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 cell 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 MIDAS,
109 although fewer pointers were found in sum (Figure 2A). In addition, SIGAR inferred a small
110 number of "cryptic" pointers, defined as repeats longer than 20bp (Figure 2A) that differ from
111 MIDAS-annotated MDS-MDS junctions. It is possible that some represent short regions of
112 paralogy.

113 SIGAR only annotated half as many pointers as MIDAS (Table S1) because SIGAR was
114 only applied to uniquely mapping regions of the MAC genome, in order to minimize the rate of
115 false discovery. Such regions comprise 60.8-63.7% of the MAC genome, and contain 61.0-
116 62.7% of all pointers (see Methods). SIGAR successfully annotated 80.4-82.3% of these
117 pointers. We conclude that SIGAR recovers a large majority of pointers from the portion of the
118 genome to 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 20X
123 MIC coverage, SIGAR was still able to recover nearly half (48.74%) of MIDAS pointers, and
124 96.16% of SIGAR pointers were also identified by MIDAS (Table 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 results suggest that pointers can vary between alleles. We identified 11,462
130 pointers, validated by both MIDAS and SIGAR, which possess at least another pointer at the
131 same junction inferred by SIGAR with high confidence (at least 2 reads supporting each
132 boundary). Table S2 shows examples of these pointer alleles in *Oxytricha*. We highlight one
133 example where pointer alleles correlate with SNPs in the MIC reads (Figure S1). The allele-
134 specific information enriched in reads could only be detected by SIGAR (which examines
135 individual sequencing reads) but not assembly-based software like MIDAS, because genome
136 assemblies tend to collapse alleles into haploid sequences. This finding suggests that MAC
137 chromosomes in *Oxytricha* arise from both precursor MIC alleles, which use different pointer
138 sequences during MIC genome rearrangements.
139

140 SIGAR infers scrambled genome architecture

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

148 chromosome architectures and their associated pointers, even in the absence of a MIC genome
149 assembly. When applied to *Oxytricha* datasets, SIGAR detected 8741 *Oxytricha* pointers at
150 scrambled junctions. Figure 3 shows an example where SIGAR successfully inferred 3 out of 5
151 scrambled junctions in an *Oxytricha* MAC chromosome (Figure 3). SIGAR found 12 MIC reads
152 covering both MDS1 and MDS3, while another 12 reads covered both MDS2 and MDS4 (Figure
153 3A and Figure S2).
154

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

157 SIGAR was developed to provide insights into germline rearrangement features across
158 ciliates for which pure MIC DNA preparations are not readily available. We therefore applied
159 SIGAR to 5 hypotrich ciliates published in Chen et al. (2015), all of which have gene-sized
160 MAC nanochromosomes like *Oxytricha*. Importantly, these sequencing datasets were derived
161 from small numbers of whole cells, some from species that were not cultivated in the lab. We
162 were able to infer pointers in all of these species using SIGAR (Table 1). TA is the most favored
163 pointer among all 5 species. Though *Sterkiella* and *Urostyla* do not have TA as the most
164 abundant pointer, both include A, T and TA as the three most abundant. With increasing
165 evolutionary distance from *Oxytricha*, we observed more pointers containing TA as a substring.
166 The GC content of short IES (under 0.2) is significantly lower than adjacent MDS regions (~0.3),
167 consistent with surveys from other ciliate MIC genomes (Prescott 1994; Chen et al 2014). The
168 length distributions of pointers in 5 species are similar, except for *Sterkiella,* which exhibits an
169 abundance of 5-20bp pointers (Figure S3).
170

171 **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	-	-	-

172

173 SIGAR also revealed evidence for novel scrambled loci in the 5 hypotrich ciliates (Table
174 1). Figure S4 shows the mapping view of an inferred scrambled locus in *Paraurostyla*, with
175 pointers detected on each side. Short IESs were inferred for all species, except *Urostyla* (Table 1,
176 Figure S3). Though short IESs were annotated in the *Urostyla grandis* genes for DNA pol α
177 (Chang et al. 2005) and actin I (Hogan et al. 2001), we were unable to recover any short IES
178 from the current dataset with limited MIC DNA. It is possible that *Urostyla* may contain more
179 intergenic IESs, which do not reside next to MDSs, as in *Tetrahymena* (Hamilton et al. 2016).
180 Such regions cannot be detected by SIGAR, which identifies IESs adjacent to MDSs.

