Running head: An Alternative Method to Enrich *Tetrahymena* Micronuclear DNA

Title: An Alternative Method to Enrich *Tetrahymena* Micronuclear DNA

Authors' names: Abigail A. Howell[a,b], Brandon Neldner[c], Jonathon T. Hileman[a], Timothy J. Licknack[a,b], Wesley E. Swenson[a], Gillian H. Gile[a], Rebecca A. Zufall[d], Reed A. Cartwright[a,b]

Authors' addresses:
a Arizona State University, School of Life Sciences, 427 E Tyler Mall Tempe, AZ, USA 85287-4501
b Arizona State University, The Biodesign Institute, Tempe, AZ, USA
c Thermo Fisher Scientific, Tempe, AZ, USA
d University of Houston, Department of Biology and Biochemistry, Houston, TX, USA

Correspondence
A. Howell, School of Life Sciences, Arizona State University, 427 E Tyler Mall Tempe, AZ, USA 85287-4501, USA Telephone number: 480-292-2575; e-mail: aahowel3@asu.edu

Abstract

The ciliate *Tetrahymena thermophila* is a unicellular eukaryotic model organism with two distinct nuclei: a diploid germline micronucleus and a ~45x somatic macronucleus. Centrifugation methods used to isolate micronuclei are time and resource intensive. An alternative approach to study the genome of micronuclei is genomic exclusion; however it results in a significant portion of the original micronucleus (~30%) being lost during chromosome fragmentation and removal of Internally Eliminated Sequences (IES). Whole cell sequencing is also not a viable option for studying micronuclear specific evolution, as ~95% of extracted DNA is derived from the macronucleus. Through detergent-free nuclei isolation and flow cytometry, we have developed a successful enrichment method for the micronuclei of *Tetrahymena* for genomic analysis. We have validated MIC enrichment during flow sorting through (a) Fisher's exact tests of uniquely mapped reads to the micro and macronuclear reference genomes, (b) mean coverage depth of IES and Macronuclear-destined regions after alignment to the micronuclear reference genome, and (c) IES retention scores.

Keywords: Flow Cytometry, Tetrahymena, Mutation, Internally Eliminated Sequences (IES), Micronuclei

1. Introduction

Tetrahymena thermophila is a microbial eukaryote with an extensive genetic toolkit, contributing to the discovery of the microtubule motor, dynein (Gibbons and Rowe 1965), catalytic RNA (Kruger et al. 1982), telomere structure and telomerase (Greider and Blackburn 1985), histone acetyl transferase (Brownell et al. 1996), and programmed excision of transposon-related DNA from the somatic genome (Taverna et al. 2002), among numerous other areas of research. As with all ciliates, *Tetrahymena* exhibit nuclear dimorphism, containing a diploid germline micronucleus (MIC) and ~45x somatic macronucleus (MAC). The life cycle of *T. thermophila* is made up of two stages: asexual reproduction by binary fission and sexual conjugation between individuals of different mating types. During normal vegetative growth, transcription occurs in the MAC, and the MIC remains transcriptionally silent. When dividing asexually, the MIC divides mitotically as most eukaryotic cells would, while the MAC divides through a process called amitosis, where chromosomes are randomly segregated, potentially leading to unequal chromosome numbers during MAC division (Allen and Nanney 1958; Doerder *et al.* 1975; Orias and Flacks 1975; Nanney and Preparata 1979). Through amitosis, a heterozygous MAC can become homozygous after multiple rounds of division through the stochastic fixation of one allele, a process known as phenotypic assortment (Orias and Flacks 1975; Nanney and Preparata 1979; Merriam and Bruns 1988). When stressed, *Tetrahymena* undergo meiosis, mate, and form new zygotic nuclei from which new MIC and MAC develop. The MAC is not a direct copy of the MIC, and during the development of the MAC chromosomes are fragmented and thousands of Internally Eliminated Sequences (IESs) are removed in the MAC but retained in the transcriptionally silent MIC (Yao *et al.* 1987). This dual life cycle of sexual and asexual reproduction is thought to be an adaptive response to stress, increasing genetic diversity in an individual to increase the odds of survival (Fjerdingstad et al. 2007).

The nuclear dimorphism and dual lifecycle of *T. thermophila* provide multiple experimental advantages. The separate micronuclear and macronuclear genomes allow for lethal mutations and segmental deletions to be maintained in the MIC until mating, which makes physical mapping of mutations, DNA polymorphisms, or MIC-limited DNA elements possible without their expression in the MAC. Phenotypic assortment, another unique attribute of *T. thermophila*, allows for recessive mutations to be fully expressed after a series of asexual generations and facilitates knockdown/knockout experiments of otherwise essential genes in the MAC (Hai *et al*. 2000). When combined with a drug resistance gene, MAC knockouts can be driven to near homozygosity through stepwise increases in drug concentrations, a model which is difficult to obtain in other organisms without an additional Cre-lox system to selectively knock-down essential genes in certain tissues. Additionally, *Tetrahymena's* fast growth rate (~2- to 3-hr doubling time), sequenced MIC and MAC genomes, and established mapped genetic markers make them an ideal model system for forward and reverse genetics studies for gene discovery (Ruehle *et al*. 2016). Despite their ubiquity as a genetic model system, there is no simple, efficient method for separating the micro and macronuclei of *Tetrahymena.*

For some studies, it is desirable to isolate the micronuclei of *Tetrahymena*. This includes research on scan RNAs (scnRNAs) produced in the MIC that regulate DNA elimination during conjugation (Schoeberl et al., 2012), the enzymatic and chemical mapping of nucleosome distribution in the micronuclei (Chen et al., 2016), or the role of heterochromatic histone posttranslational modifications (PTMs) during meiosis and mitosis (Papazyan et al., 2014). However, the centrifugation protocol used in these studies to isolate the micronuclei requires at minimum 1-2 liters of culture for a sufficient yield (Sweet and Allis 2006; Duan et al., 2021).

