Journal of Theoretical Biology 410 (2016) 171–180

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Recurring patterns among scrambled genes in the encrypted genome of the ciliate *Oxytricha trifallax*

Jonathan Burns [a,b,1], Denys Kukushkin [a,1], Xiao Chen [b], Laura F. Landweber [b,c], Masahico Saito [a], Nataša Jonoska [a,*]

[a] Department of Mathematics and Statistics, University of South Florida, Tampa, FL 33620, United States
[b] Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, United States
[c] Departments of Biochemistry & Molecular Biophysics and Biological Sciences, Columbia University, New York, NY 10032, United States

HIGHLIGHTS

- Five patterns describe over 80% of Oxytricha's completely annotated scrambled genes.
- A rearrangement pattern can have up to four equivalent representations.
- Inversions, translocations, and alternating odd–even patterns are most frequent.
- Nested instances of the five patterns account for over 96% of all scrambled genes.

ARTICLE INFO

Article history:
Received 28 April 2016
Received in revised form
21 June 2016
Accepted 30 August 2016
Available online 2 September 2016

Keywords:
Scrambled genes
Genome rearrangement
Homologous recombination
DNA inversion
Translocation

ABSTRACT

Some genera of ciliates, such as *Oxytricha* and *Stylonychia*, undergo massive genome reorganization during development and provide model organisms to study DNA rearrangement. A common feature of these ciliates is the presence of two types of nuclei: a germline micronucleus and a transcriptionally-active somatic macronucleus containing over 16,000 gene sized "nano-chromosomes". During conjugation the old parental macronucleus disintegrates and a new macronucleus forms from a copy of the zygotic micronucleus. During this process, macronuclear chromosomes assemble through DNA processing events that delete 90–98% of the DNA content of the micronucleus. This includes the deletion of noncoding DNA segments that interrupt precursor DNA regions in the micronucleus, as well as transposons and other germline-limited DNA. Each macronuclear locus may be present in the micronucleus as several nonconsecutive, permuted, and/or inverted DNA segments. Here we investigate the genome-wide range of scrambled gene architectures that describe all precursor-product relationships in *Oxytricha trifallax*, the first completely sequenced scrambled genome. We find that five general, recurrent patterns in the sets of scrambled micronuclear precursor pieces can describe over 80% of *Oxytricha*'s scrambled genes. These include instances of translocations and inversions, and other specific patterns characterized by alternating stretches of consecutive odd and even DNA segments. Moreover, we find that iterating patterns of alternating odd–even segments up to four times can describe over 96% of the scrambled precursor loci. Recurrence of these highly structured genetic architectures within scrambled genes presumably reflects recurrent evolutionary events that gave rise to over 3000 of scrambled loci in the germline genome.

© 2016 Elsevier Ltd. All rights reserved.

The ciliates *Oxytricha* and *Stylonychia* are microbial eukaryotes that undergo massive genome rearrangement during their sexual life-cycle. We provide a short description of the biological phenomenon and refer the reader to recent reviews (Yerlici and Landweber, 2014; Bracht et al., 2013) and references therein for more details. Ciliates have two types of nuclei, a germline micronucleus (MIC) and a transcriptionally-active somatic macronucleus (MAC). The macronuclear DNA of *O. trifallax* consists of multiple copies of over 16,000 different gene-sized nano-chromosomes (Swart et al., 2013). During sexual reproduction, the parental MAC disintegrates and a new MAC forms from a copy of the new zygotic micronucleus. In the midst of this process, macronuclear chromosomes assemble through DNA processing events involving global deletion of 90–98% of the germline DNA, eliminating nearly

* Corresponding author.
E-mail address: jonoska@mail.usf.edu (N. Jonoska).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jtbi.2016.08.038
0022-5193/© 2016 Elsevier Ltd. All rights reserved.

