SIGAR: Inferring features of genome architecture and DNA rearrangements by split read mapping

Authors: Yi Feng[1], Leslie Y. Beh[1], Wei-Jen Chang[2], Laura F. Landweber[1*]

1 Departments of Biochemistry and Molecular Biophysics and Biological Sciences, Columbia University, New York, NY, USA

2 Department of Biology, Hamilton College, Clinton, NY, USA

*Author for Correspondence: Laura Landweber, Departments of Biochemistry and Molecular Biophysics and Biological Sciences, Columbia University, New York, NY, USA, 212-305-3898, Laura.Landweber@columbia.edu

Abstract

Ciliates are microbial eukaryotes with distinct somatic and germline genomes. Post-zygotic development involves extensive remodeling of the germline genome to form somatic chromosomes. Ciliates therefore offer a valuable model for studying the architecture and evolution of programmed genome rearrangements. Current studies usually focus on a few model species, where rearrangement features are annotated by aligning reference germline and somatic genomes. While many high-quality somatic genomes have been assembled, a high quality germline genome assembly is difficult to obtain due to its smaller DNA content and abundance of repetitive sequences. To overcome these hurdles, we propose a new pipeline SIGAR (**S**plit-read **I**nference of **G**enome **A**rchitecture and **R**earrangements) to infer germline genome architecture and rearrangement features without a germline genome assembly, requiring only short germline DNA sequencing reads. As a proof of principle, 93% of rearrangement junctions identified by SIGAR in the ciliate *Oxytricha trifallax* were validated by the existing germline assembly. We then applied SIGAR to six diverse ciliate species without germline genome assemblies, including *Ichthyophthirius multifilii*, a fish pathogen. Despite the high level of somatic DNA contamination in each sample, SIGAR successfully inferred rearrangement junctions, short eliminated sequences and potential scrambled genes in each species. This pipeline enables pilot surveys or exploration of DNA rearrangements in species with limited DNA material access, thereby providing new insights into the evolution of chromosome rearrangements.

Key words: ciliates, local alignment, pointers, scrambled, structural variations

33 # Introduction

34 Ciliates are model organisms for studying genome rearrangement. They exhibit nuclear
35 dimorphism: each cell contains a somatic macronucleus (MAC) and a germline micronucleus
36 (MIC). The MAC consists of high-copy number chromosomes that are transcriptionally active in
37 vegetative growth. In contrast, the MIC genome is inert, and only involved in sexual conjugation.
38 After mating, a new MAC genome rearranges from a copy of the zygotic MIC, together with
39 massive DNA elimination (Chen et al. 2014; Hamilton et al. 2016). The retained, macronuclear-
40 destined sequences (MDS) must be properly ordered and oriented, and sometimes even
41 descrambled (Chen et al. 2014; Sheng et al. 2020), to form functional MAC chromosomes
42 (Figure 1A).

43 Most genome rearrangement studies focus on model organisms like *Tetrahymena*
44 (Hamilton et al. 2016), *Paramecium* (Guérin et al. 2017) and *Oxytricha* (Chen et al. 2014), which
45 possess well-assembled MIC and MAC reference genomes for annotation of DNA
46 rearrangements. Recent years have seen a bloom of *de novo* MAC genome assemblies in diverse
47 ciliate species, including *Stentor* (Slabodnick et al. 2017), *Euplotes* (Wang et al. 2016; Chen et
48 al. 2019), hypotrich ciliates (Chen et al. 2015) and *Tetrahymena* genus species (Xiong et al.
49 2019). MAC chromosomes are generally significantly shorter than MIC chromosomes, with
50 some species exhibiting gene-sized "nanochromosomes". Thus, many high quality MAC
51 genomes have been assembled using short next-generation sequencing reads. By taking
52 advantage of third generation sequencing long reads, some MAC genomes have been assembled
53 with unprecedented completeness (Wang et al. 2020; Sheng et al. 2020), sometimes obviating the
54 need for assembly when the average read length exceeds the typical chromosome length
55 (Lindblad et al. 2019). In contrast, sequencing and assembling the long MIC genomes is

experimentally and computationally complex. Besides having Mbp scale chromosomes at much lower copy number, the MIC also contains repetitive elements and centromeric regions that can best be resolved by third generation long reads. Purification of MIC genomic DNA that is free of MAC contamination is also a challenge. The MAC to MIC DNA ratio in cells ranges from 46 to 800 (Prescott 1994), which means that only 0.1-2% of the DNA in whole cells originates from the MIC. There do exist experimental methods to separate MIC and MAC nuclei, e.g. sucrose gradient centrifugation (Chen et al. 2014) and flow cytometry (Guérin et al. 2017). However, these techniques were developed for specific model ciliates, and are not generalizable across diverse species. Moreover, some ciliates are not free-living in nature (Coyne et al. 2011), or uncultivatable in the lab. The difficulty of obtaining high quality MIC-enriched DNA from these species presents additional obstacles to understanding germline genome architecture. Single-cell techniques can be helpful to analyze germline scaffolds, but require whole genome and transcriptome assemblies (Maurer-Alcalá et al. 2018).