Isolating the micronuclei of *Tetrahymena* would also be beneficial for mutation accumulation studies that seek to characterize complex mutations events (insertions, deletions, and copy number variants) which could be responsible for the fitness decline observed in previous experiments (Long et al., 2013; Long et al., 2016). Growing large volumes of culture for centrifugation and micronuclei isolation is undesirable in these studies, as increasing the number of generations and population size could bias the mutation results. To generate sufficient micronuclear material for sequencing, these studies utilized a process known as genomic exclusion (GE). In GE, a *Tetrahymena* cell is mated with a star strain of *Tetrahymena* which lacks a MIC, causing the resulting daughter cells to be generated from a single meiotic product of the original cell. The MAC in these daughter cells are descended from the original micronucleus, making all data reflective of only the MIC (Allen 1963). However, a significant portion of the original MIC, and potential mutations, are lost during somatic genome rearrangement. Whole cell sequencing is also not a viable option for studying micronuclear specific evolution, as approximately 95% of DNA in the cell consists of macronuclear DNA.

In this study we present an alternative method to enrich samples for *Tetrahymena* micronuclei using propidium iodine staining and flow sorting (Figure 1). Flow cytometry has been previously demonstrated to sort subpopulations of nuclei to high purity in *Paramecium*, a close relative of *Tetrahymena* (Guérin et al. 2017). Further, propidium iodine has shown to be successful in staining both nuclei in previous experiments with *Tetrahymena* (Po-Hsuen et al. 2015). This protocol minimizes the time and resources associated with growing large amounts of culture required in previous nuclei isolation protocols (Sweet and Allis 2006; Duan et al., 2021). We validate the flow sorted samples using high throughput DNA sequencing on the Illumina MiSeq platform. Our findings demonstrate that flow cytometry is a viable method for enriching samples for micronuclei in *Tetrahymena* that utilizes less than 25mL of culture compared to the liters of culture required of other methods.

2. Materials and Methods

2.1 Strains and Media

The *T. thermophila* strain used in this study was SB210-E from the *Tetrahymena* Stock Center (Cornell University). Cultures were grown in 25mL of Neff Medium (Cassidy-Hanley 2012) with penicillin and streptomycin (250 µg/mL each) and amphotericin B (0.25 µg/mL) at 30^{o}C shaken at 140rpm for 5 days or until cell concentrations reached ~10^{6}-10^{7}cells/mL.

2.2 Cell Preparation

To prepare cells for flow sorting, one 25mL culture was centrifuged at 10,000rpm for 1 minute to concentrate. The bottom 7.5mL of the concentrated culture and pelleted cells were removed and added to a fresh tube with 1.125mL Galbraith's solution (Galbraith et al., 1983). The solution was vortexed briefly for 30 seconds and then passed through a 40µm filter to remove debris and washed with an additional 500µl of Galbraith's solution. 1 µl of Propidium Iodide (PI) (1mg/mL) was then added per mL of solution to stain the nuclei. Here, a 10µl sample was observed under a fluorescent microscope at 535nm (40x) to confirm staining with PI. Cells were then lysed with a tight pestle Dounce homogenizer for 15 turns. A 10µl sample of the homogenate was again observed under a fluorescent microscope at 535nm (40x) to ensure the micronuclei had been removed from their "cup" next to the macronuclei and that nuclei remained intact.

2.3 Flow Cytometry and Cell Sorting

Stained samples were sorted on a BD Biosciences FACS (Fluorescence-Activated Cell Sorting) Aria IIu (San Jose, CA) utilizing a 100µm nozzle and Beckman Coulter IsoFlow Sheath Fluid (Brea, CA) with a sheath pressure of 20 psi. Forward Scatter (FSC) and Side Scatter (SSC) were measured using standard filters off the 488 nm laser for approx. 1 hour. Propidium iodide (PI) was excited and measured off the 561nm laser. Nuclei were sorted into MAC-enriched and MIC-enriched samples based on FSC, SSC, and intensity of PI signals (Figure 2) into 1.5mL microcentrifuge tubes containing phosphate buffered saline (PBS). MIC and MAC are primarily distinguished by the PI signal. Data were collected using BD FACS Diva 8.0.1 software (San Jose, CA). Data were analyzed with FlowJo v10.6 (BD Biosciences, San Jose CA).

2.4 Genomic DNA extraction and sequencing

After sorting, genomic DNA was extracted from both samples using phenol-chloroform, following a protocol provided by Pacific Biosciences (http://www.pacb.com). Samples were concentrated to 30µl at 11.2 and 10.5 ng/µl for the MIC and MAC, respectively. Paired-end sequencing was performed on the Illumina MiSeq Nano V2 platform (250 cycles) at the DNASU core facility at the Biodesign Institute at Arizona State University. Samples were multiplexed with the final number of reads per sample being 1,048,024 reads for the MAC FACS sample and 904,282 reads for the MIC FACS sample. Genomic DNA for whole cell control samples was extracted by phenol-chloroform, and one paired-end library (average insert size 280bp) was generated on the Apollo 384 liquid handler using KAPA Biosystem's LTP library preparation kit (KK8232) following the manufacturer's instructions. Sequencing was performed on the Illumina NEXseq platform (150 cycles) at the DNASU core facility at the Biodesign Institute at Arizona State University. The final number of reads for the whole cell control sample was 132,788,552. Sequencing reads are available from the NCBI's SRA database under a BioProject with accession number PRJNA735576.