all so-called "junk" DNA, including short DNA segments (internal eliminated sequences, IESs) that interrupt precursor of macronuclear gene loci in the micronucleus. Because IESs often interrupt coding regions, each macronuclear gene may appear as several nonconsecutive segments (macronuclear destined sequences, MDSs) in the micronucleus. Moreover, the precursor order of these MDS segments for over 3500 genes can be permuted in the micronucleus (Chen et al., 2014). During macronuclear development, the IESs that interrupt MDSs are deleted, and the MDSs rearrange to form the gene-sized chromosomes. Assembly of the new macronuclear chromosomes may require any combination of the following three events: descrambling of segment order, DNA inversion, and IES removal. Telomere addition also occurs at both nanochromosome ends. Several theoretical models have studied these DNA recombination processes (Ehrenfeucht et al., 2004, 2002; Kari and Landweber, 1999; Prescott et al., 2003; Angeleska et al., 2007). The more recent models include an additional template RNA (or DNA) that guides the recombination process (Prescott et al., 2003; Angeleska et al., 2007), as experimentally demonstrated in Nowacki et al. (2008). A theoretical model based on spatial graphs was also introduced in Angeleska et al. (2007). This model describes a micronuclear locus as a spatial graph, called an assembly graph, where vertices represent the homologous recombination sites, while a macronuclear chromosome is represented as a special path in the graph.

Recently the whole macronuclear (MAC) (Swart et al., 2013) and micronuclear (MIC) genomes for *O. trifallax* have been reported (Chen et al., 2014), allowing for a genome-wide annotation of the rearrangement maps between the MIC and the MAC sequences. Although the somatic genome sequences for other species of ciliates have been reported as well, e.g., Aeschlimann et al. (2014), Eisen et al. (2006), Ricard et al. (2008), Vinogradov et al. (2012). Chen et al. (2014) provides the first published germline genome for any ciliate species. This allows us to draw a thorough comparison of the genetic sequences between the two types of nuclei, and to study the scrambled genetic patterns on a genome-wide scale.

In this paper we analyze the patterns of scrambled genes of *Oxytricha trifallax*. Specifically, we identify the distributions of scrambled and nonscrambled genes, and for scrambled genes, we identify the precursor scrambled patterns with two representations: (a) DNA sequence rearrangement maps and (b) assembly graph representations. Five basic types of scrambled patterns are present in the sequences: (1) translocation, (2) inversion, (3) translocation together with inversion, and (4), (5) two types of global rearrangement maps of successive odd–even arrangements. These descriptions account for 81.7% of all scrambled genes. Moreover, combinations of two or three of these steps or operations can account for over 90% of all scrambled genes. We further analyze scrambled patterns in the remaining complex cases and find that most of these loci (96.4%) can derive from nested cases (up to 5 iterations) of the previous operations, reflecting multiple layers of DNA rearrangements during evolution.

1. Materials and methods

1.1. The analyzed data set $\mathcal{D}$

The MDS annotations analyzed in this paper are presented in Chen et al. (2014), and can be downloaded directly from http://trifallax.princeton.edu/cms/raw-data/genome/mic/Oxytricha_trifallax_micronuclear_genome_MDS_IES_maps.gff, where the whole genome of *Oxytricha trifallax* is reported. This file contains MDS annotations for the 16,220 MAC contigs whose precursor MDSs map completely (>90% covered) to single MIC loci, which constitutes over 88% all the contigs in the complete MAC assembly. The MDSs of the scrambled MAC genes are disproportionately mapped to multiple MIC loci (Chen et al., 2014), but at least 2021 scrambled MAC loci are contained within single MIC loci. The data annotation includes the locations of MDSs in the MIC genome, labeled with their relative order of appearance, and whether the sequence of each MAC contig contains either or both telomeres (indicating the extent of the completion of the MAC contig assembly). The presence of two telomere sequences in a MAC contig indicates a complete somatic chromosome sequence (most of which are short and contain a single gene). MAC contigs can exhibit alternative fragmentation, with telomeres present at different sites, leading to multiple isoforms of a gene (Swart et al., 2013).

