Using high-throughput sequencing to study genome rearrangement in *Oxytricha*

Kelsi Anne Lindblad

A DISSERTATION PRESENTED TO THE FACULTY OF PRINCETON UNIVERSITY IN CANDIDACY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY THE DEPARTMENT OF QUANTITATIVE AND COMPUTATIONAL BIOLOGY

Advisor: Laura Landweber

April 2017
Abstract

_Oxytricha trifallax_ is a ciliate, a single-celled eukaryote with two genomes. The germline genome is transcriptionally silent throughout most of the cell’s life, and after sexual conjugation 90-95% of the germline sequence is eliminated to form a streamlined somatic genome that will serve the cell’s transcriptional needs. This genome rearrangement process, which involves the deletion of hundreds of thousands of germline-limited sequences, is still poorly understood. In this thesis I use high-throughput sequencing to explore genome rearrangement from three different angles. In Chapter 2 I use RNA-seq to investigate lncRNA copies of somatic chromosomes, which _Oxytricha_ produces during conjugation. I find that these “template” RNAs, which may help the cell arrange somatic segments in the correct order and orientation, are produced throughout development, with different species appearing at different times. This suggests that the cell may rearrange its nascent somatic chromosomes in a particular order, rather than simultaneously. In Chapter 3 I use long-read sequencing to update _Oxytricha_’s somatic genome assembly to a version where more than half of the chromosomes were sequenced without assembly. I find thousands of chromosome isoforms that would be masked by traditional assembly approaches, including some that involve variable recombination between multiple germline loci, a previously unknown source of variation in the somatic genome. Finally, in Chapter 4 I characterize the activity of TBE transposons in the _Oxytricha_ germline. I find these transposons often insert near or within regions of somatic sequence. TBEs therefore continue to shape germline fragmentation. These results demonstrate the power of high-throughput sequencing to tackle complex, genome-scale problems in non-model systems.
Acknowledgements

There are many people who helped me along the path to finishing my PhD, and although I can’t possibly give all of them the thanks they deserve, I’ll do my best to acknowledge as many as I can. First I thank my advisor, Laura, whose support and guidance made all of this possible. Thank you for sharing your scientific insight, your wise advice, and your constant encouragement. And, of course, for all the chocolate. I thank my committee members, Julien and Barbara, for sharing their insight both with regards to my projects and for next steps beyond grad school. In addition I thank Ned Wingreen, who taught some of my favorite grad school courses with amazing patience and generosity; and Josh Shaevitz for a terrific lab rotation and for helping me out of more than one sticky situation. I am also hugely indebted to Jen Brick, who has been unflaggingly kind and helpful from the day I first came to interview at Princeton.

My lab members have also been a great source of inspiration and comradery during my time at Princeton. Their encouragement, feedback, and commiseration during difficult times were greatly appreciated. It’s been a pleasure working with you all. In particular I thank Leslie Beh, who worked together with me on several projects both in this thesis and not and whose good sense of humor and serious scientific chops made all of those collaborations a pleasure. I would also thank Jonathan Burns for his hospitality whenever I visited his group at USF and his patience in translating the theoretical parts of his group’s work for me. Finally, I thank former lab member John Bracht, who was my rotation project mentor and whose infectious enthusiasm and willingness to try off-the-wall ideas helped get me interested in working with Oxytricha in the first place.

I also thank my fellow QCB students. I was privileged to work with such a great group of friends and scientists, and as much as I enjoyed hearing about the diverse work they were doing
here at Princeton, I’m even more excited to see where they go next. In particular I thank Ari Strandburg-Peshkin, my friend and roommate, for serious science discussions as well as simply hanging out, and for putting up with my questionable dish-washing habits. May all your data be clean and your study subjects cooperative. Dave Robinson fielded far too many of my stats questions without asking for anything in return except perhaps an audience for a good programming language rant; many thanks.

My non-QCB friends have been immensely helpful as well, if not always so much with science, then with life, with art, with everything else. Alex, Blair, Hlín, Kyra, Meredith, and all the rest of you—thanks.

I thank my family for their love and support, both during grad school and throughout my entire life. I never would have gotten here without them. To my mom and my brother, for all the great conversations we’ve had and for always being willing to lend a listening ear, and to my dad, for always being ready to talk science and whose example sent me down this path in the first place, thank you. I love you all.

Finally, I am grateful to the National Institutes of Health for providing financial support for my studies.
# Contents

Abstract ......................................................................................................................... iii
Acknowledgements ........................................................................................................ iv
Table of Contents .......................................................................................................... vi
List of Tables ................................................................................................................ ix
List of Figures .............................................................................................................. x

1. Introduction .............................................................................................................. 1

2. Thousands of RNA-cached copies of whole chromosomes are present in the ciliate Oxytricha during development .................................................. 5
   2.1 Abstract ........................................................................................................... 5
   2.2 Background .................................................................................................... 6
   2.3 Results .......................................................................................................... 11
      2.3.1 Oxytricha produces thousands of full-length RNA copies of somatic chromosomes .............................................................. 11
      2.3.2 The absence of some templates suggests that not all IncRNAs contain both telomeres ..................................................... 14
      2.3.3 Heterogeneity in template abundance during nuclear development .............................................................................. 16
   2.4 Discussion ....................................................................................................... 16
   2.5 Methods ......................................................................................................... 17
      2.5.1 Cell Culture and Harvesting ................................................................. 17
      2.5.2 RNA Isolation and Template Amplification ....................................... 17
      2.5.3 Library Preparation and High-Throughput Sequencing ................ 23
      2.5.4 RNA-seq Analysis .............................................................................. 23
      2.5.5 RT-PCR Validation ............................................................................. 24
      2.5.6 Motif Finding ......................................................................................... 25
      2.5.7 Statistical Analysis .............................................................................. 25
3. Capture of complete chromosomes in single sequencing reads reveals widespread structural variation within a genome

3.1 Abstract ................................................................. 26
3.2 Background ........................................................... 27
3.3 Results ................................................................. 29

3.3.1 Over half the Oxytricha somatic genome can be completely sequenced without assembly ............... 29
3.3.2 Long-read sequencing discovers novel chromosome isoforms ......................................................... 26
3.3.3 Hybrid error correction produces the most complete somatic genome assembly ............................... 35

3.4 Discussion .............................................................. 36
3.5 Methods ................................................................. 37

3.5.1 Cell Growth and Culture ......................................... 37
3.5.2 RNA Isolation and Template Amplification ............ 37
3.5.3 Genome Assembly .................................................. 38
3.5.4 Analysis of Alternative Chromosome Fragmentation .... 40
3.5.5 RNA-Seq .............................................................. 41
3.5.6 Gene Prediction ....................................................... 42
3.5.7 Subsampling Analysis ............................................. 42
3.5.8 Statistical Analysis ................................................ 43

4. Active transposons contribute to genome fragmentation in Oxytricha

4.1 Abstract ................................................................. 44
4.2 Background ........................................................... 45
4.3 Results ................................................................. 47

4.3.1 The Oxytricha germline contains active TBE transposons ............................................................... 47
4.3.2 TBE transposition events produce strain-specific germline gene architectures .................................. 51
4.4 Discussion .............................................................. 52
4.5 Methods ................................................................. 54
4.5.1 Sequencing and Assembly of the *Oxytricha trifallax* JRB510 germline genome ........................................ 54
4.5.2 Identification of Strain-Specific TBE Insertions ............. 55
4.5.3 Analysis of TBE Distribution and Phylogeny .................. 55

Bibliography .......................................................... 56
List of Tables

3.1 Assembly statistics for long- and short-read genome assemblies ................. 31
4.1 Comparison of the JRB310 and JRB510 germline genome assemblies ............... 47
List of Figures

2.1 The template model of genome rearrangement .................................................. 9
2.2 Strategy for genome-wide survey of template IncRNAs .................................. 10
2.3 Template-seq reads cover entire chromosomes .............................................. 11
2.4 IncRNA levels show little correlation with the abundance of other RNAs .......... 13
2.5 Motifs enriched in chromosomes without IncRNAs ...................................... 14
2.6 RT-PCR confirms the presence of template RNAs not detected in RNA-seq .... 15
2.7 Different templates are present at different developmental stages ................. 16
2.8 Subsampling analysis suggests that the IncRNA population is well-sampled ..... 18
2.9 DNA contamination is low in IncRNA data .................................................. 22
3.1 SMRT sequencing reads are long enough to capture complete Oxytricha chromosomes ............................................................................................................. 29
3.2 Long-read sequencing reveals underlying structural variation among chromosomes ........................................................................................................... 32
3.3 Hybrid error correction outperforms long read self-correction ....................... 34
4.1 New transposon insertions reflect background frequencies ............................. 49
4.2 Some TBE families show expansion in JRB510 .......................................... 50
4.3 Transposon insertions occur near somatic-destined segments ....................... 51
4.4 Novel TBE inserts lead to strain-specific fragmentation patterns ................... 52
Chapter 1

Introduction

Genome editing and remodeling are important processes in many eukaryotes, including vertebrates, which rely on V(D)J recombination to generate diversity in their immune systems; many nematodes, insects, agnathans, and marsupials that undergo programmed chromosome diminution; and cancers characterized by aberrant chromosome breakage and fusion. One of the most convenient systems for studying large-scale genome rearrangements are ciliates, a group of deeply diverged microbial eukaryotes found everywhere from temperate soils to the Arctic Ocean. Despite the diversity of their lifestyles and environments, all ciliates share a dual-genome architecture, with a transcriptionally-silent germline (micronuclear) genome and a streamlined somatic (macronuclear) genome responsible for the majority of transcription during asexual growth and reproduction (Prescott 1994). After sexual conjugation ciliates undergo an extensive process of genome rearrangement that involves deleting regions of germline-limited DNA (IESs) that interrupt sequences retained in the mature soma (MDSs).

*Oxytricha trifallax* demonstrates a particularly extreme example of the process; between 90 and 95% of its macronuclear genome is eliminated during rearrangement. Further, the order of the MDSs in its micronuclear genome do not always reflect their order or orientation in the macronuclear sequence—some MDSs must be reversed or reordered in order to produce mature macronuclear chromosomes. These somatic chromosomes are highly fragmented, consisting of tens of thousands of “nanochromosomes” averaging 3.2kb long (Swart et al. 2013). Each tiny chromosome usually bears only a single gene with very little noncoding sequence. Somatic
chromosomes are variably amplified, with up to tens of thousands of copies in a single macronucleus (Prescott 1994).