2.5 Bioinformatic Analyses

Overview

The goal of the following bioinformatic analyses is to quantify the enrichment of micronuclei. Our first analysis compared the proportion of uniquely mapped reads to the MIC vs. MAC reference genomes for each FACS sample to a whole cell sample. Unique regions in the MIC reference are IESs, while unique regions in the MAC reference are IES excision boundaries (Figure 3). The MIC-enriched FACS sample is expected to have a greater proportion of uniquely MIC-mapped reads than the whole cell data. To estimate contamination from the opposite genome in the FACS samples, we compared the proportion of uniquely mapped reads to the MIC vs. MAC reference genomes in simulated MIC and MAC reads relative to simulated whole cell reads. The simulated whole cell data also addresses any potential biases in the original whole cell sample, as we would expect the proportion of uniquely mapped reads to the MIC vs. MAC reference genomes for both the simulated and actual whole cell reads to be similar.

In our second analysis we compared unique and shared regions between the MIC and MAC through coverage levels of IESs, which are unique to the MIC, and Macronuclear-Destined Sequences (MDSs), which are found in both MIC and MAC, per FACS sample.

For our final analysis we modified an established method of validating flow sorting, IES retention scores (IRSs), for use in *Tetrahymena*. The retention score of an IES is given by the equation: IRS=IES+/(IES+ + IES−) (Swart et al., 2014), where IES+ represents the number of reads that contain the IES sequence and IES- represents the number of reads that contain the MAC excision boundary of the corresponding IES. Originally developed for *Paramecium*, the IRS- score can be easily calculated by counting the number of reads that contain a conserved TA dinucleotide found at the ends of Parameicum IESs (Arnaiz et al., 2012; Gratias and Bétermier, 2003). However, as *Tetrahymena* exhibits sequence diversity at IES excision sites, this method can not be exactly replicated. To modify the IES retention scores for *Tetrahymena*, we calculated the IES- as the number of reads that contain the MAC excision boundary of the corresponding

IES, which we determined as the regions in the MAC reference genome immediately adjacent to a known IES region in a pairwise alignment MIC/MAC chain file generated by the software transanno (github.com/informationsea/transanno) and minimap2 (Li 2018). Only viable IESs were used for this analysis, which included those that are within 10bps of the adjacent MAC scaffolds in a MIC/MAC chain file. IESs that overlap with MAC scaffolds were discarded.

Fisher's exact tests of uniquely mapped reads

To validate enrichment of micronuclei by flow sorting, reads from the MIC-enriched FACS and MAC-enriched FACS samples were trimmed with trimmomatic v0.38 (Bolger et al. 2014) and aligned to a combined MIC (http://datacommons.cyverse.org/browse/iplant/home/rcoyne/public/tetrahymena/MIC), MAC (*Tetrahymena* Genome Database http://ciliate.org/), and mitochondrial (NCBI Reference Sequence: NC_003029.1) reference using BWA mem v0.7.12 (Li and Durbin 2010). The sequenced MIC genome (Hamilton et al. 2016) consists of 5 chromosomes totaling 157Mb, and was generated using Illumina whole genome shotgun sequencing with PCR-free fragment libraries. The micronuclei in Hamilton et al. (2016) were isolated using differential sedimentation as described in Gorovsky et al. (1975). The MAC genome sequence (Sheng et al. 2020) is made up of 181 chromosomes capped with two telomeres with all gaps entirely closed (103.3Mb), and was sequenced using 300× long Single Molecule, Real-Time reads. The macronuclei in Sheng et al. (2020) were isolated using a modified differential centrifugation protocol described in Chen et al. (2016). The 47,577 bp mitochondrial genome of *T. thermophila* has also been sequenced (NCBI Reference Sequence: NC_003029.1) (Brunk et al. 2003). All unmapped reads were removed from the BAM files of each dataset using SAMtools v1.10 (Li et al. 2009). Additionally, all secondary and chimeric alignments, which could indicate mapping to a region shared between the two genomes, were also removed. The number of reads from each BAM file that aligned uniquely to a region in the MIC or MAC reference genome were then calculated using SAMtools. As a control, whole cell reads, equivalent in read number to the MIC-enriched and MAC-enriched FACS samples, were randomly sampled from a previous sequencing run (BioProject PRJNA735576) using seqtk (Shen et al., 2016). The proportion of uniquely mapped reads from the whole cell data to each reference genome was also calculated using SAMtools. Two Fisher's exact tests were performed using the whole cell reads compared with the MIC-enriched FACS sample reads and the whole cells reads compared with the MAC-enriched FACS sample reads.

Simulated MIC-enriched FACS, and MAC-enriched FACS, and whole cell reads

To estimate contamination from the opposite genome in the FACS samples, we compared the proportion of uniquely mapped reads to the MIC vs. MAC reference genomes in simulated MIC and MAC reads relative to simulated whole cell reads. Simulated MIC-enriched FACS and MAC-enriched FACS reads were sampled from their respective reference genomes using ART (v1.5.0) (Huang et al., 2012) and aligned to the combined MIC, MAC, and mitochondrial reference using BWA mem v0.7.12 (Li and Durbin 2010). Simulated 250-bp paired-end whole cells reads were created also using ART (v1.5.0). DNA fragment size and standard deviation were estimated from the whole cell sample reads using deepTools bamPEFragmentSize (Ramírez et al., 2014). While the usual whole cell ploidy of *T. thermophila* is 2:45 MIC:MAC, considering we used an asynchronized population of cells in our experiments the MAC is estimated to be at an average of 64x (between 45x in the G1 phase and 90x in the G2 phase) (Woodard et al., 1972), while the MIC is essentially 4x with no apparent G1 phase (Cole and Sugai, 2012). Therefore, we simulated the whole cell ploidy by sampling the MIC reference at 4x coverage and the MAC reference at 64x coverage. Simulated reads were aligned to the combined MIC, MAC, and mitochondrial reference using BWA mem v0.7.12 (Li and Durbin 2010). The simulated whole cell data also addresses any potential biases in the original whole cell sample, as we would expect the proportion of uniquely mapped reads to the MIC vs. MAC reference genomes for both the simulated and actual whole cell reads to be similar.

