Multiple Layers of Nested Genes in the Complex Genome of *O. trifallax*.

Jasper Braun, Derek Covert, Nataša Jonoska and Masahiko Saito
University of South Florida: Department of Mathematics and Statistics

National Institutes of Health

Introduction

Massive genome rearrangement processes are known to occur during the sexual reproduction of the ciliate *Oxytricha trifallax*. During these processes, a transcriptionally active somatic nucleus (macronucleus) is developed from a copy of its germline nucleus (micronucleus). In the micronucleus, macronuclear genes are found in segments that are potentially scrambled and/or inverted.
In [2], it was shown that the internally eliminated sequences (IESs) between consecutive macronuclear destined segments (MDSs) of one gene can contain segments from other genes.

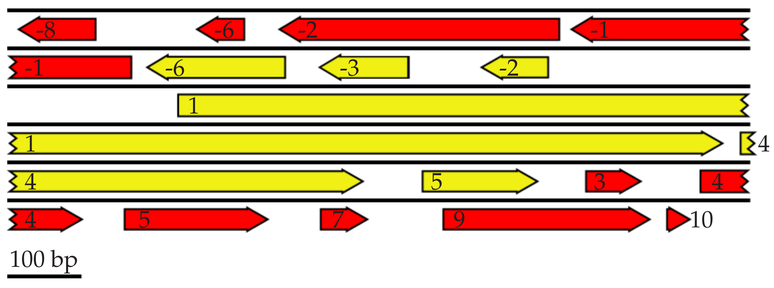

Figure 1: Partial germline map of the micronuclear contig OXYTRI_MIC_93112. The entire macronuclear contig OXYTRI_MAC_3908 (yellow) is found on an IES of the contig OXYTRI_MAC_20116 (red)[3]. The numbers indicate the corresponding ordering of the MDSs in the macronuclear contig. The negative sign indicates inverted segments.

Insertion Depth Index

We propose an insertion depth index (IDI) as a measure for the depth of the nested MDS appearances within a given IES. For an IES A, let $m(A)$ be the set of MDSs of other chromosomes that appear in A and $s(A)$ the set of IESs between MDSs in $m(A)$. Then $\mathrm{IDI}(A) = 0$ if $m(A)$ is empty, and $\mathrm{IDI}(A) = 1$ if $s(A)$ is empty. Otherwise:

$$\mathrm{IDI}(A) = 1 + max\ \mathrm{IDI}(B) \qquad (B \in s(A)).$$

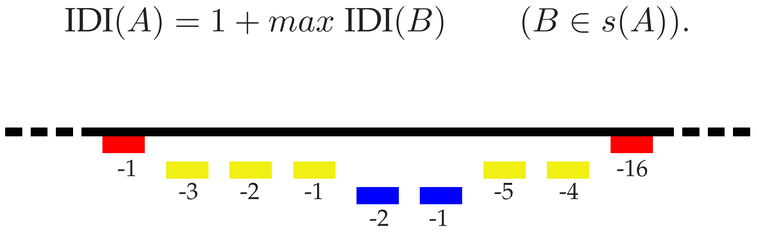

Figure 2: An example of an IES with IDI 2. The two contigs OXYTRI_MAC_9583 (yellow) and OXYTRI_MAC_6683 (blue) are nested inside OXYTRI_MAC_6331 (red) on the micronuclear contig OXYTRI_MIC_67077[3]. Not drawn to scale.

Data

From the germline and somatic genomes in [2] and [1], only macronuclear contigs with telomeres on both ends were considered. Further processing (as described in [4]) involved:

1. excluding macronuclear contigs that are subsequences of others,
2. merging consecutive MDSs if adjacent in a micronuclear contig, and
3. excluding macronuclear contigs if nonconsecutive MDSs overlap in a micronuclear contig.

Nested Genes

Multiple layers of nested MDSs from different macronuclear contigs in combination with specific scrambling patterns allow for the possibility of unscrambling major portions of the genes in a single step. Figure 3 depicts how such a rearrangement could occur on the micronuclear contig OXYTRI_MIC_93112[3] whose germline map is shown in figure 1.

