia macronuclear development. Epigenetics Chromatin *7*, 4.

Georgaki, A., Strack, B., Podust, V., and Hubscher, U. (1992). DNA unwinding activity of replication protein A. FEBS Lett *308*, 240-244.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat Rev Genet *15*, 829-845.

Greider, C.W., and Blackburn, E.H. (1987). The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell *51*, 887-898.

Greslin, A.F., Prescott, D.M., Oka, Y., Loukin, S.H., and Chappell, J.C. (1989). Reordering of nine exons is necessary to form a functional actin gene in Oxytricha nova. Proc Natl Acad Sci U S A *86*, 6264-6268.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol *52*, 696-704.

Haag, J.R., and Pikaard, C.S. (2011). Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. Nat Rev Mol Cell Biol *12*, 483-492.

Heyer, W.D., Rao, M.R., Erdile, L.F., Kelly, T.J., and Kolodner, R.D. (1990). An essential Saccharomyces cerevisiae single-stranded DNA binding protein is homologous to the large subunit of human RP-A. EMBO J *9*, 2321-2329.

Hirt, H., Manias, D.A., Bryan, E.M., Klein, J.R., Marklund, J.K., Staddon, J.H., Paustian, M.L., Kapur, V., and Dunny, G.M. (2005). Characterization of the pheromone response of the Enterococcus faecalis conjugative plasmid pCF10: complete sequence and comparative analysis of the transcriptional and phenotypic responses of pCF10-containing cells to pheromone induction. J Bacteriol *187*, 1044-1054.

Hogan, D.J., Riordan, D.P., Gerber, A.P., Herschlag, D., and Brown, P.O. (2008). Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol *6*, e255.

Huang, H., Smothers, J.F., Wiley, E.A., and Allis, C.D. (1999). A nonessential HP1-like protein affects starvation-induced assembly of condensed chromatin and gene expression in macronuclei of Tetrahymena thermophila. Mol Cell Biol *19*, 3624-3634.

Jacobs, S.A., Taverna, S.D., Zhang, Y., Briggs, S.D., Li, J., Eissenberg, J.C., Allis, C.D., and Khorasanizadeh, S. (2001). Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. EMBO J *20*, 5232-5241.

Jahn, C.L., and Klobutcher, L.A. (2002). Genome remodeling in ciliated protozoa. Annu Rev Microbiol *56*, 489-520.

Jahn, C.L., Ling, Z., Tebeau, C.M., and Klobutcher, L.A. (1997). An unusual histone H3 specific for early macronuclear development in Euplotes crassus. Proc Natl Acad Sci U S A *94*, 1332-1337.

Jordan, G.E., and Piel, W.H. (2008). PhyloWidget: web-based visualizations for the tree of life. Bioinformatics *24*, 1641-1642.

Kanno, T., Bucher, E., Daxinger, L., Huettel, B., Kreil, D.P., Breinig, F., Lind, M., Schmitt, M.J., Simon, S.A., Gurazada, S.G*., et al.* (2010). RNA-directed DNA methylation and plant development require an IWR1-type transcription factor. EMBO Rep *11*, 65-71.

Khurana, J.S., Wang, X., Chen, X., Perlman, D.H., and Landweber, L.F. (2014). Transcription-independent functions of an RNA polymerase II subunit, Rpb2, during genome rearrangement in the ciliate, Oxytricha trifallax. Genetics *197*, 839-849.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.

Klobutcher, L.A., Jahn, C.L., and Prescott, D.M. (1984). Internal sequences are eliminated from genes during macronuclear development in the ciliated protozoan Oxytricha nova. Cell *36*, 1045-1055.

Kloetzel, J.A. (1970). Compartmentalization of the developing macronucleus following conjugation in stylonychia and euplotes. J Cell Biol *47*, 395-407.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Li, C.F., Pontes, O., El-Shami, M., Henderson, I.R., Bernatavichute, Y.V., Chan, S.W., Lagrange, T., Pikaard, C.S., and Jacobsen, S.E. (2006). An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in Arabidopsis thaliana. Cell *126*, 93-106.

Libault, M., Tessadori, F., Germann, S., Snijder, B., Fransz, P., and Gaudin, V. (2005). The Arabidopsis LHP1 protein is a component of euchromatin. Planta *222*, 910-925.