Mean read depth of IESs and MDSs

To determine coverage levels of IES, which are unique to the MIC, and Macronuclear-Destined Sequences (MDSs), which are found in both MIC and MAC, per FACS sample, we used the locations of IESs in the MIC supercontigs from Hamilton et al. (2016), Supplementary file 1C and Supplementary file 3A. The depth of coverage for the IESs and MDSs was calculated using Samtools depth (Li et al. 2009) for each sample.

IES retention scores

We further verified enrichment of micronuclei by calculating a retention score for each viable IES. Viable IESs include those that are within 10bps of the adjacent MAC scaffolds in a MIC/MAC chain file created using the software transanno (github.com/informationsea/transanno) and minimap2 (Li 2018). IESs that overlap with MAC scaffolds were discarded. MIC-enriched and MAC-enriched FACS data were aligned to a reference sequence consisting of the MAC reference and the sequence of each individual IES using BWA mem v0.7.12 (Li and Durbin 2010). IES retention scores (IRSs) were determined for each IES by counting the number of reads that contain the IES sequence (IES+) and the number of reads that contain the MAC excision boundary of the corresponding IES (IES-). The retention score of an IES is given by the equation: IRS=IES+/(IES+ + IES−) (Swart et al., 2014).

3. Results and Discussion

3.1 Flow-cytometric assay

The MIC and MAC of *T. thermophila* are distinct in both the physical size of their nuclei (~3 µm and ~10-15µm in diameter, respectively) (Figure 4) as well as their genome size (a diploid 157 Mb genome and a 45x ploid 103 Mb genome, respectively). Therefore, we FACS-sorted the nuclei based on forward scatter (FSC) and side scatter (SSC) values as well as the intensity of the PI signal. The MIC fraction of the sample represented 4.87% of the total events of the flow sorting and 76.55% of the nuclear sample (circled points in Fig. 2). The MAC fraction of the sample represented 1.46% of the total events of the flow sorting and 23.44% of the nuclear sample (Figure 2). We then validated the FACS enrichment of MIC through comparisons of uniquely mapped reads, mean coverage depth of IESs and MDSs, and IES retention scores.

3.2 Percentage of uniquely mapped reads

The proportion of uniquely mapped reads (i.e., reads that aligned solely to either the MIC or MAC reference in the combined MIC, MAC, and mitochondrial reference) for the whole cell sample was 0.36 MIC and 0.64 MAC (Table 1). The 36/64 proportion of uniquely mapped reads to the MIC and MAC references represents the baseline proportion that the MIC and MAC FACS data were compared to in order to validate enrichment.

Using the whole cell proportion of uniquely mapped reads to the MIC and MAC references as a baseline, we observe that there is clear enrichment for MIC sequences in the MIC-enriched FACS sample based on the sample's 83/17 proportion of uniquely mapped reads (Fisher's exact test, $p < 0.001$). There is also enrichment for MAC sequences in the MAC-enriched FACS sample based on the sample's proportion of 14/86 uniquely mapped reads to the MIC and MAC references (Fisher's exact test, $p < 0.001$).

3.3 Simulated whole cell, MIC-enriched FACS, and MAC-enriched FACS reads

As there are far more unique regions in the MIC in the form of IESs (approx. 54 Mb) compared to the unique regions of the MAC (IES excision junctions), we found it surprising that there was a higher proportion of uniquely mapped reads to the MAC reference from the whole cell sample in our baseline proportion (36/64 MIC/MAC). To understand the 36/64 baseline proportion of uniquely mapped reads to the MIC and MAC references from the whole cell samples, we simulated whole cell reads using ART (v1.5.0) (Huang et al., 2012) to reflect the 4:64 MIC:MAC ploidy of a *T. thermophila* cell. After alignment to the combined MIC, MAC, mitochondrial, and rDNA chromosome reference using BWA mem v0.7.12 (Li and Durbin 2010), the proportion of uniquely mapped reads to the MIC and MAC references in the simulated whole cell reads was 34/66 (Table 1) compared to the 36/64 (Table 1) of the original whole cell samples, confirming the original proportion as an accurate baseline measurement (Fisher's exact test, $p=1$). To investigate further why the number of uniquely mapped reads to the MAC reference was higher than the uniquely mapped reads to the MIC reference in the whole cell samples, despite all of the MAC's genomic content stemming from the MIC with the exception of IES excision sites, we again simulated whole cell reads but at a 1:1 MIC:MAC ploidy. Table 1 illustrates that the majority of uniquely mapped reads after alignment to the combined reference genome do originate from the MIC (90%), likely due to the thousands of IESs present in the MIC which are absent in the MAC. The high number of uniquely mapped reads to the MAC reference in normal whole cell samples is attributed to the high ploidy of the MAC. The number of reads mapping uniquely to the MIC and MAC references from the 1:1 MIC:MAC ploidy simulated whole cell reads (146028:17558) when multiplied by the 4:64 MIC:MAC ploidy, equal a proportion of 34/66, identical to the 34/66 proportion of the 4:64 MIC:MAC ploidy simulated whole cell reads.