The mechanisms behind the genome rearrangements necessary to produce such a highly fragmented, streamlined genome remain poorly understood. *Oxytricha* uses Piwi-associated RNAs to mark segments of the micronuclear sequence that will be retained in the macronucleus (Fang et al. 2012; Zahler et al. 2012). Long “template” RNAs, complete copies of somatic chromosomes, may then help the cell arrange the retained segments in the correct order (Nowacki et al. 2008). In the ciliates *Tetrahymena* and *Paramecium* a domesticated transposase is necessary for the actual excision of IESs (Boudry et al. 2009; Cheng et al. 2010), but although knocking down TBE transposons causes genome rearrangement defects in *Oxytricha* (Nowacki et al. 2009), the role of transposases in this organism’s genome editing process remains unclear. Overall, genome rearrangement in ciliates is a complex process requiring the interaction between several different regulatory systems, no one of which we fully understand.

In the past, studies of ciliate genome rearrangement were constrained to one or a handful of loci that could be individually sequenced and examined. However, recent advances in high-throughput sequencing technology are beginning to allow whole-genome analyses of *Oxytricha’s* MDSs, IESs, and rearrangement process. Both macronuclear (Swart et al. 2013) and micronuclear (Chen et al. 2014) genome assemblies have been published for *Oxytricha*, and I leverage them throughout this thesis to address questions that would be impossible to approach without the benefit of high-throughput sequencing.

In Chapter 2 I analyze an RNA-seq survey of “template” IncRNAs abundant during genome rearrangement. These RNAs contain the complete sequence of somatic nanochromosomes, telomere to telomere, and are believed to help guide rearrangements by
providing a correct “template” for MDSs to align with. While previous experiments (Nowacki et al. 2008) detected template RNAs for a few chromosomes, this chapter describes the first attempt to characterize them genome-wide. I analyzed RNA-seq data from a developmental time course and found that different groups of template RNAs are present at different times during development. They are often transient, suggesting that they are transcribed throughout development, rather than in a single burst as previously thought. This may indicate that macronuclear chromosomes are rearranged in waves rather than all at once.

The short length of *Oxytricha*’s macronuclear chromosomes makes it possible for long-read sequencing technologies to capture their entire sequences in single reads. In Chapter 3 I use Single-Molecule Real Time (SMRT) sequencing to produce an update of the *Oxytricha* macronuclear genome assembly where over half of the chromosomes are sequenced in individual reads, requiring no assembly. This allowed me to capture many structural isoforms and polymorphic chromosomes that would be masked by traditional assembly approaches. I found that some chromosomes may occur in multiple forms, composed of different sets of paralogous MDSs that occur up to hundreds of thousands of base pairs away from one another in the micronuclear genome. This flexibility in selecting MDSs for the final somatic sequence represents a previously unknown source of variation in the somatic genome, which may increase the cell’s ability to adapt and evolve in response to novel environmental challenges.

In Chapter 4 I examine the role of a particular class of transposable elements, TBE transposons, in IES formation in the *Oxytricha* micronuclear genome. The examination of a few micronuclear loci led to early hypotheses that IESs are ancient remnants of transposon invasions (reviewed in Klobutcher and Herrick 1997), but it has only recently become possible to characterize the full set of IESs in a genome. In this chapter I produce a micronuclear assembly
for a strain of *Oxytricha* closely related to the published JRB310 (Chen et al. 2014) and compare the complement of TBE transposons in each. I find that TBEs remain active in the *Oxytricha* germline, and that they preferentially insert near MDSs. Many strain-specific TBE insertions actually occur inside MDSs, creating new IESs and resulting in strain-specific fragmentation patterns. Thus, transposons continue to shape *Oxytricha*’s germline architecture and may be major drivers of somatic gene fragmentation in the micronucleus.

This thesis provides an initial look at several different aspects of genome rearrangement in *Oxytricha*, but many questions remain. The discovery of different populations of template RNAs at different time points during development raises the question of what determines when a chromosome’s template is produced and what function staggering template production, and possibly chromosome rearrangements, may have. The overall role of template RNAs in genome rearrangement is also unknown—do they play any roles besides guiding MDS ordering, perhaps as precursors to the piRNAs that mark segments for retention? It also remains a mystery how the cell brings together MDSs over great genomic distances, and the variable MDS arrangements I observed in my examination of SMRT-generated nanochromosome sequences also raises the question of how the complement of MDSs used in a given chromosome is even chosen. Further, transposable elements outside of TBEs remain largely unexplored in *Oxytricha*, on top of the fact that almost nothing is known about the few transposon-derived genes in the macronuclear genome (Swart et al. 2013). Given the importance of transposases in genome rearrangement for other ciliate systems, this line of inquiry is a particularly tempting target for future study. I hope the work I present in this thesis will provide a solid foundation for future investigation into the lifestyle of this fascinating organism.
Chapter 2

Thousands of RNA-cached copies of whole chromosomes are present in the ciliate Oxytricha during development

2.1 Abstract

The ciliate Oxytricha trifallax maintains two genomes: a germline genome that is active only during sexual conjugation and a transcriptionally active, somatic genome that derives from the germline via extensive sequence rearrangement. Previously, we found that long, non-coding (Inc) RNA “templates”—telomere-containing, RNA-cached copies of mature chromosomes—provide the information to program the rearrangement process. Here we used a modified RNA-seq approach to conduct a genome-wide search for endogenous, telomere-to-telomere RNA transcripts. We find that during development Oxytricha produces long, noncoding RNA copies for over 10,000 of its 16,000 somatic chromosomes, consistent with a model in which Oxytricha transmits an RNA-cached copy of its somatic genome to the sexual progeny. Both the primary sequence and expression profile of a somatic chromosome influence the temporal distribution and abundance of individual template RNAs. This suggests that Oxytricha may undergo multiple rounds of DNA rearrangement during development. These observations implicate thousands of long RNA molecules in the wiring and maintenance of an elaborate somatic genome architecture.

1 Adapted from Lindblad KA*, Bracht JR*, Williams AE, Landweber LF. (In Review.) Thousands of RNA-cached copies of whole chromosomes are present in the ciliate Oxytricha during development. * = equal contribution
2.2 Background

Long non-coding RNAs (IncRNAs), defined as transcripts >200nt with no protein coding function, were once thought to represent primarily nonfunctional “junk” transcription. However, the discovery of ~10,000 IncRNA loci in the human genome (Derrien et al. 2012) and evidence that IncRNAs play active roles in processes as diverse as chromatin remodeling (Gupta et al. 2010), transcriptional interference (Latos et al. 2012), and post-transcriptional modification (Yoon et al. 2012), suggests that they have important roles in biological systems that modern techniques are finally making amenable to study. Originally controversial, the idea of an RNA cache was proposed as a means of epigenetic transmission of sequence information across generations (Lolle et al. 2005). Our laboratory experimentally demonstrated that epigenetically-inherited IncRNAs are essential for genome remodeling in the ciliate Oxytricha trifallax (Nowacki et al. 2008). Here, we present evidence that Oxytricha produces complete RNA copies of thousands of its somatic chromosomes during nuclear differentiation and development.

Like all ciliates, Oxytricha is a microbial eukaryote with two kinds of nuclei per cell. The smaller micronucleus (MIC) contains the germline, which provides haploid gametic nuclei for sexual conjugation. The larger macronucleus (MAC) contains the somatic genome, which is the source of gene transcription during asexual growth and reproduction. While the micronuclear genome consists of long diploid chromosomes, the MAC genome contains over 16,000 different chromosomes, most of which bear only a single gene and are a median length of 2,515 bp (mean 3.2 kb) at a typical copy number of ~1,900n (Prescott 1994).

After sexual exchange, the exconjugant daughter cell produces a new MAC from a copy of the MIC through a series of dramatic genome rearrangements (reviewed in Yerlici and Landweber 2014). This process eliminates 90-95% of the MIC genome, including all satellite
repeats, transposable elements, and germline-exclusive genes, as well as internally eliminated sequences (IESs) that interrupt the precursor gene segments in the MAC. These retained DNA regions are called the macronuclear-destined segments (MDSs). Short direct repeat sequences, called *pointers*, border consecutive MDS-IES junctions and may help guide the MDS joining events that build the somatic chromosomes. Approximately 20% of *Oxytricha*'s genes are “scrambled,” containing at least one MDS that is permuted or inverted in the MIC genome, relative to its location in the MAC (Chen et al. 2014). Because they have a different order or orientation in the MIC than in the MAC, they must rearrange before they recombine to form mature MAC chromosomes.

The process of genome remodeling in *Oxytricha* and other ciliates is guided by an RNA-based system of epigenetic inheritance. While the distantly related ciliates *Paramecium* and *Tetrahymena* use PIWI-associated “scnRNAs” to mark regions of the MIC for elimination (Lepère et al. 2008, Mochizuki et al. 2002) and other small RNAs that may facilitate IES removal (Sandoval et al. 2014, Noto et al. 2015), *Oxytricha* uses PIWI-associated RNAs (piRNAs) to mark segments of the MIC genome for retention rather than elimination (Fang et al. 2012; Zahler et al. 2012). These 27nt piRNAs derive from both strands and show peak expression between 18 and 24 hours after the beginning of conjugation (Fang et al. 2012; Zahler et al. 2012). Functional experiments demonstrated that injection of piRNAs targeting MIC-limited sequences can program retention of those sequences in the new MAC, and that the DNA sequence retention extends across sexual generations (Fang et al. 2012); however, piRNAs have not been demonstrated to program MDS joining or to be capable of substitution transfer to the rearranging molecule (Nowacki et al. 2008). Hence, the piRNAs appear to be incapable of programming DNA rearrangement (Fang et al. 2012).
Long RNA copies of MAC chromosomes, on the other hand, can (re)program chromosomal rearrangements (Nowacki et al. 2008). *Oxytricha* produces these long RNAs during a burst of genome-wide bidirectional transcription early in cell development (Khurana et al. 2014); RT-PCR detects long, telomere-containing copies of whole chromosomes between 5 and 30 hours after conjugation (Nowacki et al. 2008). Injecting synthetic RNA or DNA versions of a chromosome with incorrectly-ordered MDSs leads conjugating cells to produce progeny whose chromosomes follow the aberrant ordering in their new MAC, and this effect persists across multiple sexual generations (Nowacki et al. 2008). Point mutations introduced by an injected RNA copy of a chromosome (Nowacki et al. 2008) and, in the related ciliate *Stylonychia lemnæ*, substitutions in the telomeric sequence, can be passed to chromosomal telomeres and persist through multiple asexual generations (Fuhrmann et al. 2016). This suggests the long RNAs produced during conjugation may act as templates, as originally proposed in (Prescott et al. 2003 and Angeleska et al. 2007), to guide rearrangement of genome segments during nuclear development (Fig. 2.1).