We filtered the data so that consecutive MDSs for a single MAC contig that overlap or have no nucleotide gap (adjacent) in the MIC are merged (correcting previous annotation artifacts). Further, a MAC contig is excluded from the dataset for analysis if either (a) it contains non-consecutive MDSs that overlap in a MIC contig, that is, two segments that are distant in the MAC contig overlap in a MIC contig, or (b) the MAC contig is an alternative fragmentation of a longer MAC contig. See Supplementary Materials for a detailed description of the data processing.

Let $\mathcal{D}$ represent the processed dataset. Specifically, $\mathcal{D}$ is the set of MDS sequences describing all MAC contigs in the annotation provided in Chen et al. (2014) that remain after the above filtering steps, and represents 97.5% of the MAC contigs with >90% coverage in the MIC, together with their corresponding MIC loci. Table 1 shows different cases of MAC contigs in $\mathcal{D}$, which contains 15,811 MAC contigs in total. There are 2958 MAC contigs whose MDSs overlap with MDSs of other MAC loci, for example when the telomere addition site often overlaps between the end of one locus and the beginning of the next locus (Chen et al., 2014). If a MAC contig C overlaps with another contig C', then both of them are counted as overlapping contigs. The remaining contigs belong to the *non-overlapping* category.

A MAC contig C is called *interleaving* if two or more of its precursor MDSs are separated by a segment containing an MDS of another contig. There are 790 interleaved contigs and 12,063 non-interleaved contigs among the non-overlapping MAC contigs. There are 1252 MIC contigs in the previously published genome assembly that contain the precursor MDSs for only a single MAC contig and all are non-interleaving. The remaining 10,811 MAC contigs reside within MIC contigs that contain precursor sequences for multiple MAC loci.

1.2. Definition of a rearrangement pattern

For a macronuclear (MAC) contig and a micronuclear (MIC) contig that contain corresponding MDSs, a *micronuclear rearrangement map* is a description of both the position and the orientation of each MDS in the micronuclear contig, relative to the

Table 1
Summary of the processed *O. trifallax* MIC MDS annotation reported in Chen et al. (2014). Each MAC contig is categorized according to the proximity of its MDSs with those of another MAC contig within a MIC locus.

Oxytricha trifallax MAC contigs and their corresponding MIC loci (16,220) Filtered MAC contigs and their MIC loci (15,811), data $\mathcal{D}$			
Overlapping (2958)	Non-overlapping (12,853)		
	Interleaving (790)	Non-interleaving (12,063)	
		Sole MAC contig on a MIC contig (1252)	Multiple MAC contigs on a MIC contig (10,811)