Finally, we also simulated reads generated only from the MIC reference and reads generated only from the MAC reference to estimate contamination from the opposite genome in the FACS samples. The difference between the proportion of uniquely mapped reads to the MIC and MAC references in the simulated MIC-enriched FACS samples (99.93/0.03, Table 1) and the proportion of the actual MIC-enriched FACS sample (83/17, Table 1) indicates there is contamination from the MAC. Possibly, this is due to the degradation of the MAC observed during sorting or damage to the MAC during homogenization. Modifying the atmospheric conditions in the flow sorter to adjust the pH can also be explored as a means to limit cross contamination of nuclei due to the degrading MAC, as pH has been previously implicated in reducing cell viability (Cossarizza et al. 2017). We observed a similar level of MIC contamination in the MAC-enriched FACS sample based on the proportion of uniquely mapped reads to the MIC and MAC references in the simulated MAC-enriched FACS samples (0.05/99.5, Table 1) compared to the proportion of the actual MAC-enriched FACS sample (14/86, Table 1). Compared to previously published studies utilizing differential centrifugation (contamination of MAC ranging from 1-3%, Xiong et. al 2015; Xiong et. al 2016), our MIC-MAC fraction cross contamination is significantly higher. This could be due to the length of time the cultures were grown, as the MIC is more tightly attached to the MAC of stationary-phased cells which can reduce the purity of the fractions (Gorovsky et al., 1975). For MIC purification it is ideal that cell density does not exceed 2.5×105 cells mL−1 (mid-log phase) (Chen et al. 2016), therefore we might expect better results from flow cytometry had cultures been prevented from growing to such high density. However, for many purposes, e.g. mutation detection, absolute purity of the FACS samples is not required, as expectations of read origin and the likelihood of *de novo* mutations can be adjusted based on the estimated level of contamination and further supported by sequencing of both fractions. Beyond measuring the proportion of uniquely mapped reads, MIC purity could also be measured by a second round of flow cytometry, comparing the percentage of MIC in the total nuclear sample after the first and second round of sorting as done in Guérin et al. (2017).

3.4 Mean coverage depth of IESs and MDSs

For the second validation test we calculated sequencing coverage depth in MDSs vs. IESs for the whole cell sample, MIC-enriched FACS sample, and MAC-enriched FACS sample (Table 2). This analysis compares unique and shared regions between the MIC and MAC as opposed to only unique regions in the first validation test. For the whole cell sample we would expect an IES:MDS ratio of 4:68 or 0.06:1, as IESs occur only in the diploid MIC (4x in G1 phase) while Mac-destined regions occur in the MAC at 64x coverage (between 45x in the G1 phase and 90x in the G2 phase) and in the MIC (4 + 64 = 68). Our actual ratio in the whole cell sample may vary based on the number of IESs that are accidentally maintained in the MAC during genome rearrangement or missing from the data due to insufficient sequencing depth. Using the whole cell ratios as a baseline, we observed that in the MIC FACS data there was an IES:MDS read depth ratio of 0.45:1 which shows enrichment in the MIC. While we would expect this ratio to be 1:1 as all regions of the MIC should be diploid, there could be MAC contamination or IESs could be located in poorly assembled regions in the reference. For the MAC-enriched FACS sample there was minimal IES coverage with a IES:MDS coverage ratio of 0.084:1, as all IESs should be eliminated from the MAC. MIC contamination and IES retention are possible sources of sequencing coverage in IES regions.

In our FACS samples there was also evidence of bacterial contamination. 71% of the total number of reads in the MIC-enriched FACS sample and 34% of the total number of reads in the MAC-enriched FACS sample did not map to the combined MIC, MAC, and mitochondrial reference, compared with only 1% of reads in the whole cell data when mapped to the combined MIC, MAC, and mitochondrial reference, which was from a separate experiment with deeper sequencing. The high percentage of unmapped reads in the MIC-enriched and MAC-enriched FACS samples prompted us to BLAST the unmapped reads from each sample to explore potential sources of contamination. From the blast results we found that the major source of contamination was bacterial with a mix of *Azoarcus*, *Acidovorax*, *Alicycliphilus*, *Diaphorobacter*, *Alicycliphilus,* and *Pseudomonas*. Contamination could have occurred in cultures, during flow sorting, or during DNA extraction. While contamination should not affect the IES/MDS measures presented here, high contamination levels would be a considerable drawback for costlier long-read sequencing. Bacterial contamination can be limited by filtering with a <5µm filter before flow sorting as well incorporating the antibiotics neomycin, kanamycin, or tetracycline (100 µg/mL each) into the cell culture as suggested by Cassidy-Hanley (2012).

Also important for utilizing flow sorting for long read sequencing is the quality of DNA. While the quality of the DNA collected from the MIC and MAC FACS samples (shearing, degradation) was not collected for this study it can be obtained in future experiments using a Bioanalyzer. Suggested methods of maintaining DNA quality during flow sorting include increasing the surface area of single cell suspensions in order to maximize contact between cells and digestive enzymes (Reichard and Asosingh 2018) and sanitizing all equipment to prevent degradation by contaminating nucleases (Ormerod and Imrie 1990).