Here we report the first global survey of template RNAs in *Oxytricha trifallax* and the first such survey of telomere-to-telomere RNA transcripts in any eukaryote. We detect the presence of over 9,000 different lncRNAs corresponding to complete MAC chromosomes. These findings support the template RNA-guided model of DNA rearrangement and underscore the importance of long RNAs in the programming and maintenance of *Oxytricha*’s genome architecture.
Figure 2.1: The template model of genome rearrangement. As the parental MAC degrades, A) bi-directional transcription of the parental chromosomes produces lncRNA copies. In the developing MAC, B) Otiwi1 (blue circles) complexes with 27nt piRNAs to mark MDSs (numbered) for retention. C) The lncRNA templates are transported to the developing MAC, where they guide the correct configuration. D) DNA breaks permit recombination between MDSs, with concomitant loss of MIC-limited DNA (light gray). Telomeres (black bars) cap sequence ends to form a mature MAC chromosome.
Figure 2.2: Strategy for genome-wide survey of template lncRNAs. Whole-chromosome RNA copies of somatic chromosomes were selectively amplified using telomeric primers and reverse transcribed into DNA. Contaminating hexamer repeat sequences were digested with frequent-cutting restriction enzymes, and the remaining sequences sheared and libraries prepared for Illumina sequencing.
2.3 Results

![Graph A](image1.png) ![Graph B](image2.png)

**Figure 2.3: Template-seq reads cover entire chromosomes.** While the aggregate of average read depth across all chromosomes (A) varies along the span of a chromosome, all portions are well covered, including noncoding, sub-telomeric regions. There is no significant difference in the proportions of bases belonging to different sequence categories (B) between RNA-seq reads in this study and the genomic background, which indicates that template RNAs are noncoding, containing both introns and intergenic DNA, and that they can cover entire chromosomes from telomere to telomere.

2.3.1 *Oxytricha* produces thousands of full-length RNA copies of somatic chromosomes.

We used a novel PCR-based procedure (Fig. 2.2) to globally amplify RNA molecules containing telomeric repeats at both ends across a developmental time course. The six time points include: zero hours post-mixing (0h), after compatible *Oxytricha* strains are combined but before mating begins; six hours post-mixing, shortly after the first putative template RNAs were detected in Nowacki et al. (2008); twelve hours post-mixing, which previous studies suggested might be the peak of template RNA production; and forty-eight and sixty hours post-mixing, when few or no templates were previously observed. We sequenced the amplified RNAs using Illumina paired-end sequencing and mapped the resulting 100bp reads to a subset of the
*Oxytricha* MAC genome containing high-confidence chromosomes short enough to be amplified by the PCR step in the sequencing pipeline. Ultimately, we recovered RNA-seq read pairs corresponding to 10,507 different chromosomes, representing more than 2/3 of all completely assembled chromosomes in the somatic genome. This includes RNA-seq reads detected from 3,230 chromosomes as early as the time of cell mixing (0h), much earlier than expected. This set is enriched in chromosomes with a high level of expression during development, and transcription might occasionally begin at or within either telomere, where RNA polymerase localizes (Khurana et al. 2014), and extend to the other telomere, since subtelomeric regions are so short in *Oxytricha* (often < 50 bp; Swart et al. 2013). Thus, in these cases, some reads could actually derive from mRNAs for genes with developmental expression. Overall, our RNA-seq data show little correlation with mRNA data collected over a similar developmental time course (Swart et al. 2013). In addition, our RNA-seq reads map across whole chromosomes, including introns and sub-telomeric regions, with no enrichment for coding regions (Fig. 2.3), suggesting that mRNA contamination is not a major concern in our dataset. Furthermore, fewer than 2% of reads map to germline-limited sequence at all time points. (The only exception is the presence in the 18 and 36 hour time points of a germline-limited repeat that contains two telomere-like sequences and thus is captured by the sequencing pipeline). This suggests that our experimental approach successfully captured non-coding RNAs that, as predicted, span entire chromosomes contain telomeric sequences at both ends, and derive from the mature somatic genome rather than the germline. In addition, we found no correlation between lncRNA levels and piRNA levels at any time point in development (Fig. 2.4), suggesting that the molecules we recovered are a functionally distinct class of RNA, rather than merely piRNA precursors, but they could serve dual roles.
Figure 2.4: lncRNA levels show little correlation with the abundance of other RNAs. (A) The weak relationship between lncRNA levels and mRNA levels (Spearman's $\rho = 0.181 - 0.281$) across development indicates that our method captures noncoding RNA rather than mRNA and that template production is likely independent of normal gene transcription. (B) There is negligible correlation between lncRNA counts and piRNA counts at any point in development (Spearman's $\rho = -0.009 - 0.088$), which suggests that the two classes of RNA are largely independent, although it is possible that a subset of the longer RNAs are precursors to the piRNAs.
Figure 2.5: Motifs enriched in chromosomes without lncRNAs. Sequence logos indicate motifs significantly enriched in either the upstream or downstream noncoding regions of chromosomes for which our survey found no corresponding lncRNAs. Vertical bars indicate the median position of each motif within the noncoding region. If these motifs cause RNA polymerase to terminate transcription before reaching the end of a chromosome, it would result in templates that lack telomeres on both ends. Our RNAseq pipeline would not detect such RNAs unless they contain an internal telomeric sequence.

2.3.2 The absence of some templates suggests that not all lncRNAs contain both telomeres.

To investigate why template RNAs were absent for a subset of *Oxytricha*’s MAC chromosomes, we searched for motifs associated with those chromosomes. This identified one motif enriched in the 5' noncoding regions of nanochromosomes that had no identified templates, and a second motif enriched in the 3' noncoding region of those same chromosomes (p < 2.2e-16 for both motifs) (Fig. 2.5). Overall 48.1% of chromosomes with no mapped lncRNA reads had at least one instance of one of the motifs, versus 37.0% of contigs with mapped lncRNA reads. Contigs without lncRNA data also have significantly more copies of the motifs per chromosome than those with lncRNA reads (one-sided Welch’s t = 6.646, df = 3487.63, p = 1.738e-11).

We propose that these motifs disrupt transcription of the template RNA and prevent RNA polymerase II (Khurana et al. 2014) from reaching the far telomere. The AT-rich 5' motif, in particular, could mimic a transcription termination signal. Since, in principle, a template RNA only needs to span all of a chromosome's MDS junctions to guide DNA rearrangements, the
presence of both telomeres may not be a strict requirement for function. However, because our survey only captured RNA molecules that contain telomeric sequences on both ends (see Methods), such prematurely-terminated templates would not appear in our dataset.

Furthermore, experiments using RT-PCR and gene-specific primers recovered sense, antisense, or both strands corresponding to lncRNA transcripts with a telomere sequence at one end for five (out of five) MAC chromosomes whose template RNAs were absent from our RNA-seq survey (two examples shown in Fig. 2.6). This evidence of RNA templates that our current methods were unable to detect is consistent with the proposal that the cell produces lncRNA transcripts for all of its chromosomes, and that these RNAs may include the presence of one or both telomeres.

Figure 2.6: RT-PCR confirms the presence of template RNAs not detected in RNA-seq. On the left are two representative chromosomes that had no RNA-seq reads but were within the size range of our lncRNA sequencing method. On the right is a positive control amplification from a chromosome that did have RNA-seq reads. The combination of a gene-specific primer on one side plus a generic primer on the telomere side produces some background, as expected, but bands of the appropriate size are recovered in both forward (For.) and reverse (Rev.) reactions (marked with a red arrowhead). This suggests that the cell does produce template RNAs for these chromosomes, although it is possible that some only contain one telomeric sequence.
2.3.3 Heterogeneity in template abundance during nuclear development

We find that different populations of lncRNA templates with telomeric sequences at both ends are present at different developmental stages. While the greatest absolute number and diversity of templates appears twelve hours post-cell mixing, thousands of templates appear only in samples from later time points. In addition, hundreds of templates were only detected at a single time point (Fig. 2.7), suggesting the possibility of rapid lncRNA production followed by swift degradation. While the telomere-to-telomere transcripts we surveyed are individually present at low abundance, resampling simulations indicate that our read depth was high enough such that we are not missing many lncRNAs as a result of stochastic loss (Fig. 2.8).

This variation in individual template RNA levels across development suggests that
chromosome rearrangement may occur in waves, rather than all at once. Chromosomes whose templates appear in early development (at zero or six hours post mixing of mating types) have significantly higher RNA-seq expression levels during genome rearrangement (one-sided Welch’s $t = 3.450$, df = 5181.362, $p = 0.00028$) than those whose templates were only detected later in development. Chromosomes whose gene products are required early might assemble first, especially if the cell relies on zygotic transcription from the developing macronucleus while the parental nucleus degrades. On the other hand, chromosomes that bear genes whose products are not required until vegetative growth, i.e. until nuclear differentiation is complete, are less critical and can be rearranged later during development.

In addition, some features of a chromosome’s rearrangement map appear to correlate with when and how long its RNA templates are present. Chromosomes whose templates appear in only a single time point have a significantly greater number of pointers (the short direct repeats in the germline where somatic segments recombine) than those present at all time points (one-sided Welch’s $t = 2.183$, df = 845.314, $p = 0.01465$), although they have fewer scrambled pointers, which require reordering during rearrangement, relative to chromosomes with templates found at all time points surveyed (one-sided Welch’s $t = -5.0907$, df = 1842.702, $p = 1.966e-7$). Those chromosomes whose templates are absent from our RNA-seq dataset have a greater number of MDSs than those that are present (one-sided Welch’s $t = -33.718$, df = 16281.52, $p < 2.2e-16$) as well as more scrambled pointers (one-sided Welch’s $t = -3.645$, df = 12449.62, $p = 0.00013$). Conversely, chromosomes with only a single MDS, which require no rearrangement or IES elimination but do require telomere addition, appear at significantly more time points than those with more than one MDS (one-sided Welch’s $t = 2.691$, df = 10.086, $p = 0.011$). Overall,
chromosomes that require the greatest number of MDS joining or rearrangement events have the most transiently available template RNAs.