To overcome these challenges and provide insight into germline genome architecture in the absence of a fully-assembled MIC genome, we propose a new pipeline, SIGAR (**S**plit-read **I**nference of **G**enome **A**rchitecture and **R**earrangements) using economical short, next generation reads and MAC genome assemblies. Rather than using a MIC assembly, SIGAR takes advantage of short MIC reads whose alignment diverges at MDS-MDS junctions in MAC chromosomes (Figure 1B). Here we validate SIGAR by showing high concordance of its results with published *Oxytricha* MIC genome annotations (Burns et al. 2016). We then used SIGAR to infer rearrangement features in 5 hypotrich ciliates and *Ichthyophthirius multifiliis*, yielding novel insights into MIC genome architecture in diverse phylogenetic lineages, all without genome assembly. This new pipeline will promote the use of published datasets to reveal more cases of

79 DNA rearrangement, and offers the possibility to explore germline genome evolution in diverged
80 ciliate species and natural isolates.
81

82 # Results

83 ## SIGAR strategy

84 SIGAR infers MIC genome structure and DNA rearrangement features by identifying
85 short MIC reads that partially map to MAC chromosomes. It first removes MDS-only reads by
86 end-to-end mapping to enrich for MIC-specific reads in the dataset (Figure 1B). MIC reads that
87 pass this filter should contain DNA sequences that are eliminated during rearrangement (IES,
88 Internal Eliminated Sequence). The MIC reads covering MDS-IES breakpoints will split at MDS
89 boundaries when mapped to MAC chromosomes (Figure 1B). SIGAR verifies that split read
90 junctions indeed correspond to rearrangement breakpoints by searching within the split read for
91 short sequence motifs, called "pointers", which are microhomologous repeated sequences in the
92 MIC that are retained as a single copy in the MAC after rearrangement (Figure 1A). SIGAR is
93 able to infer more germline genome information from reads that map to two or more MDSs, even
94 partially (Figure 1B). If the two blocks map adjacently to each other in the same direction, the
95 region in between is recognized as a nonscrambled IES. Otherwise, the split read often indicates
96 the presence of a scrambled region in the MIC.
97

98 ## Validation of SIGAR by genome-assembly based annotations

99 To validate our proposed strategy to infer MIC genome features using short reads, we
100 mapped *Oxytricha* Illumina MIC reads (Chen et al. 2014) to the original MAC genome assembly
101 (Swart et al. 2013). SIGAR was then used to infer pointers from split reads (i.e. partially

102 mapping reads), and the results were compared with existing MIC genome annotations (Chen et
103 al. 2014; Burns et al. 2016). Pointer annotations in the MIC genome assembly were previously
104 generated using MIDAS (Burns et al. 2016) which compares MIC and MAC genomes using
105 BLAST. We find that 92.9% of the SIGAR-inferred pointers were validated by MIDAS. 59.70%
106 contain the same first and last nucleotide in the MIDAS annotation, 27.74% share one identical
107 boundary, and 3.8% share no boundary but have midpoints within 5bp of the MIDAS-inferred
108 pointer (Table S1). The length distribution of SIGAR-inferred pointers is similar to that of
109 MIDAS-inferred pointers, although fewer pointers were found in total (Figure 2A). In addition,
110 SIGAR inferred a small number of "cryptic" pointers, defined as repeats longer than 20bp
111 (Figure 2A) that differ from MIDAS-annotated MDS-MDS junctions. It is possible that some
112 represent short regions of paralogy.

113 SIGAR annotated half as many pointers as MIDAS (Table S1) because SIGAR was only
114 applied to uniquely mapping regions of the MAC genome, in order to minimize the rate of false
115 discovery. Such regions comprise 60.8-63.7% of the MAC genome, and contain 61.0-62.7% of
116 all pointers (see Methods). SIGAR successfully annotated 80.4-82.3% of these pointers. We
117 conclude that SIGAR recovers a large majority of pointers from the portion of the genome to
118 which it was applied, even in the absence of a reference germline genome assembly.

119 In ciliates, the MIC DNA is significantly less abundant than MAC DNA, which can make
120 it experimentally challenging to obtain pure micronuclei for DNA isolation. To test the
121 robustness of SIGAR to variation in MIC coverage, we simulated 100 bp Illumina HiSeq reads
122 from the *Oxytricha* MIC genome and calculated the number of inferred pointers. With only 5X
123 MIC coverage, SIGAR was still able to recover 41.98% of MIDAS-inferred pointers, and
124 97.53% of SIGAR-inferred pointers were also identified by MIDAS (Table S1, Figure S1).

125 SIGAR can also infer the presence of short IESs that are contained within single Illumina
126 reads (Figure 1B). Although the small read length constrains the size of IESs that can be detected
127 from single reads, we were able to infer 20,599 short IESs with maximum length of 41bp using
128 the 100bp read dataset (Figure 2B).

129 SIGAR's results demonstrate that pointers can vary between alleles. 11,462 *Oxytricha*
130 pointers validated by both MIDAS and SIGAR possess at least another pointer at the same
131 junction inferred by SIGAR with high confidence (at least 2 reads supporting each boundary)
132 (Table S2). Figure S2 highlights an example where pointer alleles correlate with SNPs in the
133 MIC reads. Allele-specific information could only be detected by SIGAR (which examines
134 individual sequencing reads) but not by assembly-based software like MIDAS, because genome
135 assemblies tend to collapse alleles. This suggests that MAC chromosomes in *Oxytricha* arise
136 from both MIC alleles, which can use different pointer sequences during MIC genome
137 rearrangements.