Lipps, H.J., Sapra, G.R., and Ammermann, D. (1974). The histones of the ciliated protozoan Stylonychia mytilus. Chromosoma *45*, 273-280.

Maldonado, E., Shiekhattar, R., Sheldon, M., Cho, H., Drapkin, R., Rickert, P., Lees, E., Anderson, C.W., Linn, S., and Reinberg, D. (1996). A human RNA polymerase II complex associated with SRB and DNA-repair proteins. Nature *381*, 86-89.

Martin-Tumasz, S., and Brow, D.A. (2015). Saccharomyces cerevisiae Sen1 Helicase Domain Exhibits 5'- to 3'-Helicase Activity with a Preference for Translocation on DNA Rather than RNA. J Biol Chem *290*, 22880-22889.

Matzke, M.A., Kanno, T., and Matzke, A.J. (2015). RNA-Directed DNA Methylation: The Evolution of a Complex Epigenetic Pathway in Flowering Plants. Annu Rev Plant Biol *66*, 243-267.

Matzke, M.A., and Mosher, R.A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat Rev Genet *15*, 394-408.

Mischo, H.E., Gomez-Gonzalez, B., Grzechnik, P., Rondon, A.G., Wei, W., Steinmetz, L., Aguilera, A., and Proudfoot, N.J. (2011). Yeast Sen1 helicase protects the genome from transcription-associated instability. Mol Cell *41*, 21-32.

Mochizuki, K., Fine, N.A., Fujisawa, T., and Gorovsky, M.A. (2002). Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. Cell *110*, 689-699.

Mochizuki, K., and Gorovsky, M.A. (2004). Conjugation-specific small RNAs in Tetrahymena have predicted properties of scan (scn) RNAs involved in genome rearrangement. Genes Dev *18*, 2068-2073.

Moore, S.P., Erdile, L., Kelly, T., and Fishel, R. (1991). The human homologous pairing protein HPP-1 is specifically stimulated by the cognate single-stranded binding protein hRP-A. Proc Natl Acad Sci U S A *88*, 9067-9071.

Murti, K.G. (1973). Electron-microscopic observations on the macronuclear development of Stylonychia mytilus and Tetrahymena pyriformis (Ciliophora-Protozoa). J Cell Sci *13*, 479-509.

Murti, K.G. (1976). Organization of genetic material in the macronucleus of hypotrichous ciliates. Handbook of genetics *5*, 113-137.

Nishimoto, T., Eilen, E., and Basilico, C. (1978). Premature of chromosome condensation in a ts DNA- mutant of BHK cells. Cell *15*, 475-483.

Nowacki, M., Shetty, K., and Landweber, L.F. (2011). RNA-Mediated Epigenetic Programming of Genome Rearrangements. Annu Rev Genomics Hum Genet *12*, 367-389.

Nowacki, M., Vijayan, V., Zhou, Y., Schotanus, K., Doak, T.G., and Landweber, L.F. (2008). RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature *451*, 153-158.

Nowacki, M., Zagorski-Ostoja, W., and Meyer, E. (2005). Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in Paramecium tetraurelia. Curr Biol *15*, 1616-1628.

Petukhova, G.V., Romanienko, P.J., and Camerini-Otero, R.D. (2003). The Hop2 protein has a direct role in promoting interhomolog interactions during mouse meiosis. Dev Cell *5*, 927-936.

Pochart, P., Woltering, D., and Hollingsworth, N.M. (1997). Conserved properties between functionally distinct MutS homologs in yeast. J Biol Chem *272*, 30345-30349.

Ponticelli, A.S., and Smith, G.R. (1989). Meiotic recombination-deficient mutants of Schizosaccharomyces pombe. Genetics *123*, 45-54.

Postberg, J., Heyse, K., Cremer, M., Cremer, T., and Lipps, H.J. (2008). Spatial and temporal plasticity of chromatin during programmed DNA-reorganization in Stylonychia macronuclear development. Epigenetics Chromatin *1*, 3.

Prescott, D.M. (1994). The DNA of ciliated protozoa. Microbiological reviews *58*, 233-267.