**Figure 2.8: Subsampling analysis suggests that the IncRNA population is well-sampled.** The number of chromosomes with corresponding IncRNA data saturates rapidly at all time points, with 60% of the data collected being sufficient to recover 90% of the unique IncRNAs at all time points. This suggests that additional read depth would not substantially increase the number of IncRNAs identified in this study, and that chromosomes with no IncRNAs identified in this study either are not transcribed into IncRNAs or were missed due to experimental limitations, rather than missed due to low abundance and stochastic loss.
2.4 Discussion

*Oxytricha* produces RNA copies of thousands of its somatic chromosomes during macronuclear development. The levels of these template RNAs fluctuate during development, and this temporal heterogeneity suggests that not all chromosomes undergo DNA rearrangement at the same time. Thus the corresponding need for template RNAs would vary for individual chromosomes.

Consistent with the findings of Khurana et al. (2014), both the absolute number of template RNAs and the number of distinct template RNA sequences peak twelve hours after cell mixing, when RNA polymerase is found poised near both ends of the macronuclear chromosomes. However, we find some templates present earlier in the developmental cascade, even at the time of cell mixing. The early assembly of these chromosomes may be important for production of the corresponding gene products during development. Given that DNA replication is central to genome rearrangement in most ciliates (Ammermann et al. 1974), we hypothesize that the observed variations in genome rearrangement timing may reflect underlying variations in DNA replication timing, as reported in human (Koren et al., 2014).

If the cell produces template RNAs serially rather than through a single burst of transcription, then the time at which a chromosome’s templates become available may influence the order of DNA deletion and descrambling events or template-guided DNA repair (Nowacki et al. 2008). Similarly, Möllenbeck et al. (2008) observed that chromosomes go through distinct stages of rearrangement, with simple DNA deletions often occurring before translocations during the process of DNA rearrangement. Furthermore, the data in Möllenbeck et al. (2008) are consistent with the possibility that some rearrangements might occur before RNA templates are abundant or even available, because the earlier DNA deletions were accompanied by higher
levels of error at rearrangement junctions, which RNA template-guided DNA repair (Nowacki et al. 2008) may later restore. While there were no RNA studies in Möllenbeck et al. (2008), it suggested a temporal component to DNA processing, whereas the current study reveals a temporal component to RNA template presence.

In addition, we found that the persistence of template RNAs reflects the degree of fragmentation of the corresponding germline locus, and chromosomes that require a larger number of rearrangement events tend to have templates that appear for shorter durations. While one might expect chromosomes with more complex scrambling patterns to be more challenging (and time-consuming) to descramble and therefore require longer-lived templates, it is intriguing that such complex chromosomes are produced instead by some of the most transient lncRNAs in our dataset. Overall, the heterogeneity of the lncRNA population during development suggests that different somatic chromosomes may differ in their rearrangement pathways, and that the entire population of molecules does not differentiate in lockstep. Such a strategy may reflect the cell’s need to properly assemble over 16,000 chromosomes to build its somatic genome. In the future, it would be fruitful to extend the types of single-locus studies of DNA rearrangement pathways in Möllenbeck et al. (2008) to a genome-wide level of analysis, to test the hypothesis that the timing of chromosome rearrangement correlates with the timing of template RNA accumulation.

The relationship between template RNAs and the other RNA molecules that participate in genome rearrangement is also a subject of ongoing inquiry. On the basis of sequence, the telomere-to-telomere RNA transcripts of MAC chromosomes may perform myriad roles, for example, providing the lncRNAs that not only guide DNA rearrangement, but also establish gene dosage levels (Nowacki et al. 2010) and supply the possible precursors to piRNAs that protect
regions of the germline DNA from elimination (Fang et al. 2012; Zahler et al. 2012). The present study identified template RNAs for both scrambled and nonscrambled loci, as well as for “IES-less” chromosomes that require no DNA deletion and only excision from the MIC genome, plus telomere addition. Surprisingly, lncRNAs in the latter category tended to persist longer than those for chromosomes with multiple MDSs. Combined with the previous observation that DNA rearrangements in *Oxytricha* are highly error-prone (Möllenbeck et al. 2008) and that SNPs from injected template RNAs can transfer to the MAC chromosomes of daughter cells (Nowacki et al. 2008), this implicates the lncRNAs in a third role, providing the templates for RNA-guided DNA proofreading (Nowacki et al. 2008), allowing the cell to identify and correct or degrade aberrant rearrangements (Bracht et al. 2012). In this scenario, late expressed templates may participate in one of the final steps of genome rearrangement, allowing the cell to correct errors.

### 2.5 Methods

#### 2.5.1 Cell Culture and Harvesting

*Oxytricha trifallax* mating compatible strains JRB310 and JRB510 were cultured in inorganic salt media according to established protocol (Chang et al. 2004) with *Chlamydomonas reinhardii* and *Klebsiella oxytoca* as food sources. The cells were left overnight to exhaust their food, then filtered through gauze to remove any remaining algae before mixing them in equal proportions to initiate conjugation (Khurana et al. 2014). *Oxytricha* cells do not synchronize during conjugation. Cells were estimated to be within approximately six hours of one another in terms of development throughout the time course.
2.5.2 RNA Isolation and Template Amplification

We collected whole-cell RNA from conjugating cells at six developmental time points (0, 6, 12, 18, 36, and 60 hours post-mixing) using a TRIzol extraction kit (Invitrogen). We DNase treated (Turbo DNase) 10µg RNA, precipitated it in phenol-chloroform, and resuspended it in 24µL nuclease-free water (Agencourt Bioscience Corp., Beverly, MA). Reverse transcription reactions were performed with Superscript III enzyme on 3µg RNA, using a telomeric primer (5'− CCCCCAAACCCCAAAACCC -3'). In addition, control reactions at each time point without reverse transcriptase showed no qualitative product (Fig. 2.9), but were sequenced as negative controls. For PCR we used eight replicates per time point to reduce jackpot effects. Fragments were amplified using FastStart enzyme with 0.5µM of the same telomeric primer used for reverse transcription through 40 cycles of touchdown PCR (70-55° for 30 cycles, then 10 cycles at 55°). Replicates were pooled and digested with BsrGI and Scal-HF (NEB) to eliminate contaminating hexamer repeats observed in previous experiments (data not shown).

Figure 2.9: DNA contamination is low in lncRNA data. Lanes on the left, generated after reverse transcription with a telomeric primer, show amplification at all time points, whereas no product is visible in any of the lanes without reverse transcriptase. NT, no template negative control. Marker is 1 Kb Plus DNA Ladder (ThermoFisher).
2.5.3 Library Preparation and High-Throughput Sequencing

Touchdown PCR products were sheared to 400bp using the Covaris MiniTube system and gel purified with the QIAquick Gel Extraction Kit to obtain 350-450bp fragments. Library preparation was carried out using a standard Illumina protocol, with end repair, A-tailing, and adaptor ligation, followed by a second round of gel purification. We eluted (Minelute, Qiagen) purified product into 10µL nuclease-free water and amplified 5µL by 12 cycles of PCR before gel purification, validation on a Bio-Analyzer, and sequencing on an Illumina GAIIx HiSeq instrument. Raw RNA-seq read files are available on SRA under accession SRP079066.

2.5.4 RNA-seq Analysis

We pruned the complete *Oxytricha trifallax* macronuclear genome (Swart et al. 2013) to a subset of 14,162 chromosomes considered high confidence (with both telomeres assembled) and less than 5,000bp in length (available at http://trifallax.princeton.edu/data/pacbio_twotelo_l5000bp.fa), as the incubation time we used for PCR precluded the capture of RNAs corresponding to longer chromosomes. Because template RNAs are expected to cover the entire chromosome without splicing, we used the non-spliced aligner BWA MEM (Li 2013) to align reads to the subset. Overall, we found 10,507 *Oxytricha* chromosomes with at least one mapped IncRNA read. From these we selected for further analysis (Figures 3, 5, 7) a high-confidence subset of 4,744 chromosomes with at least sixteen pairs of reads mapped in the proper orientation and with the correct insert size and with at least thirty-two times as many mapped reads in the +RT dataset as the -RT dataset.
2.5.5 RT-PCR Validation

We followed the protocol of Nowacki et al. (2008). Briefly, Turbo DNase treatment (Thermo AM 2238) was followed by reverse transcription of 3μg RNA isolated at 12 hrs using a long primer containing telomeric sequence and a “-RT” control. Subsequent to one-sided PCR used a short anchor primer plus one gene-specific primer, testing forward and reverse strands independently. PCR was carried out with Phusion enzyme (NEB) for 35 cycles of: 98°C 10 sec (40 sec in the first cycle), 55° 30 sec, 72° 25 sec; followed by 72° for 5 minutes. Primer sequences are as follows (all 5' to 3'):

Reverse transcriptase primer for anchor PCR
ACTATAGGGCACGCGTGTCGACGCGCCGGCTGGTCCCCAAAACCCCAAAACCCC
AAAA

Anchor primer
ACTATAGGGCACGCGTG

Gene-specific primers:

Contig8682.0_Forward
GGTTATTGATGCACTTAAATTACACTG

Contig8682.0_Rev
CCACATGCATGATACTGGATTTC

Contig13450.0_Forward
CATATCAACGAGTTGAGAGATTC

Contig13450.0_Rev
TCGAAGAAAGGCTTCTTGAATTGAG

Contig10887.0_Forward
CTTAAGCTTTCTTTAGTTCTTC

Contig10887.0_Rev
CTCATAACTGCTCGACGGTTA
2.5.6 Motif Finding

We limited our motif search to chromosomes with only one gene, as noncoding regions on multi-gene chromosomes are difficult to classify as either upstream or downstream: some regions are 5' to one gene while 3' to another. The single-gene chromosomes were binned into two categories based on whether or not our survey found an RNA template for them at any time during development, the telomeres were removed, and the upstream (all sequence 5' of the transcription start site) and downstream (all sequence 3’ of the transcription termination signal) regions were extracted. We also removed noncoding regions < 5nt long from the dataset according to the requirements of the motif-finding algorithm. We used FIRE-1.1a (Elemento et al. 2007) for discriminatory motif finding with the parameters “--nodups=1” and “--exptype=discrete.”

2.5.7 Statistical Analysis

We used the R programming environment (R Development Core Team 2008) for all statistical analyses, with the ggplot2 package (Wickham 2009) for visualization and the subSeq package (Robinson and Storey 2014) for the resampling analysis. For all statistical analyses raw RNA read counts were normalized by library size, chromosome length, and DNA copy number.