181 We also applied SIGAR to *Ichthyophthirius multifiliis (Ich),* an oligohymenophorean
182 ciliate related to *Tetrahymena* (Coyne et al. 2011). *Ich* lives a parasitic life in fish epithelia,
183 leading to "white spot" disease (Coyne et al. 2011). Furthermore, for ciliates that are hard to
184 cultivate, SIGAR offers an ideal tool to infer properties of MIC genome architecture and DNA
185 rearrangement, given the lack of availability of high quality MIC DNA preparations. Since the
186 *Ich* MAC genome is ~85% A+T, we only considered inferred pointers with at least 5 well-
187 mapped reads supporting each boundary. We found that the most abundant pointers are not only

188 AT rich, but also contain surprisingly long TA tandem repeats (Figure S5). The pointer length
189 distribution shows a peak at 10bp, with the 10bp pointer "TATATATATA" and
190 "ATATATATAT" among the most abundant pointers in *Ich*.
191

192 # Discussion

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

211 this commonly observed 2bp TA pointer (Klobutcher and Herrick 1997, Chen et al. 2014), long
212 TA repeats are present at *Ich* rearrangement junctions (Figure S5). Given that pointer sequences
213 may help recruit enzymes that mediate IES removal and are necessary for excision in
214 *Paramecium* (Aury et al. 2006), it is plausible that their extended length constitutes an adaptive
215 feature to improve recognition amidst an AT-rich genome.

216 In addition to ciliates, many organisms in nature exhibit programmed genome
217 rearrangement, such as lampreys (Smith et al. 2018) and songbirds (Kinsella et al. 2019).
218 Furthermore, aberrant structural variations in mammalian cells are frequently observed in
219 diseases like cancer (Stankiewicz and Lupski 2010; Forment et al. 2012). We expect SIGAR to
220 be directly applicable across a wide range of genomes that exhibit rearrangements in both
221 healthy and diseased states. SIGAR, which only requires one reference genome and short reads
222 from a rearranged genome, could be a convenient tool to investigate all types of DNA
223 rearrangement, providing insight into genome stability and instability.
224

225 Methods

226 SIGAR pipeline

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

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

234 Step 2. Filtered reads, which mainly consist of MIC reads are aligned to MAC contigs by
235 BWA MEM local mapping (Li and Durbin 2009) with lowest mapping quality of 10 (default).
236 MAC regions with abnormal high coverage, calculated by pileup.sh in BBtools
237 (sourceforge.net/projects/bbmap/), were excluded from downstream analysis. The intermediate
238 output of this step is used for visualization in this paper by IGV (Robinson et al. 2011).
239 Step 3. Parsing of the alignment output was implemented by Python. The main idea is
240 parsing the CIGAR strings and "SA" tag in the alignment output. For example, the alignment of
241 one read has CIGAR "40S60M", meaning that the initial 40 bp are soft clipped from mapping. A
242 rearrangement junction is therefore inferred at 40-41bp of the read. The "SA" tag represents
243 supplementary alignment of the read, indicating that at least two mapping blocks are found in a
244 single read. These "SA" tagged reads are used for further inference of IES or scrambled loci.
245 All source codes and manual for SIGAR are available at https://github.com/yifeng-
246 evo/SIGAR.
247

248 **Genomes and datasets**

249 The *Oxytricha trifallax* (strain JRB310) MAC genome data in this study are from MDS-
250 IES-DB (http://knot.math.usf.edu/mds_ies_db/; Swart et al. 2013; Burns et al. 2016) and MIC
251 Illumina reads are from GenBank SRX365496, SRX385993, SRX385994, SRX385995 and
252 SRX385996 (Chen et al. 2014). Simulated MIC reads were produced by ART (Huang et al.
253 2012). To estimate the uniquely mapped region used in SIGAR analysis, we simulated 30bp
254 reads and 92bp reads using MAC genome as the reference to mimic the partially aligned region
255 in split reads. The 30bp and 92bp are the minimum and maximum of the SIGAR split reads
256 alignment length respectively. We mapped these reads to MAC genome by BWA MEM and

filtered by mapq 10, the default setting of SIGAR. The ratio of reference bases covered are 63.69% for 92bp reads and 60.77% for 30bp reads, which indicates that only ~60% MAC genome is considered in the SIGAR analysis.