138

139 **SIGAR infers scrambled genome architecture**

140 Scrambled loci differ in order and/or orientation between the MIC and MAC versions
141 (Figure 1A). Programmed rearrangements therefore entail MDS inversion and/or translocation.
142 Scrambled genes have been described in most hypotrich ciliates (Chang et al. 2005; Hogan et al.
143 2001), *Chilodonella uncinata* (Katz and Kovner 2010) and some Postciliodesmatophora ciliates
144 (Maurer-Alcalá et al. 2018). A recent report also validated the presence of scrambled loci in
145 *Tetrahymena* (Sheng et al. 2020), providing further evidence that this is common to ciliate
146 genomes. By using split read mapping, we find that SIGAR can infer some scrambled
147 chromosome architectures and their associated pointers, even in the absence of a MIC genome

148 assembly. When applied to *Oxytricha* datasets, SIGAR detected 8741 *Oxytricha* pointers at
149 scrambled junctions. Figure 3 shows an example where SIGAR successfully inferred 3 out of 5
150 scrambled junctions in an *Oxytricha* MAC chromosome (Figure 3). SIGAR found 12 MIC reads
151 covering both MDS1 and MDS3, while another 12 reads covered both MDS2 and MDS4 (Figure
152 3A and Figure S3).
153

154 Inferring MIC genome architecture and rearrangement features in diverse
155 ciliate species

156 SIGAR was developed to provide insights into germline rearrangement features across ciliates
157 for which pure MIC DNA preparations are not readily available. We therefore applied SIGAR to
158 5 hypotrich genomes published in Chen et al. (2015), all of which have gene-sized MAC
159 nanochromosomes like *Oxytricha*. Importantly, these sequencing datasets were derived from
160 small numbers of whole cells, some from species that were not cultivated in the lab. We were
161 able to infer pointers in all of these species using SIGAR (Table 1). TA is the most favored
162 pointer among all 5 species. Though *Sterkiella* and *Urostyla* do not have TA as the most
163 abundant pointer, both include A, T and TA as the three most abundant. With increasing
164 evolutionary distance from *Oxytricha*, we observed more pointers containing TA as a substring.
165 The GC content of short IESs (<0.2) is significantly lower than adjacent MDS regions (~0.3),
166 consistent with surveys from other ciliate MIC genomes (Prescott 1994; Chen et al 2014). The
167 pointer length distributions are similar in these species, except for *Sterkiella,* which exhibits an
168 abundance of 5-20bp pointers (Figure S4).
169

170 **Table 1**. Rearrangement features recovered by SIGAR for 5 surveyed hypotrich ciliates

Phylogeny (Chen et al. 2015)	Species	# of pointers	Most abundant pointer	% of "TA" pointers	% of pointers with "TA" substring	# of scrambled pointers	# of IES	IES G+C%	G+C% of MAC contigs with IES	IES G+C < MDS G+C p-value (t-test)
	Oxytricha trifallax									
	Sterkiella histriomuscorum	12950	A	1.20	46.55	60	601	16	28	1e-69
	Stylonychia lemnae	2509	TA	3.99	43.76	21	945	19	31	1e-120
	Laurentiella sp.	1315	TA	5.10	48.14	19	509	14	28	1e-81
	Paraurostyla sp.	1318	TA	7.28	52.66	17	653	11	30	1e-158
	Urostyla sp.	1489	A	1.61	58.50	29	0	-	-	-

171

172 SIGAR also revealed evidence for novel scrambled loci in the 5 hypotrich ciliates (Table
173 1). Figure S5 shows the mapping view of an inferred scrambled locus in *Paraurostyla*, with
174 pointers detected on each side. Short IESs were inferred for all species, except *Urostyla* (Table 1,
175 Figure S4). Though short IESs were identified in *Urostyla* for *DNA pol α* (Chang et al. 2005)
176 and *actin I* (Hogan et al. 2001), we were unable to recover short IESs in the current dataset with
177 limited MIC DNA. We note that intergenic IESs, which are common in *Tetrahymena* (Hamilton
178 et al. 2016) cannot be detected by SIGAR, which identifies IESs adjacent to MDSs.

179 We also applied SIGAR to *Ichthyophthirius multifiliis (Ich),* an oligohymenophorean
180 ciliate related to *Tetrahymena* (Coyne et al. 2011). *Ich* lives as a parasite in fish epithelia,
181 causing "white spot" disease (Coyne et al. 2011). Furthermore, for ciliates that are hard to
182 cultivate, SIGAR offers an ideal tool to infer properties of MIC genome architecture and DNA
183 rearrangement, given the lack of availability of high quality MIC DNA preparations. Since the
184 *Ich* MAC genome is ~84.1% A+T, we required pointers to have at least 5 well-mapped reads
185 supporting each boundary. We found that the most abundant pointers are not only AT rich, but
186 also contain surprisingly long TA tandem repeats (Figure S6). The pointer length distribution

187 shows a peak at 10bp, with the 10bp pointer “TATATATATA” and “ATATATATAT” among
188 the most abundant pointers in *Ich*.
189