2.5.8 RNA Dataset Comparisons

We used publicly-available mRNA data from Swart et al. 2013 (downloaded from http://trifallax.princeton.edu/cms/databases/raw-data/transcriptome/reads/RNA-seq/WUGSC) and piRNA data from from Fang et al. 2012 (GSE35018) for comparison with IncRNA levels during development.
Chapter 3

Capture of complete chromosomes in single sequencing reads reveals widespread structural variation within a genome\(^2\)

3.1 Abstract

Whole-genome shotgun sequencing, which stitches together millions of short sequencing reads into a single genome, ushered in the era of modern genomics and led to a rapid expansion of the number of genome sequences available. Nevertheless, genome assembly remains difficult and prone to error. Ultimately, only a sequencing technology capable of capturing complete chromosomes in a single run could resolve all ambiguities. Even “third generation” sequencing technologies produces reads far shorter than most eukaryotic chromosomes. However, the ciliate *Oxytricha trifallax* has a somatic genome with thousands of chromosomes averaging only 3.2 kbp, making it an ideal candidate for exploring the benefits of sequencing whole chromosomes without assembly.

We used single-molecule real-time sequencing to update the published *Oxytricha trifallax* JRB310 genome assembly. In this version over 50% of the completed chromosomes derive from single reads. The improved assembly includes over 12,000 new chromosome isoforms, which demonstrate that somatic chromosomes derive from variable rearrangements between somatic segments encoded up to 191,000 base pairs away. However, while long reads reduce the need for

\(^2\) Adapted from Lindblad KA, Bracht JR, Sebra RP, Hutton ER, Landweber LF. (In Review.) Capture of complete chromosomes in single sequencing reads reveals widespread structural variation within a genome.
assembly, a hybrid approach that supplements long-read sequencing with short reads produced the most complete assembly overall.

This assembly provides the first example of complete eukaryotic chromosomes captured by single sequencing reads and demonstrates that traditional approaches to genome assembly can mask considerable structural variation.

### 3.2 Background

Whole-genome shotgun sequencing, first pioneered during the human genome project, has become such common practice that over 16,000 genome assemblies are available from NCBI (as of June 6, 2016). Despite its ubiquity, genome assembly is still a challenge, requiring the computation of overlaps among millions of short reads. This makes it difficult to place repetitive elements, resolve the length of microsatellite repeats, or capture haplotypes over large genomic regions. Despite improvements in assembly algorithms, the only way to completely overcome these issues would ultimately be to use a sequencer capable of reading the sequence of each chromosome in full. Although current sequencing technologies fall far short of this mark, read lengths have increased substantially. Pacific Biosciences’ single-molecule real-time (SMRT) sequencing platform achieves read lengths as high as 50,000 base pairs (Berlin et al. 2015), while reads over 200,000 base pairs long have been reported from Oxford Nanopore’s MinION (Ip et al. 2015). The higher resolution provided by these long reads has made it possible to produce high-quality reference sequences that capture structural variation that short-read sequencing cannot resolve (Chaisson et al. 2014) and even automate the completion of microbial genomes (Koren et al. 2015).

While it is not yet possible to produce reads long enough to capture most eukaryotic
chromosomes, *Oxytricha trifallax*'s tiny “nanochromosomes” fall well within the range of recent long-read sequencing technologies. Like all ciliates, *Oxytricha* has two nuclear genomes, a transcriptionally silent germline and a compressed somatic genome used for most of the cell's transcription. The somatic genome is highly fragmented, consisting of ~20,000 different chromosomes averaging 3.2kb in length (Swart et al. 2013), which are produced from the germline through a process of genome rearrangement that eliminates 90-95% of the germline sequence.

In addition to small chromosome size, *Oxytricha*'s somatic genome displays several features that make traditional assembly difficult. Approximately 25% of chromosomes contain one or more internal sites used for telomere addition, which terminates the chromosome. The use of internal telomere addition sites produces a family of smaller isoforms that contain only part of the chromosome’s sequence (Swart et al. 2013). Furthermore, somatic chromosomes exhibit copy number variation that can range over an order of magnitude, which is well outside the assumptions of most assembly programs and sequencing techniques.

In 2013, our lab published a high-quality assembly of *Oxytricha*’s somatic genome using a combination of Sanger, 454 and Illumina data. Here we present an updated version incorporating SMRT sequencing. The improved assembly includes over 15,000 complete chromosomes captured in single reads, entirely without assembly. We find that long reads are ideal for capturing the large number of structural variants in the *Oxytricha* somatic genome and discuss the relative merits of different sequencing strategies for producing the highest-quality assembly for an extensively fragmented genome.
3.3 Results

Figure 3.1: SMRT sequencing reads are long enough to capture complete *Oxytricha* chromosomes. The length distribution of corrected SMRT subreads is similar to the length distribution of *Oxytricha* chromosomes. To improve readability, a random subsample of 50,000 SMRT subreads is shown, and the twelve chromosomes longer than 20,000bp have been omitted from the plot.

3.3.1 Over half the *Oxytricha* somatic genome can be completely sequenced without assembly

We isolated *Oxytricha* somatic DNA for SMRT sequencing and recovered 599,310 subreads after long read self-correction. As expected, the distribution of sequencing read lengths closely matches the length distribution of *Oxytricha* somatic chromosomes (Fig. 3.1), and 324,445 corrected subreads contained telomeric sequences on both ends, indicating that they are complete chromosomes. These reads with two telomeres represent 11,378 distinct chromosomes or 51% of the contigs in the published assembly; thus, over half of the genome can be completely
sequenced without assembly.

We used the Celera Assembler to assemble the corrected reads that lacked telomeric sequences on both ends into contigs and combined these contigs with the single-read chromosomes to produce a long-read-only assembly (Table 3.1). Although this assembly contains over 9,000 more contigs than the previously published assembly, the majority of the new additions are alternatively fragmented isoforms of previously sequenced chromosomes (see “Long-read sequencing discovers novel chromosome isoforms”). While SMRT sequencing provided good coverage of chromosomes around the somatic genome's mean 3.2kbp length, it was unable to capture most of the shortest chromosomes, largely because short reads were filtered out at several points during the data cleaning process. The shortest gene-containing two-telomere chromosome in our assembly was 618bp, compared to 502bp in the published assembly. Meanwhile, the longest chromosome captured by a single read was 12,467bp, which bears a single long gene encoding a dynein motor protein. Overall, 50% of contigs ≥10,000bp were present in the long read data, compared to 77% of contigs between 1,000bp and 10,000bp. This indicates that SMRT sequencing was able to capture long chromosomes nearly as well as short ones.
Table 3.1: Assembly statistics for long- and short-read genome assemblies.

<table>
<thead>
<tr>
<th></th>
<th>Published Assembly</th>
<th>Pure PacBio Assembly</th>
<th>Hybrid Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly Size</td>
<td>64MB</td>
<td>87MB</td>
<td>108MB</td>
</tr>
<tr>
<td>Total Contigs</td>
<td>22450</td>
<td>31664</td>
<td>39023</td>
</tr>
<tr>
<td>Unique Predicted Proteins$^a$</td>
<td>21106</td>
<td>16251</td>
<td>35809</td>
</tr>
</tbody>
</table>

$^a$ Based on a 90% similarity cutoff

To produce a final assembly that combines the strengths of the short read assembly with long read data, we combined our pure long read (PacBio) assembly with high-confidence contigs from the published assembly to create a hybrid assembly containing all high-confidence chromosome isoforms identified by either approach. The previously published genome assembly had been judged largely complete based on its complement of tRNA genes and overlap with the CEG database of core eukaryotic proteins (Swart et al. 2013). Moreover, although the complement of unique predicted proteins is much higher for the hybrid assembly, it contains only 94 more protein domains that were not found in the previous assembly. This suggests that, rather than having missed large numbers of functional proteins in the previous assembly, the larger proteome size in the hybrid assembly is mostly accounted for by the presence of variants of existing proteins. Also, while approximately 13,500 new chromosome variants were identified in the long read data, only two entirely new, incomplete chromosomes were discovered. This suggests that the hybrid assembly is virtually complete.
Figure 3.2: Long-read sequencing reveals underlying structural variation among chromosomes. Segments of four germline contigs (1-4) rearrange to produce nine high-confidence isoforms (A-I) of one somatic chromosome, Contig14329.0. The four germline contigs most likely represent two paralogous loci; contig 2 terminates in repetitive sequences at both ends. These match repetitive sequence at the 3’ end of contig 3 and the 5’ end of contig 4. Dotted lines indicate that a contig extends beyond the region shown. Dark gray blocks on the germline contigs represent somatic sequence that is 100% identical between the two paralogs, while colored sequence represents regions that differ between the two germline loci. Colored segments in the somatic isoforms indicate the corresponding germline segments of origin; two boxes stacked vertically indicate ambiguity when the germline paralogs are identical. While most isoforms contain only sequence from one locus or the other, isoforms “A,” “B,” “D,” and “I” have variants that incorporate sequence from both loci, suggesting that alternative fragmentation and assembly can recombine segments from multiple loci, in addition to variable retention of segments within a single locus.
3.3.2 Long-read sequencing discovers novel chromosome isoforms

*Oxytricha*’s somatic chromosomes can exist in a number of forms with identical sequence but different lengths. These “alternative fragmentation” isoforms are often masked by genome assembly pipelines that merge short chromosomes into larger ones with the same sequence. Previous estimates of the level of alternative fragmentation in *Oxytricha* were based either on PCR examination of individual loci (Herrick et al. 1987) or on the inference of telomere addition sites by identifying pileups of telomere-containing reads (Swart et al. 2013). SMRT sequencing captures these variants in their entirety. Our genome-wide analysis of alternative fragmentation sites identified 23,341 distinct chromosome isoforms, with 4,676 of the 14,345 (33%) detected chromosomes demonstrating at least one alternative fragmentation site. Mapping telomere-containing Sanger sequencing reads predicted 12,213 isoforms (Swart et al. 2013), 76% of which are also found among the isoforms in the long read assembly. This indicates that SMRT sequencing captures the same kind of isoforms, but notably it finds more of them.