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

The Ichthyophthirius multifiliis MAC genome is from Coyne et al. 2011, Genbank accession number GCF_000220395.1. *Ichthyophthirius multifiliis* whole cell DNA sequences were produced as described in MacColl et al. 2015.

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 current laboratory members for discussion. We are grateful for the National Center for Genome Analysis Support (NCGAS) computing resources (supported by National Science Foundation DBI-1062432, ABI-1458641, ABI-1759906 to Indiana University). This work was supported by National Institutes of Health [grant R35GM122555] and National Science Foundation [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.

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

301 Forment, Josep V., Abderrahmane Kaidi, and Stephen P. Jackson. "Chromothripsis and cancer:
302 causes and consequences of chromosome shattering." Nature Reviews Cancer 12.10 (2012): 663-
303 670.

304 Fox, Edward J., Kate S. Reid-Bayliss, Mary J. Emond, and Lawrence A. Loeb. "Accuracy of
305 next generation sequencing platforms." Next generation, sequencing & applications 1 (2014).

306 Guérin, Frédéric, et al. "Flow cytometry sorting of nuclei enables the first global characterization
307 of Paramecium germline DNA and transposable elements." BMC genomics18.1 (2017): 327.

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

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

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

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

318 Kinsella, Cormac M., et al. "Programmed DNA elimination of germline development genes in
319 songbirds." Nature communications 10.1 (2019): 1-10.

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

323 Klobutcher, Lawrence A., and Glenn Herrick. "Developmental Genome Reorganization in
324 Ciliated Protozoa: The Transposon Link." Progress in nucleic acid research and molecular
325 biology 56.5 (1997): 1-62.

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

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

330 Li, Heng, et al. "The sequence alignment/map format and SAMtools." Bioinformatics 25.16
331 (2009): 2078-2079.

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

334 MacColl, Elisabeth, et al. "Molecular genetic diversity and characterization of conjugation genes
335 in the fish parasite Ichthyophthirius multifiliis." Molecular phylogenetics and evolution 86
336 (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.

Prescott, David M. "The DNA of ciliated protozoa." Microbiol. Mol. Biol. Rev. 58.2 (1994): 233-267.

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

Sheng, Yalan, et al. "The completed macronuclear genome of a model ciliate Tetrahymena thermophila and its application in genome scrambling and copy number analyses." SCIENCE CHINA Life Sciences (2020).