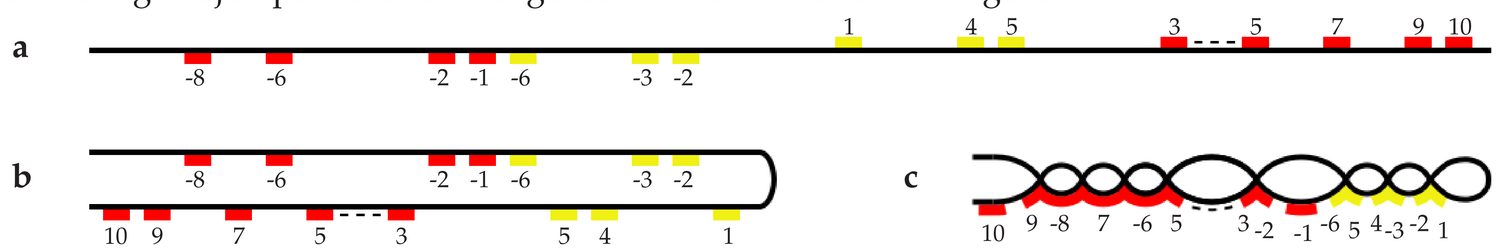

Figure 3: Alignment of consecutive MDSs on the micronuclear contig OXYTRI_MIC_93112[3]. Two MDSs connected by a dotted line indicate the presence of all consecutive MDSs between the two. **(a)** The order of the MDSs on the precursor sequence. **(b)** Consecutive MDSs begin to align. **(c)** Schematic picture of alignment. Not drawn to scale.

Results

The IESs of all macronuclear contigs in the processed data from [4] have been analyzed using custom Python scripts.
The numbers of MDSs interleaving into each IES was computed and the numbers of contigs with an IDI of at least 1, 2, 3 or 4 obtained, which is summarized in Table 1. The table also shows how many contigs are nested into IESs of others and how many are entirely contained in single IESs.
94.0% of nested genes are found entirely on a single IES. 63.6% of macronuclear contigs with a nonzero IDI are scrambled, and 22.1% of contigs which interleave into IESs of other contigs are scrambled. Additionally, the average number of MDSs of contigs with an IDI ≥ 1 is 14 whereas nested contigs have on average 10 MDSs.

One of two contigs with an IDI of 4 can be seen in figure 4. Figure 5 shows the alignment of the MDSs of two macronuclear contigs, each of which has MDSs in an IES of the other.

Table 1: Summary of the processed data

	total	scrambled
after processing	15,811[1]	2,021[1]
IDI ≥ 1	1,301	827
IDI ≥ 2	137	105
IDI ≥ 3	13	12
IDI ≥ 4	2	2
nested in other contigs	2,281	505
nested in single IES	2,144	432

[1]: Taken from [4]