3.5 IES retention scores

The final metric used to validate MIC-enrichment from flow sorting was the IES retention scores (IRS) of viable IESs. Viable IESs included those that are within 10bps of the adjacent MAC scaffolds in a MIC/MAC chain file. IESs that overlap with MAC scaffolds were discarded. The fraction of viable IESs to the potential total number of IESs (unverified) from Hamilton et al. (2016) are included in Table 3. The location of the IESs in the fully assembled MIC chromosomes were extrapolated from the position of the IESs in the MIC scaffolds (Supplementary file 3A) and the position of the MIC scaffolds within the MIC chromosomes (Supplementary file 1C). As some MIC scaffolds were incorporated multiple times into the MIC chromosome assemblies there are repeats of identified IESs, bringing the original total of 7551 IESs from Hamilton et al. (2016) to 8171. For the MIC-enriched FACS sample if sorting were perfect (and IES excision were perfect) the expected IRS is 1, as there should be no short reads that span both the left and right excision boundary while there will be reads that map to the IES itself. The MAC-enriched FACS sample IRS should be 0, as there will be reads that span both the left and right excision boundary after the IES is removed and no reads that map uniquely to any IES. Table 3 shows the average IRS for the MIC-enriched FACS data per chromosome as closer to 1 than the MAC IRSs. The IRSs of the MIC-enriched FACS sample are also demonstrated to skew towards 1 in Figure 5 (blue) while the MAC-enriched FACS sample IRSs skew towards 0 (red). This indicates that there is enrichment for MIC DNA in the MIC-enriched FACS sample and limited MIC contamination in the MAC-enriched FACS sample. The not insignificant fraction of 0 scores for the MIC-enriched FACS sample (1068) could be from IESs identified in the Hamilton et al. (2016) study that do not actually exist or were not present in our sample. While the IESs used in this study are classified as "high-confidence" after verification from at least two different identification methods (MAC read alignment to MIC reference, MIC read alignment to MAC reference, and MIC-MAC cross-assembly alignment), there is still a degree of uncertainty regarding the location of the identified IESs. This is due to the repetitive sequences within the IESs and minor contamination of the MIC sequencing libraries with MAC DNA during the assembly in Hamilton et al. (2016), which could create a mixture of inconsistent short reads at IES/MDS junctions creating false breaks in the assembly that are not the result of IESs. Additionally, research suggests (Feng et al. 2017; Jaspan et al. 2019) that the excision boundaries of IESs exhibit variability during conjugation, which could account for the IRS scores in the MIC-enriched FACS sample that are closer to 0 than 1 if the predicted IES/MDS junction is incorrect due to excision variability. This validation technique could be improved through the use of de novo IES detection in our own samples as opposed to using previously published IES locations.

Conclusions and Future Applications

The use of flow sorting that we describe here to enrich for micronuclei in *T. thermophila* is simple and requires only 25mL of culture compared to the liters required for centrifugation methods. Our validation approaches, (a) Fisher's exact tests of uniquely mapped reads to the micro and macronuclear reference genomes (b) mean coverage depth of IES and Macronuclear-destined regions after alignment to the micronuclear reference genome and (c) IES retention scores, all support MIC-enrichment from flow sorting (approximately 9x more MIC DNA compared to whole cell sequencing based on validation approach (a)), but also suggest cross contamination of nuclei between the MIC and MAC FACS samples. To improve MIC purification in future iterations of flow sorting, cell cultures can be grown to no greater than mid-log phase to prevent MICs from tightly attaching to the MAC.

While the quality of the DNA collected from the MIC and MAC FACS samples (shearing, degradation) was not collected for this study it can be obtained in future experiments using a Bioanalyzer. Suggested methods of maintaining DNA quality during flow sorting include increasing the surface area of single cell suspensions in order to maximize contact between cells and digestive enzymes (Reichard and Asosingh 2018) and sanitizing all equipment to prevent degradation by contaminating nucleases (Ormerod and Imrie 1990).

This method will allow for rapid sequencing of MIC-enriched DNA for comparison with existing MAC sequences to help elucidate the evolution and molecular mechanisms of genome rearrangement in ciliates. Additionally, flow sorting allows for more detailed mutation detection after sequencing than genomic exclusion. Using sequencing data from flow-sorted micronuclei, we will be able to characterize complex mutations events (insertions, deletions, and copy number variants) in *Tetrahymena* after mutation accumulation, which have previously been difficult to detect bioinformatically after whole cell sequencing. Further, this method can be used to sequence the MIC in species that are unable to be grown to sufficient volumes for centrifugation, for example species of Karyorelicteans, which are notoriously difficult to culture.

Data Availability

All scripts used in the bioinformatics analysis presented here are available at https://github.com/aahowel3/An-Alternative-Method-to-Enrich-Tetrahymena-Micronuclear-DNA. Sequencing reads are available from the NCBI's SRA database under a BioProject with accession number PRJNA735576.

Acknowledgements

We would like to acknowledge Joy Blain and the personnel from the Genomics KED core facility at Arizona State University for their assistance in DNA library preparation and sequencing. We would also like to thank the editor and reviewers for their helpful suggestions which improved both the analysis and discussion of this manuscript.

Literature Cited

Allen, S. L. (1963). Genomic Exclusion in Tetrahymena: Genetic Basis*. *The Journal of Protozoology*, *10*(4), 413–420. https://doi.org/10.1111/j.1550-7408.1963.tb01699.x

Allen, S. L., & Nanney, D. L. (1958). An Analysis of Nuclear Differentiation in the Selfers of Tetrahymena. *The American Naturalist*, *92*(864), 139–160. https://doi.org/10.1086/282022

Bolger, A. M., Lohse, M., & Usadel, B. (2014). *Genome analysis Trimmomatic: a flexible trimmer for Illumina sequence data*. *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brownell, J. E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D. G., Roth, S. Y., & Allis, C. D. (1996). Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell*, *84*(6), 843–851. https://doi.org/10.1016/S0092-8674(00)81063-6

Cassidy-Hanley, D. M. (2012). Tetrahymena in the Laboratory: Strain Resources, Methods for Culture, Maintenance, and Storage. *Methods in Cell Biology*, *109*, 237. https://doi.org/10.1016/B978-0-12-385967-9.00008-6