Furthermore, the ability of long read sequencing to retrieve complete sequences of the isoforms, not just their lengths, allows us to examine alternative fragmentation on a genome-wide scale. Previous studies have inferred that multiple germline loci may contribute to families of alternative fragmentation isoforms (Herrick et al. 1987; Klobutcher et al. 1988). With the current data provided by long-read sequencing, we find that some isoforms may derive from mixing and matching between these different loci, rather than from processing each locus separately. For example, one of the chromosomes with the most fragmentation isoforms in our dataset, Contig14329.0, has nine isoforms (Fig. 3.2). Of these, four incorporate sequence from two separate germline loci. The distance between these loci is unknown, but they are at least 28,299bp apart and may derive from different germline chromosomes, which would require
inter-chromosomal recombination. The other five isoforms include segments from only one locus or the other. To produce the full complement of isoforms for this chromosome the cell must therefore undertake variable processing within a single locus, as well as combine sequences from multiple loci. Several of the alternatively fragmentated isoforms also contain segments from just one locus or the other, suggesting that an unknown mechanism might regulate which isoforms a locus produces. We find that variable processing is widespread, with 2,314 alternatively fragmented chromosomes deriving from two or more paralogous germline loci.

Figure 3.3: Hybrid error correction outperforms long read self-correction. With 50x coverage of short-read data, hybrid error correction produces a more complete assembly than self-correction, even at twice the minimum recommended long-read coverage.
3.3.3 Hybrid error correction produces the most complete somatic genome assembly

To compare a pure long read self-correction strategy to one that uses short reads to correct PacBio reads, we subsampled our long read data and assessed the completeness of assemblies produced using the two correction methods. Overall, hybrid error correction outperforms long-read-only error correction at all sequencing depths (Fig. 3.3), and while the number of contigs recovered by hybrid error correction begins to saturate with eight flow cells' worth of data, the steep slope of the long-read-only curve suggests that considerably more sequencing depth would be necessary to correct all chromosomes using only long reads.

The heterogeneous copy number of chromosomes in *Oxytricha's* somatic genome may be the root cause for the inadequacy of long-read error correction. While the average somatic chromosome copy number is approximately 2,000 (Prescott 1994), some chromosomes can be amplified to over 200,000 copies. This reduces the effective coverage for low-copy number chromosomes, as more abundant chromosomes absorb a disproportionate amount of sequencing depth. Both the hybrid-corrected and self-corrected genome assemblies were biased towards chromosomes with significantly greater copy number than average, even when all flow cells were incorporated (Welch's one-sided $t$-test, $t = 4.1652$, $p = 1.559e-05$ for hybrid correction, $t = 4.7637$, $p = 1.559e-05$ for self-correction). However, hybrid error corrections resulted in a steeper decline in mean chromosome copy number across the genome as sequencing depth increased, compared to self-correction (79 for the hybrid error correction, 13 for self correction). This indicates that the hybrid error correction incorporated more low-abundance chromosomes as the amount of long read data increased, relative to the self-correction method. While we recovered 135x coverage of corrected sequence from strictly long reads, this derived from only hundreds of thousands of PacBio reads, each an individually sampled molecule, compared with
tens of millions of short-read Illumina sequences. The increased depth that can be achieved with short reads is thus more important to the completeness of the final genome assembly than the increased resolution provided by long-read sequencing.

3.4 Discussion

As long-read sequencing technology improves, it may eventually be possible to sequence complete chromosomes of most organisms in a single contiguous read. For now, *Oxytricha*'s highly fragmented genome provides the first opportunity for genome sequencing without assembly. This approach permitted the discovery of structural chromosome isoforms that were masked by traditional assembly pipelines. Long-read sequencing dramatically increased the number of alternative isoforms that we could identify. Furthermore, we conclude that many of these isoforms may derive from recombination among multiple germline loci, in addition to variable processing within a locus. Where allelic variation is present, this implies that genome rearrangement may occur between—as well as within—germline chromosomes. The observation that some isoforms derive exclusively from one locus or the other also raises the question of what regulates this selection and what determines the range of isoforms produced. The finding that piRNAs can strongly influence chromosome fragmentation patterns and lead to alternatively processed chromosomes (Bracht et al. 2016) suggests that the piRNA pathway is involved (Fang et al. 2012) in this process.

The long read sequencing in this study permitted a higher quality examination of closely-related chromosome isoforms than a previously published assembly. However, for *de novo* genome assembly, the variation in chromosome copy number in *Oxytricha* makes the approach less economical than short-read sequencing. For such a highly fragmented genome with variable
chromosome copy number, we recommend that future studies perform an initial assembly based on short read data and use long reads to investigate structural variants, the area where we reaped the most benefit for this genome.

3.5 Methods

3.5.1 Cell Growth and Culture

Cells growth, harvest, and nuclei isolation were carried out as described in (Chen et al. 2014), with the exception that the pellet was collected after the initial centrifugation step rather than from the 10% gradient fraction.

3.5.2 Library Preparation and Sequencing

Library preparation and sequencing were performed according to the manufacturer’s instructions and reflects both the P5-C3 and P6-C4 sequencing enzyme and chemistry. Aliquots of 5 µg of extracted high-quality genomic DNA were enriched for MAC DNA and verified using Qubit analysis. After quantification, DNA was diluted to 150 µL using Qiagen elution buffer at 33 µg / µL. The sample was pipetted into the top chamber of a Covaris G-tube spin column and sheared gently for 60 seconds at 4500 rpm using an Eppendorf 5424 bench top centrifuge, and the sheared DNA was re-purified using a 0.45X AMPure XP purification step.

After purification, ~1.2 ug of purified and sheared sample was repaired using DNA damage repair solution (1X DNA damage repair buffer, 1X NAD+, 1 mM ATP high, 0.1 mM dNTP, and 1X DNA damage repair mix) with a volume of 21.1 µL and incubated at 37°C for 20 minutes. DNA ends were repaired by adding 1X end repair mix to the solution and incubating at 25°C for 5 minutes, followed by a second 0.45X Ampure XP purification. Next, 0.75 µM of blunt adapter was added to the DNA, followed by 1X template preparation buffer, 0.05 mM ATP low
and 0.75 U/µL T4 ligase to ligate (final volume of 47.5 µL) the SMRTbell adapters to the DNA fragments. This solution was incubated at 25ºC overnight, followed by a 65ºC 10-minute ligase denaturation step. After ligation, the library was treated with an exonuclease cocktail to remove un-ligated DNA fragments using a solution of 1.81 U/µL Exo III 18 and 0.18 U/µL Exo VII, then incubated at 37ºC for 1 hour. Two additional 0.45X Ampure XP purifications steps were performed to remove < 2000 bp molecular weight DNA and organic contaminants.

Upon completion of library construction, samples were validated as ~5kb using an Agilent DNA 12000 gel chip. We used Sage Science Blue Pippin 0.75% agarose cassettes to select a MAC-enriched library in the range of 5,000 bp – 50,000 bp. Then primer was annealed to the size-selected SMRTbell with the full-length libraries (80ºC for 2 minute 30 followed by decreasing the temperature by 0.1º/s to 25ºC). For sequencing conducted on the first 2 SMRTcells (2014), the polymerase-template complex was then bound to the P5 enzyme using a ratio of 10:1 polymerase to SMRTbell at 0.5 nM for 4 hours at 30ºC and then held at 4ºC until ready for magbead loading, prior to sequencing using the C3 chemistry. For the later 8 SMRTcells (2015), the complex was bound to the P6 enzyme and sequenced using the C4 chemistry. The magnetic bead-loading step was conducted at 4ºC for 60 minutes. The magbead-loaded, polymerase-bound SMRTbell libraries were placed onto the RSII machine at a sequencing concentration of 100 to 110 pM and sequenced across two SMRTcells using P5-C3 and 8 additional SMRTcells using P6-C4 chemistry.

### 3.5.3 Genome Assembly

We used Pacific Bioscience’s SMRT Pipe 2.3.0 (available from https://github.com/PacificBiosciences/SMRT-Analysis/wiki/SMRT-Pipe-Reference-Guide-v2.1)
to quality trim and to filter raw SMRT sequencing reads, using default parameters but enabling the artifact filter (parameter value -1000) in order to remove chimeric reads. Reads that passed the filter were self-corrected using PBcR (Berlin et al. 2015) and default parameters.

Error correction deleted the telomeres from most reads, so we gathered all raw reads that had at least one telomere, based on matching to the regular expression [TG]*TTTTGGGGTTTT, [TG]*GGGTTTTGGGG, [AC]*AAAACCCCAAAA, or [AC]*CCCCAAAACCCC with an edit distance of two. The first and last 1,000bp of these reads were corrected using ECTools (Lee et al. 2014) and a 50x coverage subset of Illumina reads from the previously published Oxytricha somatic assembly (Swart et al. 2013). Chromosome ends corrected in this manner were aligned to the PBcR-corrected read and the missing bases filled in from the ECTools corrected read.

Some corrected reads were chimeras of multiple chromosomes, characterized by embedded telomeric sequences, or sequencing artifacts composed almost exclusively of homopolymer runs. As a result, we filtered out all corrected reads containing a homopolymer run of >10bp or a non-terminal telomeric sequence (matching the regular expression [AC]*(CCAAAACCCCAAAA) or (GGTTTTGGGGTTTT)[TG] with an edit distance of one or [AC]*CCCAAAACCCGGGGTTTTGGG[TG*] or [TG]*GGGTTTTGGGGCCCAAAACCC[AC*] with an edit distance of three).

After filtering, all reads with telomeric sequences on both ends were considered complete chromosomes and retained, while reads with one or fewer telomeres were assembled using Celera Assembler 8.3rc (Berlin et al. 2015). We combined the assembled contigs with the two-telomere reads and clustered the resulting sequences at a 90% identity threshold using VSEARCH (https://github.com/torognes/vsearch) and took the centroid contig for each of the resulting clusters to produce a final set of unique chromosomes.
We determined alternative fragmentation isoforms by extracting all two-telomere single reads and contigs from our data and masking the telomeres according to the procedure described in (Swart et al. 2013). We then used BWA MEM (Li 2013) to map the masked reads against the subset of unique chromosomes in our assembly. We grouped all reads with both start and end positions within 50bp of one another into distinct isoforms and clustered all reads assigned to each isoform at a 97% similarity threshold. We added the consensus sequence of each cluster comprising at least two contigs to the assembly.