190 # Discussion

191 Here, we have developed a novel computational tool, SIGAR, for inferring genome
192 rearrangements and germline genome structure across diverse ciliates. A separate,
193 complementary study proposed using short reads to infer the presence of eliminated DNA during
194 genome rearrangement, but requires a high quality reference MIC genome assembly instead
195 (Zheng et al. 2020). MIC genome assemblies typically pose the greatest challenge, whereas high
196 quality MAC genome assemblies are much more accessible and also considerably less expensive
197 to produce. In this sense, SIGAR will be more broadly applicable to the study of DNA
198 rearrangements in ciliates.
199 SIGAR has provided new insights into the architecture of several ciliate MIC genomes,
200 including *Ichthyophthirius multifiliis* and a group of early diverged hypotrichs, relative to
201 *Oxytricha*. We observe a widespread preference for TA pointers across all ciliates in this survey.
202 Studies of the ciliate model systems *Paramecium* and *Euplotes crassus* revealed the exclusive
203 use of TA pointers in their germline genomes (Klobutcher and Herrick 1995; Arnaiz et al. 2012).
204 It has also been shown that the terminal consensus sequences in *Paramecium* and *Euplotes* IES
205 resemble terminal sequences in Tc1/mariner transposons (Klobutcher and Herrick 1995). This
206 and other observations contributed to the hypothesis that an ancestral wave of transposons
207 invaded ciliate germline genomes. The transposons then decayed but preserved the flanking
208 “TA” as a relic and a modern requirement for accurate DNA elimination. Curiously, instead of
209 this commonly observed 2bp TA pointer (Klobutcher and Herrick 1997, Chen et al. 2014), long

210 TA repeats are present at *Ich* rearrangement junctions (Figure S6). Given that pointer sequences
211 may help recruit enzymes that mediate IES removal and are necessary for excision in
212 *Paramecium* (Aury et al. 2006), it is plausible that their extended length constitutes an adaptive
213 feature to improve recognition and recruitment of DNA binding proteins that participate in
214 genome rearrangement, amidst an AT-rich genome.

215 In addition to ciliates, many organisms in nature exhibit programmed genome
216 rearrangement, such as lampreys (Smith et al. 2018) and songbirds (Kinsella et al. 2019).
217 Furthermore, aberrant structural variations in mammalian cells are frequently observed in
218 diseases like cancer (Stankiewicz and Lupski 2010; Forment et al. 2012). We expect SIGAR to
219 be directly applicable to a wide range of genomes that exhibit rearrangements in both healthy
220 and diseased states. SIGAR, which only requires one reference genome and short reads from a
221 rearranged genome, could be a convenient tool to investigate all types of DNA rearrangement,
222 providing insight into genome stability and instability.

223

224 # Methods

225 ## SIGAR pipeline

226 SIGAR consists of three parts: 1) enrichment of MIC reads, 2) local alignment of MIC
227 reads to MAC genome, and 3) parsing the split-read alignment report. At each step the pipeline
228 provides adjustable parameters. All analyses in this paper were performed using the default
229 parameters.

230 Step 1. We map input reads to MAC genome by Bowtie2 (Langmead and Salzberg 2012)
231 end-to-end mapping to detect reads with only MDS. All reads with a mapping quality higher than
232 threshold (default 3) are removed from the downstream analysis by SAMtools (Li et al. 2009).

233 Step 2. Filtered reads, which mainly consist of MIC reads are aligned to MAC contigs by
234 BWA MEM local mapping (Li and Durbin 2009) with lowest mapping quality of 10 (default).
235 MAC regions with abnormal high coverage, calculated by pileup.sh in BBtools
236 (sourceforge.net/projects/bbmap/), were excluded from downstream analysis. The intermediate
237 output of this step is used for visualization in this paper by IGV (Robinson et al. 2011).
238 Step 3. Parsing of the alignment output was implemented by Python. The main idea is
239 parsing the CIGAR strings and "SA" tag in the alignment output. For example, CIGAR
240 "40S60M" means that the initial 40 bp are soft clipped from mapping and a rearrangement
241 junction is inferred at 40-41bp in the read. CIGAR "30M30I40M" means that a 30bp IES is
242 inferred at a nonscrambled junction (Figure 1B). "SA" tags represent supplementary alignment
243 of the read, indicating at least two mapping blocks present in a single read. These "SA" tagged
244 reads are used to infer IES or scrambled loci.
245 Once SIGAR collects the split positions in reads by parsing CIGAR and "SA" tags, it
246 infers pointers by pairwise comparison of alignments split in different directions (Figure 1B).
247 Alignments are grouped as 5' splits and 3' splits. For example, "40S60M" is a 5' split read,
248 while "30M30I40M" contains a 30bp 3' split alignment and a 40bp 5' split alignment. The
249 overlapped sequence between 5' split and 3' split is identified as a pointer. SIGAR outputs the
250 pointers and the number of reads supporting each pointer boundary.
251 All source codes and manual for SIGAR are available at https://github.com/yifeng-
252 evo/SIGAR.
253

254 **Genomes and datasets**

255 The *Oxytricha trifallax* (strain JRB310) MAC genome data in this study are from MDS-
256 IES-DB (http://knot.math.usf.edu/mds_ies_db/; Swart et al. 2013; Burns et al. 2016) and MIC
257 Illumina reads are from GenBank SRX365496, SRX385993, SRX385994, SRX385995 and
258 SRX385996 (Chen et al. 2014). Simulated MIC reads were produced by ART (Huang et al.
259 2012). To estimate uniquely mapped regions in SIGAR analysis, we simulated 30bp and 92bp
260 reads using MAC genome as reference, to mimic partially aligned regions in split reads,
261 representing the minimum and maximum SIGAR split reads alignment. We mapped these reads
262 to MAC genome by BWA MEM and filtered by mapq 10, the default setting of SIGAR. 63.69%
263 reference bases are covered for 92bp reads and 60.77% for 30bp reads, which indicates that
264 ~60% MAC genome is considered in the SIGAR analysis.