Chen, X., Gao, S., Liu, Y., Wang, Y., Wang, Y., & Song, W. (2016). Enzymatic and chemical mapping of nucleosome distribution in purified micro- and macronuclei of the ciliated model organism, Tetrahymena thermophila. *Science China Life Sciences*, *59*(9), 909–919. https://doi.org/10.1007/S11427-016-5102-X

Doerder, F. P., Lief, J. H., & Doerder, L. E. (1975). A corrected table for macronuclear assortment in Tetrahymena pyriformis, syngen 1. *Genetics*, *80*(2), 263–265. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1213326/

Feng, L., Wang, G., Hamilton, E. P., Xiong, J., Yan, G., Chen, K., Chen, X., Dui, W., Plemens, A., Khadr, L., Dhanekula, A., Juma, M., Dang, H. Q., Kapler, G. M., Orias, E., Miao, W., & Liu, Y. (2017). A germline-limited piggyBac transposase gene is required for precise excision in Tetrahymena genome rearrangement. *Nucleic Acids Research*, *45*(16), 9481–9502. https://doi.org/10.1093/nar/gkx652

Fjerdingstad, E. J., Schtickzelle, N., Manhes, P., Gutierrez, A., & Clobert, J. (2007). Evolution of dispersal and life history strategies - Tetrahymena ciliates. *BMC Evolutionary Biology*, *7*(1), 133. https://doi.org/10.1186/1471-2148-7-133

Galbraith, D. W., Harkins, K. R., Maddox, J. M., Ayres, N. M., Sharma, D. P., & Firoozabady, E. (1983). Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science*, *220*(4601), 1049–1051. https://doi.org/10.1126/science.220.4601.1049

Gibbons, I. R., & Rowe, A. J. (1965). Dynein: A protein with adenosine triphosphatase activity from cilia. *Science*, *149*(3682), 424–426. https://doi.org/10.1126/science.149.3682.424

Gorovsky, M. A., Yao, M. C., Keevert, J. B., & Pleger, G. L. (1975). Isolation of micro- and macronuclei of Tetrahymena pyriformis. *Methods in Cell Biology*, *9*(0), 311–327. https://doi.org/10.1016/S0091-679X(08)60080-1

Greider, C. W., & Blackburn, E. H. (1985). Identification of a Specific Telomere Terminal Transferase Activity in Tetrahymena Extracts. In *Cell* (Vol. 43).

Guérin, F., Arnaiz, O., Boggetto, N., Denby Wilkes, C., Meyer, E., Sperling, L., & Duharcourt, S. (2017). Flow cytometry sorting of nuclei enables the first global characterization of Paramecium germline DNA and transposable elements. *BMC Genomics*, *18*(1), 327. https://doi.org/10.1186/s12864-017-3713-7

Hai, B., Gaertig, J., & Gorovsky, M. A. (2000). Knockout heterokaryons enable facile mutagenic analysis of essential genes in Tetrahymena. *Methods in Cell Biology*, *62*(62), 513–531. https://doi.org/10.1016/s0091-679x(08)61554-x

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–594. https://doi.org/10.1093/bioinformatics/btr708

Jaspan, V. N., Taye, M. E., Carle, C. M., Chung, J. J., & Chalker, D. L. (2019). Boundaries of eliminated heterochromatin of Tetrahymena are positioned by the DNA-binding protein Ltl1. *Nucleic Acids Research*, *47*(14), 7348–7362. https://doi.org/10.1093/nar/gkz504

Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., & Cech, T. R. (1982). Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, *31*(1), 147–157. https://doi.org/10.1016/0092-8674(82)90414-7

Lee, P.-H., Meng, X., & Kapler, G. M. (2015). Developmental Regulation of the Tetrahymena thermophila Origin Recognition Complex. *PLoS Genetics*, *11*(1), e1004875. https://doi.org/10.1371/journal.pgen.1004875

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/BIOINFORMATICS/BTY191

Li, H., & Durbin, R. (2010). *Fast and accurate long-read alignment with Burrows-Wheeler transform*. *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Long, H. A., Paixão, T., Azevedo, R. B. R., & Zufall, R. A. (2013). Accumulation of spontaneous mutations in the ciliate Tetrahymena thermophila. *Genetics*, *195*(2), 527–540. https://doi.org/10.1534/genetics.113.153536

Long, H., Winter, D. J., Chang, A. Y. C., Sung, W., Wu, S. H., Balboa, M., Azevedo, R. B. R., Cartwright, R. A., Lynch, M., & Zufall, R. A. (2016). Low base-substitution mutation rate in the germline genome of the ciliate tetrahymena thermophila. *Genome Biology and Evolution*, *8*(12), 3629–3639. https://doi.org/10.1093/gbe/evw223

Long, H., Winter, D. J., Y-C Chang, A., Sung, W., Wu, S. H., Balboa, M., R Azevedo, R. B., Cartwright, R. A., Lynch, M., & Zufall, R. A. (n.d.). *Low Base-Substitution Mutation Rate in the Germline Genome of the Ciliate Tetrahymena thermophila*. https://doi.org/10.1093/gbe/evw223

Merriam, E. V., & Bruns, P. J. (1988). Phenotypic assortment in Tetrahymena thermophila: assortment kinetics of antibiotic-resistance markers, tsA, death, and the highly amplified rDNA locus. *Genetics*, *120*(2), 389–395. /pmc/articles/PMC1203518/?report=abstract

Nanney, D. L., & Preparata, R. M. (1979). Genetic Evidence Concerning the Structure of the Tetrahymena thermophila Macronucleus*†. *The Journal of Protozoology*, *26*(1), 2–9. https://doi.org/10.1111/j.1550-7408.1979.tb02722.x

Orias, E., & Flacks, M. (1975). Macronuclear genetics of Tetrahymena. I. Random distribution of macronuclear gene copies in T. pyriformis, syngen 1. *Genetics*, *79*(2), 187–206. /pmc/articles/PMC1213266/?report=abstract

Ormerod, M. G., & Imrie, P. R. (1990). Flow cytometry. *Methods in Molecular Biology*, *5*, 543–558.