To finalize the assembly, we added contigs that were captured in the published Oxytricha assembly but not in our long-read assembly. These included two-telomere contigs fewer than 600bp long and contigs either without an analog in the long-read data, or where the longest isoform in the long-read assembly was at least 75bp shorter than the version in the published assembly. In cases where the published contig was longer and the long-read version had both telomeres, we considered the long-read form an alternative fragmentation isoform and retained it in addition to adding the longer published contig. If the long-read form had fewer than two telomeres, it was discarded instead. Finally, we removed contigs where at least 50% of the contig sequence was covered by a known germline repetitive element or satellite repeat. We also removed as likely contaminants any contigs without any telomeres and which were less than 20% covered in the germline genome.

3.5.4 Analysis of Alternative Chromosome Fragmentation

To compare the alternative fragmentation isoforms found by SMRT sequencing with those predicted by older sequencing technologies, we masked all two-telomere corrected reads as described above and mapped them against the published somatic genome assembly. We then
grouped reads into distinct isoforms as described above, choosing only the longest hit for each read. In addition, because a size selection step was used in the Sanger sequencing that produced the original predicted isoforms, we filtered the resulting isoforms to include only those less than 6,000bp long. To determine whether an isoform found by one method was also discovered by the other, we used BEDTools 2.25.0 intersect (Quinlan and Hall 2010) with the options -F 90 -f 90 to count only isoforms that were at least 90% covered in both assemblies.

To analyze how somatic isoforms relate to their germline loci, we selected all isoforms supported by at least two corrected reads and aligned them to the germline genome with Megablast (Camacho et al. 2009). Isoforms containing sequence from more than one paralogous locus were identified by choosing the best hit for each germline sequence comprising the isoform, then filtering for isoforms containing segments from two or more different germline loci.

3.5.5 RNA-Seq

We prepared RNA-seq libraries from vegetatively-growing, starved, and encysted *Oxytricha* cells. The vegetative culture was grown according to the same procedure used for collecting MAC DNA. Cells for starved and cyst libraries were placed in a clean dish and incubated at 4°C and room temperature, respectively, for 5 days. RNA for the starved and vegetative samples was extracted using TRIzol® Reagent (Life Technologies™). RNA for the encysted sample was extracted using 0.25mm silica carbide beads in the UltraClean Microbial RNA Isolation Kit (MO Bio). Three replicates of vegetative cell RNA, three replicates of encysted cell RNA, and one replicate of 4°C-starved RNA were prepared with the Epicentre Stranded kit, along with a no-RNA input control. cDNA samples were amplified in 12 PCR
cycles. Library preparation and sequencing was performed by the Lewis-Sigler Institute for Integrative Genomics Sequencing Core Facility using the Illumina Truseq Library Prep Kit.

3.5.6 Gene Prediction

We used a gene prediction model trained on Oxytricha data and presented in (Swart et al. 2013) in conjunction with AUGUSTUS 2.5.5 (Stanke and Morgenstern 2005) to predict genes for all three assemblies. We used the RNA-seq data collected from vegetatively-growing, starved, and encysted cells (accession identifier goes here); previously-published RNA-seq collected from cells undergoing conjugation and genome rearrangement (at 0, 10, 20, 40, and 60 hours after cells were mixed to initiate mating) (Swart et al. 2013); and transcription start site data [in preparation] to provide hints to the gene prediction software. We mapped reads to the genomes using STAR 2.4.1c (Dobin et al. 2013) in two-pass mode, then generated hints files according to the instructions on the AUGUSTUS web site (http://augustus.gobics.de/binaries/readme.rnaseq.html). We set the bonus for transcription start site hints to $1\times 10^5$ and the malus to 0.34. Finally, we ran AUGUSTUS with the options --UTR=on and --alternatives-from-evidence=true. We predicted protein domains using Interproscan 5 RC5 (Goujon et al. 2010) using default parameters.

3.5.7 Subsampling Analysis

We took random subsets of one, two, four, six, seven, and all eight of the flow cells from our 2015 sequencing run and used them to complete de novo Oxytricha assemblies. The reads were first filtered using the same methodology used for the primary assembly, then error corrected using either the PBcR pipeline or ECTools. For the one- and two-flow cell subsets
corrected by PBcR, we used the recommended high-sensitivity parameter settings intended for low coverage assemblies (QV=52 asmOvlErrorRate=0.1 asmUtgErrorRate=0.06 asmCgwErrorRate=0.1 asmCnsErrorRate=0.1 asmOBT=1 asmObtErrorRate=0.08 asmObtErrorLimit=4.5 utgGraphErrorRate=0.05 utgMergeErrorRate=0.05). Otherwise, all settings used were the default. After error correction, reads were assembled using Celera assembler. To assess genome completeness, we mapped corrected reads and assembled contigs against the previously published *Oxytricha* assembly and counted the number of contigs at least 80% covered by either a single read or a single contig from the de novo assembly.

3.5.8 **Statistical Analysis**

We carried out all statistical analyses in the R programming environment (R Development Core Team 2013) and used the ggplot2 package (Wickham 2009) to generate figures.
Chapter 4

Active transposons contribute to genome fragmentation in *Oxytricha*\(^3\)

4.1 Abstract

The ciliate *Oxytricha trifallax* is a single-celled eukaryote with both a transcriptionally-silent germline genome and a somatic genome responsible for most transcription during asexual growth. During sexual conjugation, genome rearrangements precisely delete as much as 95% of the germline and knit together over 225,000 retained sequences to form a new somatic genome. Telomere-bearing elements (TBEs), a class of Tc1/mariner transposons, are essential to the rearrangement process. Previous research noted the similarity of the TBE insertion sequence to the direct repeats that flank thousands of deleted germline-limited sequences, which led to the hypothesis that these sequences are the descendants of ancient transposition events. We compared germline assemblies for two *Oxytricha* strains to identify novel TBE insertions and discovered that all four TBE families appear to be active. In addition, the majority of recent insertions appear to have occurred in close proximity to somatic-destined sequences, in some cases actually interrupting the somatic sequence. Such recent insertions form new germline-limited sequences that result in differential processing of somatic-destined loci between the two strains to produce the same somatic chromosome. These findings support the hypothesis that at least some IESs result from ancient transposition events and demonstrate that active transposons

---

\(^3\) Adapted from Lindblad KA, Beh LY, Landweber LF. (In Prep.) Active transposons contribute to genome fragmentation in *Oxytricha*.
continue to shape the evolution of genome architecture in the *Oxytricha* germline.

### 4.2 Background

Transposable elements comprise a substantial portion of most eukaryotic genomes, representing approximately 45% of the human genome (Lander et al. 2001) and 20% of the germline genome of the ciliate *Oxytricha* (Chen et al. 2014). While originally viewed purely as selfish elements, some transposons can provide benefits to the host cell, like the RAG1 transposase that was recruited for V(D)J recombination in the vertebrate immune system (Aragawa et al. 1998) or the centromere protein CENP-B (Casola et al. 2008). In the ciliates Paramecium (Baudry et al. 2009) and Tetrahymena (Cheng et al. 2010) a domesticated piggyBac transposase has been recruited to facilitate genome editing during nuclear development, but the influence of transposable elements in the underlying germline genome rearrangements in ciliates remains largely unexplored.

All ciliates share a dual-genome architecture, with an encrypted germline (micronuclear) genome that is silent during most of the cell’s life and a streamlined somatic (macronuclear) genome that supplies the bulk of transcription during asexual growth. After sexual conjugation a new somatic genome is produced from the germline through a dramatic process of DNA elimination and rearrangement. In the ciliate *Oxytricha trifallax* this genome editing is especially pronounced, leading to deletion of over 90-95% of the germline genome, including hundreds of thousands of internally eliminated sequences, or IESs, as well as all copies of transposable elements and other germline-limited DNA. In addition, the remaining macronuclear-destined segments (MDSs) sometimes require reordering or inversion to restore the coding content of the mature somatic genome (reviewed in Yerlici & Landweber 2014). A class of Tc1/Mariner transposons called TBEs is required for genome rearrangement, and our
lab previously showed that a knockdown of these elements results in editing errors in the mature cell (Nowacki et al. 2009). TBEs, or telomere-bearing elements, are so named because of the similarity between their terminal inverted repeats and the *Oxytricha* telomeric sequence. TBEs average 3.8kb long and are divided into four sub-families (TBE1, TBE2.1, TBE2.2, and TBE3) on the basis of their three open reading frames (Chen and Landweber 2016). They constitute roughly 15% of the *Oxytricha* germline (Chen et al. 2014). In addition, many TBEs are under purifying selection, further supporting the idea that they provide an essential function (Witherspoon et al. 1997; Nowacki et al. 2009; Chen & Landweber 2016).

In addition to their developmental importance during genome rearrangement, TBE transposons may also shape the evolution of the distinctive germline architecture and genome encryption in ciliates. One evolutionary hypothesis for the interruption of ciliate genes by IESs is that IESs are degraded remnants of transposable elements that invaded the germline (reviewed in Klobutcher & Herrick 1997). Seegmiller et al. (1996) specifically noted the similarities between IESs in one *Oxytricha* germline locus and TBE elements. In the distantly-related ciliate *Paramecium*, IESs are flanked by two base pair direct repeats similar to transposon target site duplications (Klobutcher & Herrick 1995), and many IESs also show some sequence similarity to the domesticated PiggyMAC transposon required for their removal (Arnaiz et al. 2012), which reinforce the idea that they were originally PiggyMAC insertions. However, the potential transposon origin of germline-limited segments is less clear in the *Oxytricha* lineage, where the short direct repeats (called pointers) present at recombination junctions vary dramatically in length and have no consensus sequence, although the subset of two and three base pair pointers do display features similar to the ANT target site duplication of TBE transposons (Chen et al. 2014).
We sequenced a new strain of *O. trifallax* (JRB510) and compared it to the reference strain JRB310. This allowed us to infer the sites of TBE insertion events that occurred after the divergence of the closely-related strains. This analysis reveals that TBE elements are still mobile in the *Oxytricha* genome. Moreover, they appear to preferentially insert within MDSs, leading to the creation of new strain-specific IESs, and thus new fragmentation patterns that the cell must resolve during genome rearrangement. These findings support the hypothesis that many IESs originated as transposable elements. Our comparative analysis also demonstrates the resilience of ciliate genomes, which are surprisingly tolerant of transposable element insertions and most likely continue to be shaped by them today.