265 The hypotrich MAC genomes and whole cell DNA sequences are from Chen et al. 2015,
266 accession numbers *Laurentilla sp*. LASS02000000, *Sterkiella histriomuscorum* LAST02000000,
267 *Stylonychia lemnae* ADNZ03000000, *Urostyla sp*. LASQ02000000, and *Paraurostyla sp*.
268 LASR02000000. Only telomeric MAC contigs with "CCCCAAAACCCC" or
269 "GGGGTTTTGGGG" were used in the analysis. All SIGAR annotations are within the contig
270 body, which is at least 50bp from contig ends to avoid noisy mapping on telomeric regions.

271 *The Ichthyophthirius multifiliis* MAC genome is from Coyne et al. 2011, Genbank
272 accession number GCF_000220395.1. *Ichthyophthirius multifiliis* whole cell DNA sequence
273 reads were from MacColl et al. 2015.

274

Acknowledgements

We thank Theodore Clark and Donna Cassidy-Hanley for sharing *Ich* data and Sindhuja Devanapally for comments on the manuscript. We thank Rafik Neme, Jananan Pathmanathan, Talya Yerlici, Derek Clay, Sandrine Moreira, and all laboratory members for discussion. We are grateful for the National Center for Genome Analysis Support (NCGAS) computing resources (supported by NSF DBI-1062432, ABI-1458641, ABI-1759906 to Indiana University). This work was supported by NIH grant R35GM122555 and NSF grant DMS1764366 to L.F.L.

References

Arnaiz, Olivier, et al. "The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences." *PLoS Genetics* 8.10 (2012).

Aury, Jean-Marc, et al. "Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*." *Nature* 444.7116 (2006): 171-178.

Burns, Jonathan, et al. "<mds_ies_db>: a database of ciliate genome rearrangements." *Nucleic Acids Research* 44.D1 (2016): D703-D709.

Chang, Wei-Jen, Paul D. Bryson, Han Liang, Mann Kyoon Shin, and Laura F. Landweber. "The evolutionary origin of a complex scrambled gene." *Proceedings of the National Academy of Sciences* 102, no. 42 (2005): 15149-15154.

Chen, Xiao, et al. "Genome analyses of the new model protist *Euplotes vannus* focusing on genome rearrangement and resistance to environmental stressors." *Molecular Ecology Resources* 19.5 (2019): 1292-1308.

Chen, Xiao, et al. "The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development." *Cell* 158.5 (2014): 1187-1198.

Chen, Xiao, Seolkyoung Jung, Leslie Y. Beh, Sean R. Eddy, and Laura F. Landweber. "Combinatorial DNA rearrangement facilitates the origin of new genes in ciliates." *Genome Biology and Evolution* 7, no. 10 (2015): 2859-2870.

302 Coyne, Robert S., et al. "Comparative genomics of the pathogenic ciliate *Ichthyophthirius*
303 *multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic
304 lifestyle and prospects for disease control." *Genome Biology* 12.10 (2011): R100.

305 Forment, Josep V., Abderrahmane Kaidi, and Stephen P. Jackson. "Chromothripsis and cancer:
306 causes and consequences of chromosome shattering." *Nature Reviews Cancer* 12.10 (2012): 663-
307 670.

308 Guérin, Frédéric, et al. "Flow cytometry sorting of nuclei enables the first global characterization
309 of *Paramecium* germline DNA and transposable elements." *BMC Genomics18*.1 (2017): 327.

310 Hamilton, Eileen P., et al. "Structure of the germline genome of *Tetrahymena thermophila* and
311 relationship to the massively rearranged somatic genome." *Elife* 5 (2016): e19090.

312 Hogan, Daniel J., Elizabeth A. Hewitt, Kindra E. Orr, David M. Prescott, and Kirsten M. Müller.
313 "Evolution of IESs and scrambling in the actin I gene in hypotrichous ciliates." *Proceedings of*
314 *the National Academy of Sciences* 98, no. 26 (2001): 15101-15106.

315 Huang, Weichun, Leping Li, Jason R. Myers, and Gabor T. Marth. "ART: a next-generation
316 sequencing read simulator." *Bioinformatics* 28, no. 4 (2012): 593-594.

317 Katz, Laura A., and Alexandra M. Kovner. "Alternative processing of scrambled genes generates
318 protein diversity in the ciliate *Chilodonella uncinata*." *Journal of Experimental Zoology Part B:*
319 *Molecular and Developmental Evolution* 314.6 (2010): 480-488.