Prescott, D.M. (2000). Genome gymnastics: unique modes of DNA evolution and processing in ciliates. Nat Rev Genet *1*, 191-198.

Prescott, D.M., and DuBois, M.L. (1996). Internal eliminated segments (IESs) of Oxytrichidae. The Journal of eukaryotic microbiology *43*, 432-441.

Prescott, D.M., Ehrenfeucht, A., and Rozenberg, G. (2003). Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates. Journal of Theoretical Biology *222*, 323-330.

Prescott, D.M., Murti, K.G., and Bostock, C.J. (1973). Genetic apparatus of Stylonychia sp. Nature *242*, 576, 597-600.

Rowley, M.J., Avrutsky, M.I., Sifuentes, C.J., Pereira, L., and Wierzbicki, A.T. (2011). Independent chromatin binding of ARGONAUTE4 and SPT5L/KTF1 mediates transcriptional gene silencing. PLoS Genet *7*, e1002120.

Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics *20*, 3246-3248.

Semon, M., and Wolfe, K.H. (2007). Consequences of genome duplication. Curr Opin Genet Dev *17*, 505-512.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539.

Skourti-Stathaki, K., and Proudfoot, N.J. (2014). A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. Genes Dev *28*, 1384-1396.

Smith, J., and Rothstein, R. (1995). A mutation in the gene encoding the Saccharomyces cerevisiae single-stranded DNA-binding protein Rfa1 stimulates a RAD52-independent pathway for direct-repeat recombination. Mol Cell Biol *15*, 1632-1641.

Strunnikov, A.V., and Jessberger, R. (1999). Structural maintenance of chromosomes (SMC) proteins: conserved molecular properties for multiple biological functions. Eur J Biochem *263*, 6-13.

Sun, Q., Csorba, T., Skourti-Stathaki, K., Proudfoot, N.J., and Dean, C. (2013). R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus. Science *340*, 619-621.

Swart, E.C., Bracht, J.R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J.S., Goldman, A.D., Nowacki, M., Schotanus, K.*, et al.* (2013). The Oxytricha trifallax macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol *11*, e1001473.

Tang, W., Tu, S., Lee, H.C., Weng, Z., and Mello, C.C. (2016). The RNase PARN-1 Trims piRNA 3' Ends to Promote Transcriptome Surveillance in C. elegans. Cell *164*, 974-984.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol *31*, 46-53.

Treuner, K., Ramsperger, U., and Knippers, R. (1996). Replication protein A induces the unwinding of long double-stranded DNA regions. J Mol Biol *259*, 104-112.

Tucker, S.L., Reece, J., Ream, T.S., and Pikaard, C.S. (2010). Evolutionary history of plant multisubunit RNA polymerases IV and V: subunit origins via genome-wide and segmental gene duplications, retrotransposition, and lineage-specific subfunctionalization. Cold Spring Harbor symposia on quantitative biology *75*, 285-297.

Williams, B.R., Bateman, J.R., Novikov, N.D., and Wu, C.T. (2007). Disruption of topoisomerase II perturbs pairing in drosophila cell culture. Genetics *177*, 31-46.

Wilson, M.H., and Holzbaur, E.L. (2015). Nesprins anchor kinesin-1 motors to the nucleus to drive nuclear distribution in muscle cells. Development *142*, 218-228.

Wold, M.S., and Kelly, T. (1988). Purification and characterization of replication protein A, a cellular protein required for in vitro replication of simian virus 40 DNA. Proc Natl Acad Sci U S A *85*, 2523-2527.

Yamaguchi, S., Decottignies, A., and Nurse, P. (2003). Function of Cdc2p-dependent Bub1p phosphorylation and Bub1p kinase activity in the mitotic and meiotic spindle checkpoint. EMBO J *22*, 1075-1087.

Zahler, A.M., Neeb, Z.T., Lin, A., and Katzman, S. (2012). Mating of the stichotrichous ciliate Oxytricha trifallax induces production of a class of 27 nt small RNAs derived from the parental macronucleus. PLoS One *7*, e42371.

Zahler, A.M., and Prescott, D.M. (1988). Telomere terminal transferase activity in the hypotrichous ciliate Oxytricha nova and a model for replication of the ends of linear DNA molecules. Nucleic Acids Res *16*, 6953-6972.