**Table 4.1: Comparison of the JRB310 and JRB510 germline genome assemblies**

<table>
<thead>
<tr>
<th></th>
<th>JRB310</th>
<th>JRB510</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly Size (MB)</td>
<td>566</td>
<td>474</td>
</tr>
<tr>
<td>N50</td>
<td>2932</td>
<td>27807</td>
</tr>
<tr>
<td>Number of Contigs</td>
<td>339820</td>
<td>25720</td>
</tr>
<tr>
<td>Longest Contig (bp)</td>
<td>340519</td>
<td>380827</td>
</tr>
<tr>
<td>Annotated TBEs</td>
<td>39250</td>
<td>34721</td>
</tr>
</tbody>
</table>

### 4.3 Results

#### 4.3.1 The *Oxytricha* germline contains active TBE transposons

To search for potential recent TBE transposition events, we sequenced and assembled the germline genome of *Oxytricha trifallax* strain JRB510, which is mating compatible with the reference JRB310 strain whose germline genome was published in Chen et al. (2014). The Illumina-based sequencing approach used for JRB510 produced a less complete assembly than the JRB310 germline assembly (Table 1), but it proved sufficient to perform an initial survey of transposable element differences between the two strains. We used a discordant pair mapping
approach to compare the JRB510 and JRB310 micronuclear assemblies (see Methods) and to identify TBE insertions exclusive to the JRB510 germline. We found 129 high-confidence predicted TBE insertions exclusive to the JRB510 germline genome. These putative recent insertions include members of all four TBE families (Figure 1) in similar proportions to the background frequencies in both strains.

Furthermore, some of the JRB510-exclusive TBE transposons appear to have undergone expansion after transposition. Plotting pairwise distances between JRB310 germline TBEs results in a bimodal distribution (Figure 2), suggesting that the four families have undergone two waves of expansion. A subset of JRB310 TBEs have low nucleotide diversity, which may indicate that these elements are under purifying selection, as investigated in Chen & Landweber (2016), or that they arose from recent transposition events.

By contrast, the four families of JRB510-exclusive inserts show different patterns of divergence. The TBE2 subfamilies, TBE2.1 and TBE2.2, have no peaks in their nucleotide diversity histograms, which suggests that their insertions are the result of multiple transpositions by individual TBE2.1 or TBE2.2 elements common to both JRB310 and JRB510. However, the TBE1 and TBE3 families show broad peaks of similar nucleotide diversities, likely due to a wave of amplification among those elements after strains JRB510 and JRB310 diverged. The small peak of TBE1 elements at very low nucleotide diversity among the TBE1 JRB510-exclusive insertions suggests that they might currently be undergoing another burst of amplification.
Figure 4.1: New transposon insertions reflect background frequencies. We identified 129 cases of transposon insertions in the JRB510 germline that were not present in strain JRB310. All four families of TBE transposon were represented, in similar proportions to the frequencies of TBE transposons shared between the two strains.
Figure 4.2: Some TBE families show expansion in JRB510. Peaks in the distributions of nucleotide diversities indicate amplification events acting on an element. The distributions for TBEs in JRB310 (blue) are bimodal, reflecting two distinct waves of expansion, while TBEs exclusive to JRB510 (red) show different patterns. The peaks in the distributions for JRB510-exclusive TBE1 and TBE3 elements suggest that these families have undergone expansion after transposition in the JRB510 germline.
Figure 4.3: Transposon insertions occur near somatic-destined segments. The distribution of distances from the nearest MDS is plotted for all JRB510-exclusive TBEs and a set of JRB310 TBEs subsampled to include an equal number of elements. Not shown are 58 JRB310 TBEs and 11 JRB510-exclusive TBEs that are more than 1,000bp away from any MDS.

4.3.2 TBE transposition events produce strain-specific germline gene architectures

TBE transposons are closely associated with MDS regions; 18.3% lie within 500bp of an MDS in the JRB310 germline (Chen & Landweber 2016). This tendency is even more extreme among the JRB510-exclusive insertions (Figure 3), of which 87% are within 500bp of an MDS. This rate of insertion near MDSs is also much higher than would be expected by chance, as MDSs represent only 11.1% of the JRB310 germline (Chen et al. 2014).

In addition to the trend of insertion near MDSs, over 40% of novel TBE insertions interrupt MDSs in the JRB510 germline genome. These TBE insertions are therefore strain-specific IESs (Figure 4), which must be removed during nuclear differentiation and genome rearrangement. As a result, the same locus is processed differently in the two mating-compatible strains, and a daughter cell must be able to correctly process either configuration, since both loci will coexist in the germline after mating between the two strains.
Figure 4.4: Novel TBE inserts lead to strain-specific fragmentation patterns. The germline locus for Contig18924 (which encodes a gene for cleft lip and palate transmembrane protein 1) has three somatic-destined segments in the JRB310 germline and four in JRB510 due to a TBE insertion in the first MDS.

4.4 Discussion

The discovery of active transposition of TBE elements in the *Oxytricha* germline underscores the flexibility of this organism’s genome rearrangement machinery. Because all transposons are eliminated during development of the somatic nucleus, the cell may be surprisingly tolerant of transposable element insertions that interrupt somatic coding sequence. Indeed, ciliates have adopted the most severe strategy to silence transposons, which is to eliminate them from the active nucleus. Their dependence on transposase activity for genome rearrangements prevents loss of the transposons from the germline, however, where they can still expand and be active (reviewed in Vogt et al. 2013).

There might also be long term advantages to the high degree of genome fragmentation in ciliates. A transposon that inserts within an MDS creates two new, shorter MDSs as well as a new IES, which may free the cell to combine new MDSs with existing MDSs to create new somatic chromosomes. The ability to use one MDS in several different somatic chromosomes has been demonstrated to facilitate the creation of new genes in the *Oxytricha* lineage (Chen et
al. 2015), and this in turn could lead to increased evolutionary innovation.

Alternatively, the prevalence of strain-specific TBEs near or within MDSs may be the result of either chance or preferential retention of the transposon. The presence of TBEs near a retained somatic sequence might confer some benefit, perhaps by recruiting factors that facilitate DNA processing during genome rearrangement.

In conclusion, the detection of hundreds of strain-specific TBE transposon insertions in *O. trifallax* reveals that IES formation is an ongoing process in the *Oxytricha* germline. TBE insertion can contribute to new IES sequence, though not all IESs are TBE transposon-derived, since other transposable element types may contribute. Furthermore, transposon-independent meiotic recombination events can also give rise to new insertions in somatic destined regions, including scrambled gene maps (Landweber 1998; Wong & Landweber 2006; Chang et al. 2004). The diversity of short direct repeat sequences used at recombination junctions may indicate involvement of multiple different families of transposable elements in the history of DNA insertions during ciliate genome evolution, and conversely the involvement of one or more domesticated transposases during nuclear development and genome rearrangement (Vogt et al. 2013). TBE elements are currently the best studied of *Oxytricha’s* transposon families. Therefore, further investigation of *Oxytricha’s* full complement of transposons may yield new insights into the impact of other transposon-derived factors on genome reorganization during both development and evolution.
4.5 Methods

4.5.1 Sequencing and Assembly of the *Oxytricha trifallax* JRB510 germline genome

*Oxytricha trifallax* strain JRB510 was grown in Pringsheim media (0.11 mM Na$_2$HPO$_4$, 0.08 mM MgSO$_4$, 0.85 mM Ca(NO$_3$)$_2$, 0.35 mM KCl, pH 7.0) to a density of ~1.5 x 10$^7$ cell/ml and fed daily with *Chlamydomonas reinhardtii*. The cultures were starved for 72 hr at 4°C to purge internal food particulates and subsequently filtered through a 10μm Nitex mesh to remove debris in cell media. We lysed harvested cells with 3% (w/v) sucrose, 0.2% (v/v) Triton X-100, and 0.01% (w/v) spermidine-trihydrochloride (pH 6.8) and centrifuged them through a 10%-40% discontinuous sucrose gradient at 250 x g for 10 min. The top 10ml of the 10% sucrose layer was transferred into a fresh tube and centrifuged at 4200 x g for 5 min to pellet micronuclei. The gDNA was isolated using a Nucleospin Tissue Kit (Macherey-Nagel), and loaded on a 0.3% (w/v) SeaKem Gold (Lonza) agarose gel. High molecular weight DNA migrating at the mobility limit was excised and purified using a QIAGEN gel extraction kit. Illumina library preparation and sequencing was then performed according to manufacturer’s instructions.

We sequenced three libraries: two 215 bp paired-end libraries and one 67 bp paired-end library. All three were assembled using MaSuRCA (Zimin, et al. 2013) with default parameters. To remove non-*Oxytricha* contaminants from the assembly, we first used BLAST+ 2.2.31 (Camacho et al. 2009) with the MegaBLAST algorithm to compare the assembly to a set of common *Oxytricha* contaminant genomes and the NCBI reference microbial genomes (downloaded May 28 2014). We removed all contigs at least 50% covered by hits with E-value <= 0.001. Next we used BLAST to compare the JRB510 macronuclear genome (GenBank accession ASM129792v1; Chen et al. 2015) with the assembly (parameters -task blastn -gapopen 5 -gapextend 2 -penalty -3 -reward 1). Finally, we used BWA MEM (Li 2013) to map JRB310
micronuclear reads to all contigs at least 10,000 bp long with no macronuclear BLAST hits of at least 95% similarity, then discarded any that were less than 25% covered.

4.5.2 Identification of Strain-Specific TBE Insertions

We used BWA MEM to map raw JRB510 germline reads against the JRB310 reference germline genome (Chen et al. 2014; GenBank accession ARYC00000000). To identify novel TBE insertion events, we used RetroSeq version 1.41 (Keane et al. 2013) in conjunction with a reference set of TBE coordinates, requiring a minimum read depth of 10. All other parameters used were the default. We eliminated all predicted insertions within 200 bp of a known transposable element. To produce a high-confidence set of predicted JRB510-exclusive TBEs we filtered the remaining predictions to include only those with a FL confidence tag of at least 6 and at least 20 supporting reads.

4.5.3 Analysis of TBE Distribution and Phylogeny

We used Clustal Omega version 1.2.0 (Sievers et al. 2011) to align and produce distance matrices for sequences of known JRB310 TBEs and JRB510-exclusive TBE insertions. The distance measure used was the Kimura-corrected pairwise aligned identity (Kimura 1980). MDS and IES annotations downloaded from MDS_IES_DB (Burns et al. 2016) and BEDTools (Quinlan and Hall 2010) were used to analyze the proximity of different transposons to somatic-destined sequences. Statistical analyses were carried out in the R programming environment (R Development Core Team 2013), and plots were generated with the ggplot package (Wickham 2009).
Bibliography