320 Kinsella, Cormac M., et al. "Programmed DNA elimination of germline development genes in
321 songbirds." *Nature Communications* 10.1 (2019): 1-10.

Klobutcher, Lawrence A., and Glenn Herrick. "Consensus inverted terminal repeat sequence of *Paramecium* lESs: resemblance to termini of Tc1-related and *Euplotes* Tec transposons." *Nucleic Acids Research* 23.11 (1995): 2006-2013.

Klobutcher, Lawrence A., and Glenn Herrick. "Developmental Genome Reorganization in Ciliated Protozoa: The Transposon Link." *Progress in Nucleic Acid Research and Molecular Biology* 56.5 (1997): 1-62.

Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." *Nature Methods* 9.4 (2012): 357.

Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." *Bioinformatics* 25.14 (2009): 1754-1760.

Li, Heng, et al. "The sequence alignment/map format and SAMtools." *Bioinformatics* 25.16 (2009): 2078-2079.

Lindblad, Kelsi A., et al. "Capture of complete ciliate chromosomes in single sequencing reads reveals widespread chromosome isoforms." *BMC Genomics* 20.1 (2019): 1-11.

MacColl, Elisabeth, et al. "Molecular genetic diversity and characterization of conjugation genes in the fish parasite *Ichthyophthirius multifiliis*." *Molecular Phylogenetics and Evolution* 86 (2015): 1-7.

Maurer-Alcalá, Xyrus X., Ying Yan, Olivia A. Pilling, Rob Knight, and Laura A. Katz. "Twisted tales: insights into genome diversity of ciliates using single-cell 'omics." *Genome Biology and Evolution* 10, no. 8 (2018): 1927-1938.

342 Prescott, David M. "The DNA of ciliated protozoa." *Microbiological Reviews* 58.2 (1994): 233-
343 267.

344 Robinson, James T., et al. "Integrative genomics viewer." *Nature Biotechnology* 29.1 (2011): 24-
345 26.

346 Sheng, Yalan, et al. "The completed macronuclear genome of a model ciliate *Tetrahymena*
347 *thermophila* and its application in genome scrambling and copy number analyses." *Science*
348 *China Life Sciences* (2020).

349 Slabodnick, Mark M., et al. "The macronuclear genome of *Stentor coeruleus* reveals tiny introns
350 in a giant cell." *Current Biology* 27.4 (2017): 569-575.

351 Smith, Jeramiah J., et al. "The sea lamprey germline genome provides insights into programmed
352 genome rearrangement and vertebrate evolution." *Nature Genetics* 50.2 (2018): 270-277.

353 Stankiewicz, Paweł, and James R. Lupski. "Structural variation in the human genome and its role
354 in disease." *Annual Review of Medicine* 61 (2010): 437-455.

355 Swart, Estienne C., et al. "The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic
356 genome with 16,000 tiny chromosomes." *PLoS Biology* 11.1 (2013).

357 Wang, Guangying, et al. "A strategy for complete telomere-to-telomere assembly of ciliate
358 macronuclear genome using ultra-high coverage Nanopore data." *bioRxiv* (2020)
359 https://doi.org/10.1101/2020.01.08.898502.

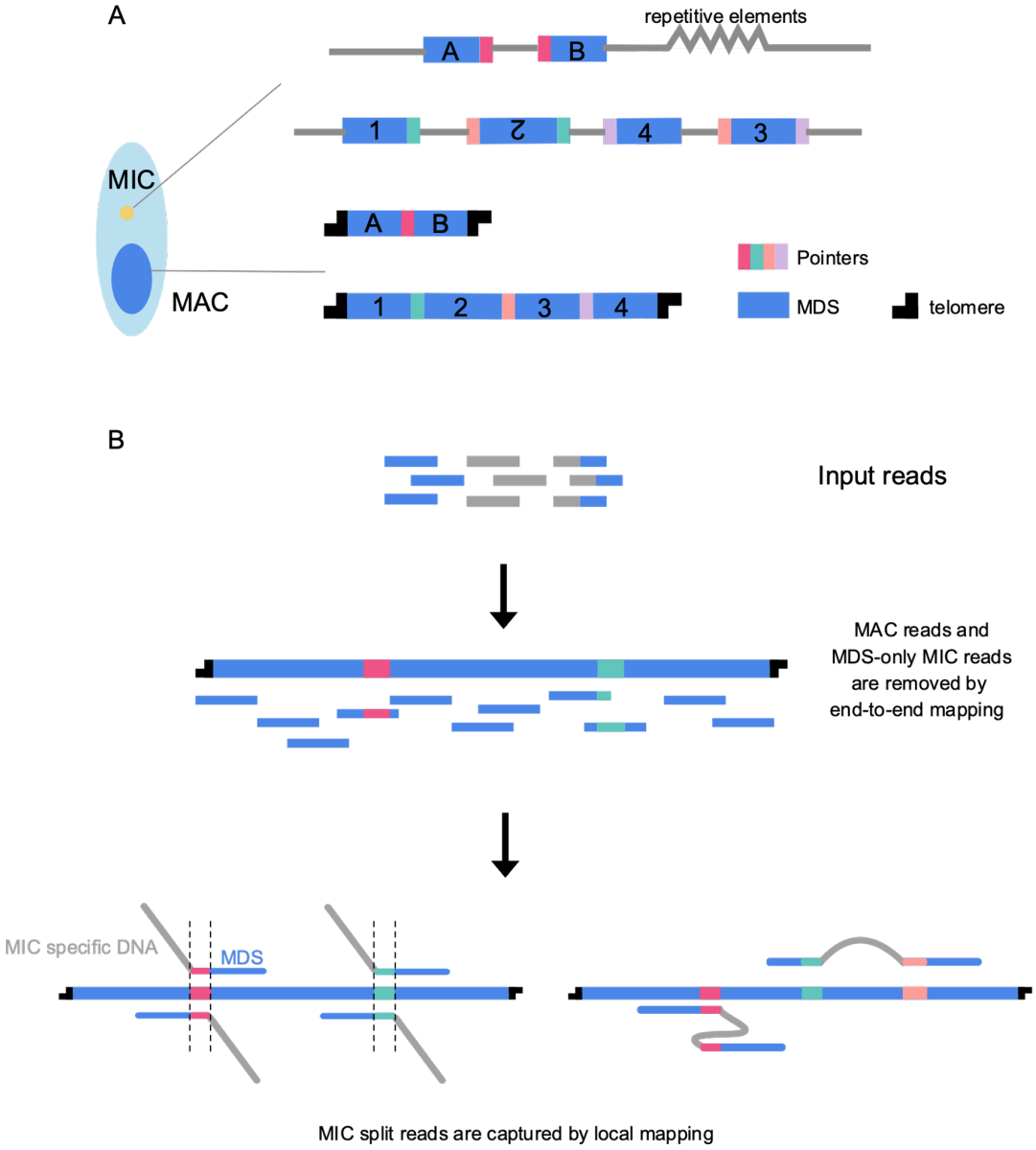
Wang, Ruanlin, Jie Xiong, Wei Wang, Wei Miao, and Aihua Liang. "High frequency of+ 1 programmed ribosomal frameshifting in *Euplotes octocarinatus*." *Scientific Reports* 6 (2016): 21139.