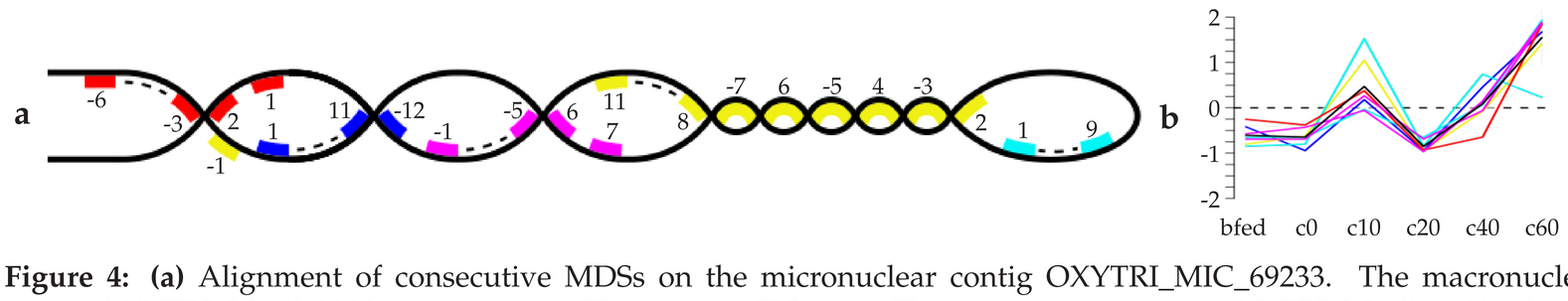

Figure 4: **(a)** Alignment of consecutive MDSs on the micronuclear contig OXYTRI_MIC_69233. The macronuclear contig OXYTRI_MAC_20394 (red) has an IES with an IDI of 4. The nested contigs are OXYTRI_MAC_14894 (blue), OXYTRI_MAC_9031 (purple), OXYTRI_MAC_18903 (yellow) and OXYTRI_MAC_20347 (light blue)[3]. Not drawn to scale. **(b)** Normalized gene expression patterns of the macronuclear contigs shown in (a). Time point bfed is the vegetative cycle, c0 is at start of conjugation and c10, c20, c40 and c60 mark the time points 10, 20, 40 and 60 hours post conjugation. The red, the yellow and one of the purple lines are hypothetical proteins. The other yellow line is phosphoglycolate/pyridoxal phosphate phosphatase, the blue line is DENN (AEX-3), and the two light blue lines are PAC2 and SET[3].

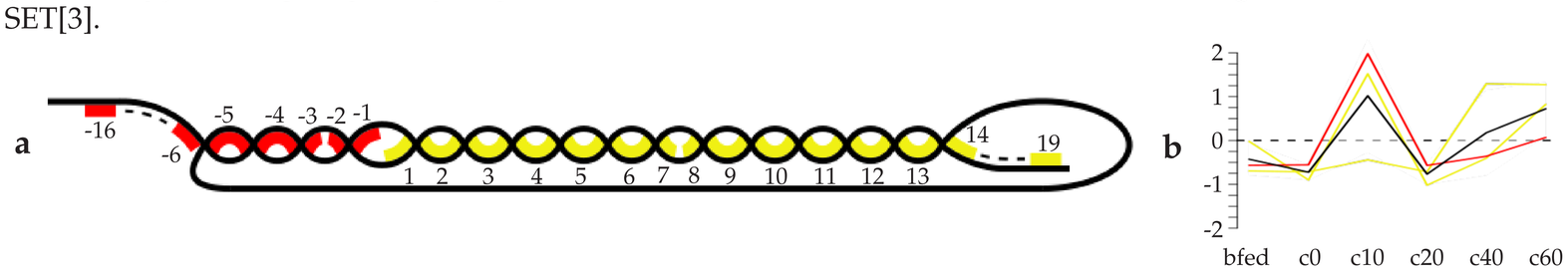

Figure 5: (a) Alignment of consecutive MDSs on the micronuclear contig OXYTRI_MIC_69076. The macronuclear contigs OXYTRI_MAC_22087 (red) and OXYTRI_MAC_11124 (yellow) are mutually nested[3]. Not drawn to scale. **(b)** Normalized gene expression patterns of the macronuclear contigs shown in (a). Time points are the same as above. Red is a hypothetical protein and the genes corresponding to yellow are a cupin superfamily protein and a dolichol phosphate-mannose biosynthesis regulatory protein (DPM2)[3].

References

[1] Swart et al. The *Oxytricha trifallax* macronuclear genome: A complex eukaryotic genome with 16,000 tiny chromosomes. *PLOS Biology*, 11, January 2013.

[2] Chen et al. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell*, 158(5):1187–1198, August 2014.

[3] Burns et al. <mds_ies_db>: a database of ciliate genome rearrangements. *Nucleic Acids Research*, 44(D1):D703–D709, November 2015.

[4] Burns et al. Recurring patterns among scrambled genes in the encrypted genome of the ciliate *Oxytricha trifallax*. *Journal of Theoretical Biology*. In revision.

Special thanks to the members of Math-Bio Lab at the University of South Florida.

This work has been supported by the NIH grant No. R01 GM109459.