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

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

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

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

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

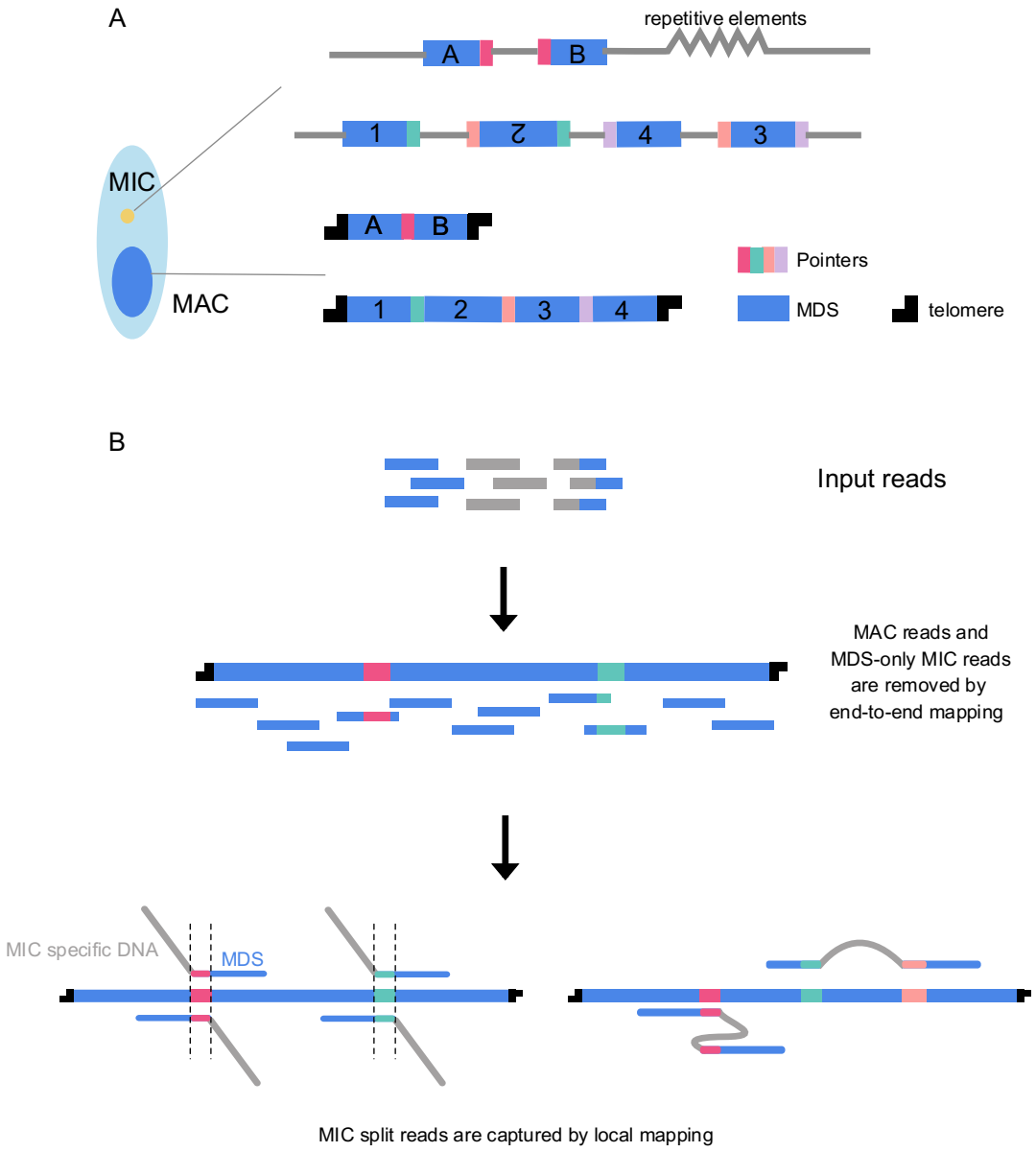
Wang, Ruanlin, Jie Xiong, Wei Wang, Wei Miao, and Aihua Liang. "High frequency of+ 1 programmed ribosomal frameshifting in Euplotes octocarinatus." Scientific Reports 6 (2016): 21139.

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.