Xiong, Jie, et al. "Hidden genomic evolution in a morphospecies—The landscape of rapidly evolving genes in *Tetrahymena*." *PLoS Biology* 17.6 (2019): e3000294.

Zheng, Weibo, Jing Chen, Thomas G. Doak, Weibo Song, and Ying Yan. "ADFinder: accurate detection of programmed DNA elimination using NGS high-throughput sequencing data." *Bioinformatics* (2020) https://doi.org/10.1093/bioinformatics/btaa226.

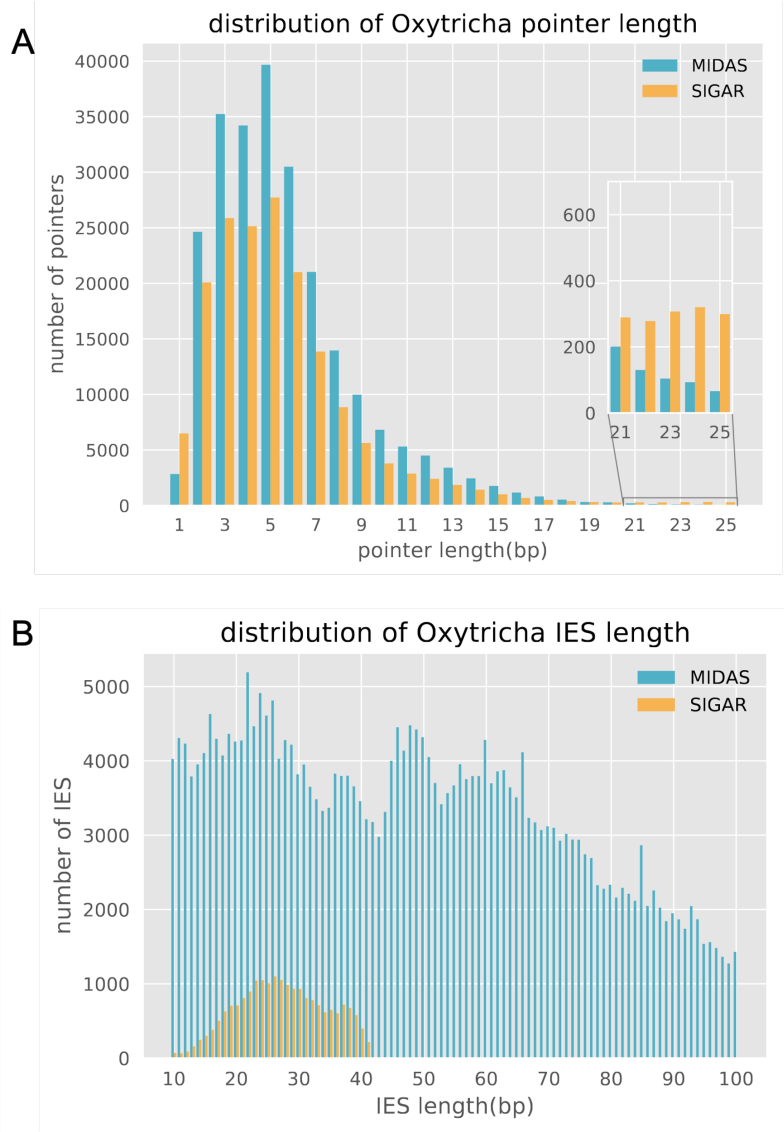
368 # Figures

369



370
371

372 **Figure 1**. Schematic of genome rearrangements in ciliates and SIGAR strategy. **A)** Ciliates have
373 separate germline (MIC) and somatic (MAC) genomes. The MAC chromosomes form from MIC
374 DNA during development by elimination of intervening DNA sequences (gray) and
375 reorganization of MDSs (blue). Some rearrangements join neighboring MDSs (e.g. A and B),
376 whereas scrambled rearrangements require translocation and/or inversion (e.g. 1-4).
377 Microhomologous pointer sequences, shown in different colors, are present at the end of MDS n
378 and the beginning of MDS n+1 on the MIC chromosome, with one copy retained in the MDS-
379 MDS junction on the MAC chromosome. **B)** SIGAR strategy. Reads that only contain MDS are
380 removed by end-to-end mapping to MAC contigs. The filtered reads are mapped locally to
381 identify MIC reads that split at MDS-MDS junctions. Such reads that map to both MDS n and
382 MDS n+1 permit inference of the pointer sequence from the overlapped region and the
383 eliminated sequence between them. Scrambled features of the germline map can sometimes be
384 inferred from reads containing at least two mapped blocks.
385

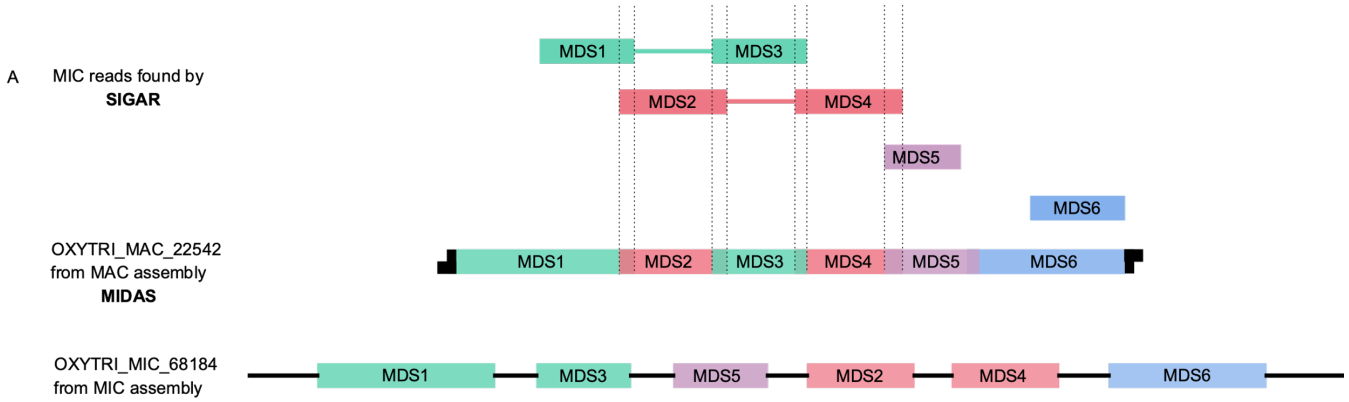

A
distribution of Oxytricha pointer length