Papazyan, R., Voronina, E., Chapman, J. R., Luperchio, T. R., Gilbert, T. M., Meier, E., Mackintosh, S. G., Shabanowitz, J., Tackett, A. J., Reddy, K. L., Coyne, R. S., Hunt, D. F., Liu, Y., & Taverna, S. D. (2014). Methylation of histone H3K23 blocks DNA damage in pericentric heterochromatin during meiosis. *ELife*, *2014*(3). https://doi.org/10.7554/eLife.02996

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, *42*(W1), W187. https://doi.org/10.1093/nar/gku365

Reichard, A., & Asosingh, K. (2019). Best Practices for Preparing a Single Cell Suspension from Solid Tissues for Flow Cytometry. *Cytometry Part A*, *95*(2), 219–226. https://doi.org/10.1002/CYTO.A.23690

Ruehle, M. D., Orias, E., & Pearson, C. G. (2016). Tetrahymena as a unicellular model eukaryote: Genetic and genomic tools. In *Genetics* (Vol. 203, Issue 2, pp. 649–665). Genetics. https://doi.org/10.1534/genetics.114.169748

Schoeberl, U. E., Kurth, H. M., Noto, T., & Mochizuki, K. (2012). Biased transcription and selective degradation of small RNAs shape the pattern of DNA elimination in Tetrahymena. *Genes and Development*, *26*(15), 1729–1742. https://doi.org/10.1101/gad.196493.112

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Sheng, Y., Duan, L., Cheng, T., Qiao, Y., Stover, N. A., & Gao, S. (2020). The completed macronuclear genome of a model ciliate Tetrahymena thermophila and its application in genome scrambling and copy number analyses. *Science China. Life Sciences*, *63*(10), 1534. https://doi.org/10.1007/S11427-020-1689-4

Swart, E. C., Wilkes, C. D., Sandoval, P. Y., Arambasic, M., Sperling, L., & Nowacki, M. (2014). Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion. *Nucleic Acids Research*, *42*(14), 8970–8983. https://doi.org/10.1093/nar/gku619

Sweet, M. T., & Allis, C. D. (2006). Isolation and Purification of Tetrahymena Nuclei. *Cold Spring Harbor Protocols*, *2006*(23), pdb.prot4500-pdb.prot4500. https://doi.org/10.1101/pdb.prot4500

Taverna, S. D., Coyne, R. S., & Allis, C. D. (2002). Methylation of histone H3 at lysine 9 targets programmed DNA elimination in Tetrahymena. *Cell*, *110*(6), 701–711. https://doi.org/10.1016/S0092-8674(02)00941-8

Xiong, J., Gao, S., Dui, W., Yang, W., Chen, X., Taverna, S. D., Pearlman, R. E., Ashlock, W., Miao, W., & Liu, Y. (2015). Distinct nucleosome distribution patterns in two structurally and functionally differentiated nuclei of a unicellular eukaryote. *BioRxiv*, 018754. https://doi.org/10.1101/018754

Xiong, J., Gao, S., Dui, W., Yang, W., Chen, X., Taverna, S. D., Pearlman, R. E., Ashlock, W., Miao, W., & Liu, Y. (2016). Dissecting relative contributions of cis- and trans-determinants to nucleosome distribution by comparing Tetrahymena macronuclear and micronuclear chromatin. *Nucleic Acids Research*, *44*(21), 10091–10105. https://doi.org/10.1093/NAR/GKW684

Yao, M. C., Zheng, K., & Yao, C. H. (1987). A conserved nucleotide sequence at the sites of developmentally regulated chromosomal breakage in tetrahymena. *Cell*, *48*(5), 779–788. https://doi.org/10.1016/0092-8674(87)90075-4

Figure Legends

Figure 1. Flowchart summarizing the isolation of *Tetrahymena thermophila* MIC and MAC (see text for details).

Figure 2. MIC and MAC are primarily distinguished by the PI signal. This flow cytometry dot plot demonstrates the profiles for MIC and MAC nuclei of *Tetrahymena thermophila* after staining with propidium iodide. The y-axis represents an autofluorescent signal generated in the blank channel and the x-axis represents the fluorescent signal of PI stained samples. Signals in the lower left hand corner of the dot plot outside the gated MIC and MAC signals (circled) represent debris in the samples.

Figure 3. **Figure 3**. MIC enrichment can be quantified through uniquely mapped reads. A. Schematic depiction of a *Tetrahymena* cell, with micronucleus indicated by red fill and macronucleus indicated by green fill. B. Alignment of reads to the MIC (top) and MAC (bottom) reference genomes. Reads derived from internally eliminated sequences (IES) map uniquely to the MIC genome (pink regions) while reads that span the MAC excision boundary of an IES map uniquely to the MAC genome.

Figure 4. The MIC and MAC of *Tetrahymena thermophila* are distinct in both physical size of their nuclei as well as their genome size. This figure shows whole cell *Tetrahymena thermophila* stained MIC and MAC (left) and MIC and MAC after homogenization (right).

Figure 5. Histogram of IES Retention scores for the MIC-enriched FACS Sample (blue) and MAC-enriched FACS Sample (red). IES Retention scores for the Micronuclear FACS Sample skew towards 1, indicating there are a high number of IESs in the sample. IES Retention scores for the Macronuclear FACS Sample skew towards 0, indicating there are a low number of IESs in the sample.