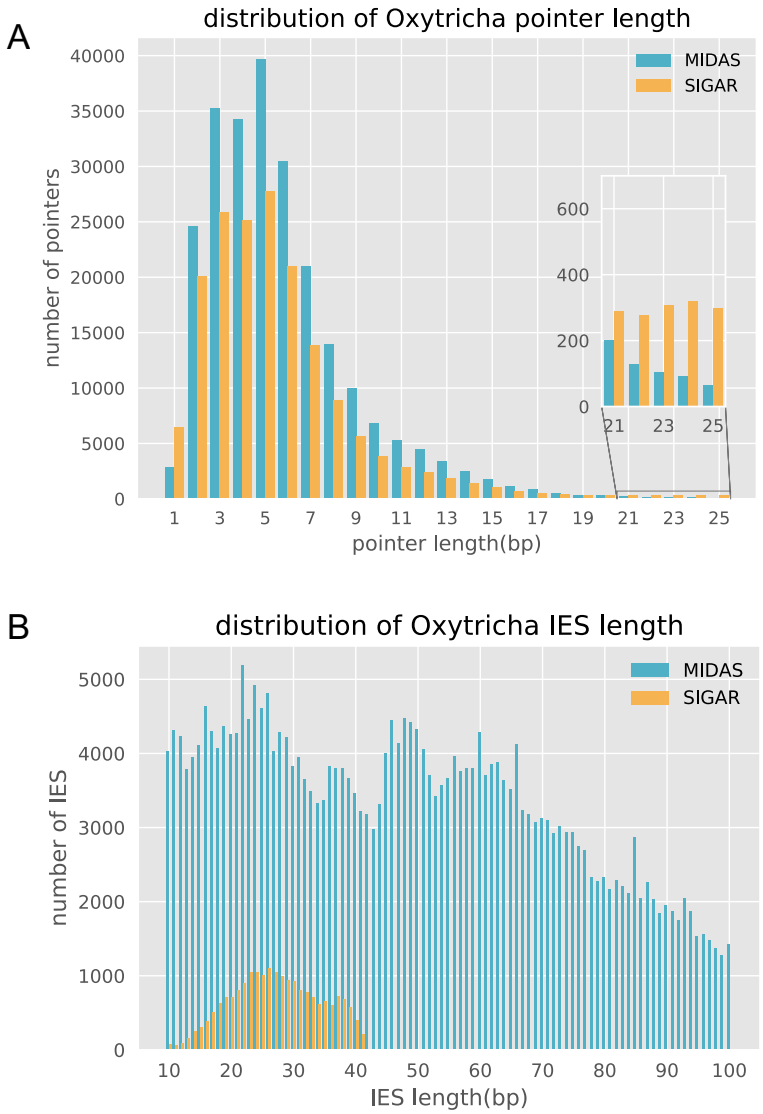
366 # Figures

367



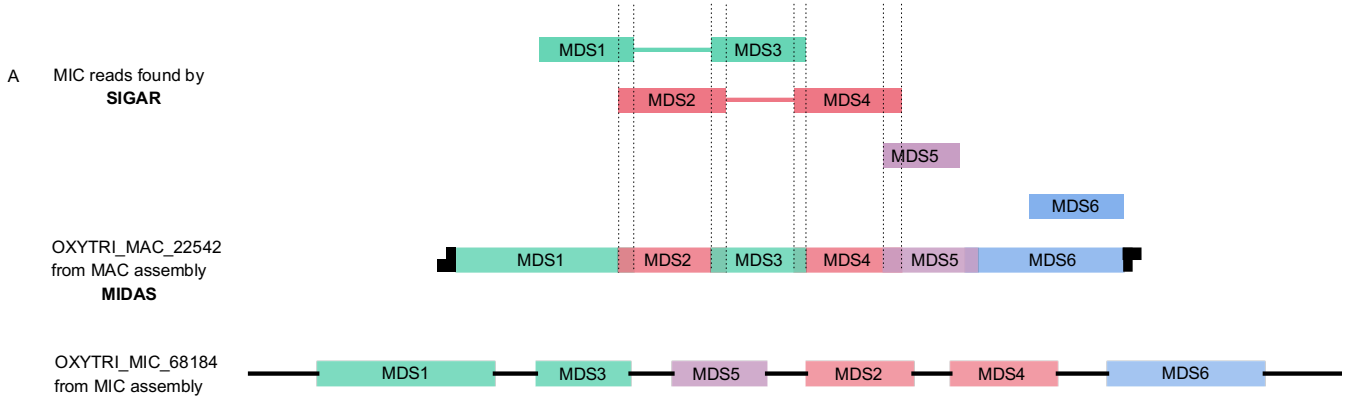
368
369

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


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

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 S2 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.
402

403 *Significance statement*

404 The current study of genome rearrangement in ciliates is restricted to a few model species due to
405 the lack of high-quality germline genome assemblies for most species. We present a new
406 computational pipeline to infer properties of DNA reorganization and germline genome
407 architecture without the requirement of a germline genome assembly, using only short reads
408 mapped to a reference somatic genome. The pipeline has the advantage of being applicable to a
409 range of evolutionarily diverged organisms with limited culturing capacity or DNA sequence
410 availability.

411

412 *Data Availability Statement*

413 All source codes and manual for SIGAR are available at https://github.com/yifeng-evo/SIGAR.

414