MIDAS
SIGAR
number of pointers
pointer length(bp)
B
distribution of Oxytricha IES length
MIDAS
SIGAR
number of IES
IES length(bp)

386

387

388 **Figure 2**. Comparison of pointer and IES length distributions between methods. **A)** The pointer
389 length distribution and **B)** IES length distribution inferred for *Oxytricha* by MIDAS vs. SIGAR.
390 Note that SIGAR used ~110X Illumina reads from Chen et al (2014), but MIDAS used an
391 additional 15X PacBio reads for MIC genome assembly. Only ~60% of the MAC genome is
392 considered by SIGAR as uniquely mapped regions for analysis, and the inferred IES length is
393 restricted by read length (100bp in *Oxytricha* dataset). The pointer length distribution is only
394 shown for 1-25bp, and IESs between 10-100bp.

A



B

	SIGAR		MIDAS	
Pointers	Positions	Sequences	Positions	Sequences
MDS1-2	243-259	ATAAAGACTTGGCTACA	243-259	ATAAAGACTTGGCTACA
MDS2-3	312-319	GATGGCAC	312-324	GATGGCAC**G**TTAT*
MDS3-4	364-370	TAAAGCT	364-370	TAAAGCT
MDS4-5	386-399	TTACTGGAAGGAAC	386-399	TTACTGGAAGGAAC
MDS5-6	-	-	412-424	CTTAATATTCAAA

*: Mismatches are allowed in pointers predicted by MIDAS, here the bolded "G" is a mismatch

396 **Figure 3**. A representative scrambled region in *Oxytricha* inferred by SIGAR. **A)** SIGAR
397 identified scrambled MIC reads that could be validated by MIDAS. The green, red, purple and
398 blue reads are diagrams of split reads found by SIGAR mapped to OXYTRI_ MAC_22542 (see
399 Figure S3 for reads mapping view). Green and red reads show a scrambled structure and the
400 associated pointers are also inferred. **B)** The pointers on OXYTRI_ MAC_22542 found in
401 SIGAR are the same as previous annotations by MIDAS.