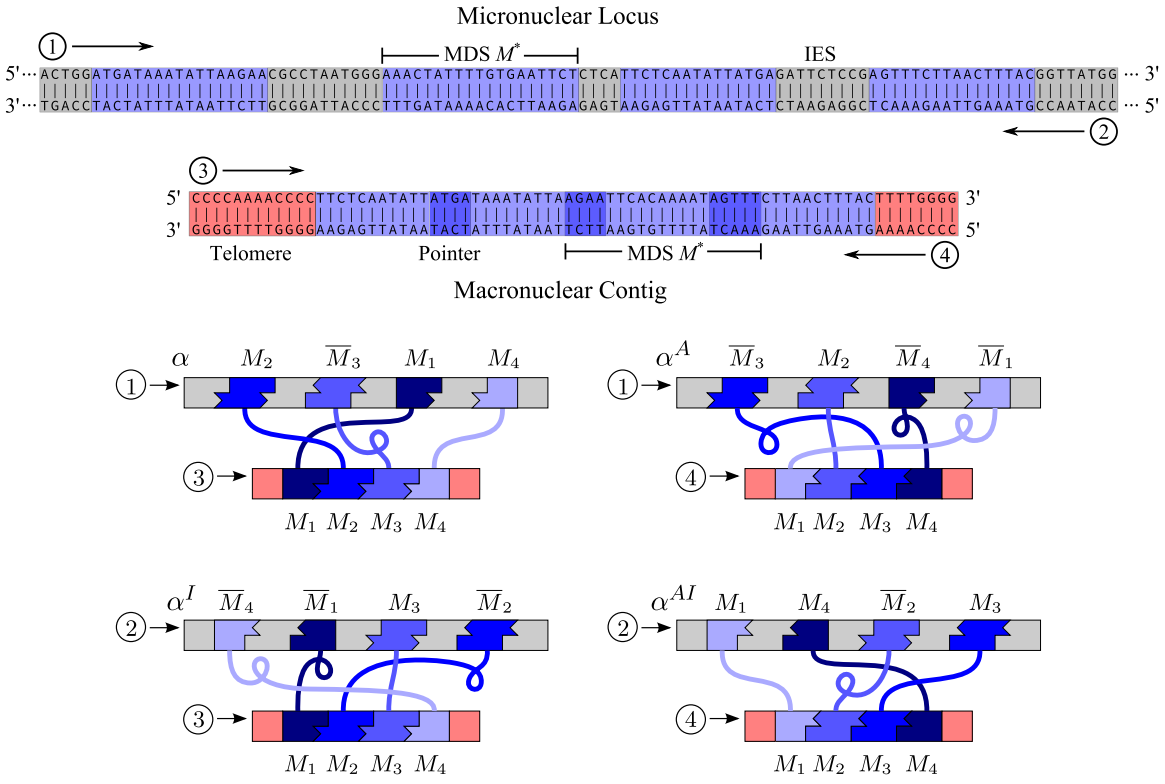


Fig. 1. A recombination pattern that could be annotated with the four possible rearrangement maps $\alpha = M_2\overline{M}_3M_1M_4$, $\alpha^I = \overline{M}_4\overline{M}_1M_3\overline{M}_2$, $\alpha^A = \overline{M}_3M_2\overline{M}_4\overline{M}_1$, and $\alpha^{AI} = M_1M_4\overline{M}_2M_3$, which correspond to the choice of micronuclear and macronuclear strands ① & ③, ② & ③, ① & ④, and ② & ④, respectively. Corresponding schematic diagrams of rearrangement maps are depicted below where the MAC MDSs are labeled from left to right. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

macronuclear contig. Up to four different sequence rearrangement maps, according to choices of the DNA strands in the MIC and MAC contigs, describe the same pattern of rearranged sequence. For instance in Fig. 1, the nucleotide sequence for the MDS $M^\star$ = AAACTATTTTGTGAATTCT, on strand ①, appears in both the macronuclear and micronuclear contigs, but its annotation position and orientation are dependent on the DNA strands chosen as reference. Reading from 5′ to 3′, $M^\star$ is either third MDS, labeled M_3, when the annotation uses strand ③, or it is the second MDS, labeled M_2, when the annotation uses strand ④. Once a MAC strand ③ or ④ is chosen, depending on the longest open frame (or the transcription unit in the case of RNA genes), the order of the MDS $M^\star$ in the MAC contig is fixed. The orientation of $M^\star$ is described relative to a MIC reference strand, ① or ②. If ① and ③ are chosen as the reference strands, then $M^\star$ appears inverted, denoted by $\overline{M}_3$, in the micronuclear contig relative to the macronuclear contig, whereas if ② and ④ are chosen, $M^\star$ appears non-inverted, and is denoted M_2.

In general, if a MAC contig consists of n MDSs, then the ith MDS with respect to one of the strands of the MAC contig might also be described as an inverted $(n+1-i)$ th segment with respect to the other strand of the MAC contig. Mathematically, we define a permutation ρ: $\{1, \ldots, n\} \to \{1, \ldots, n\}$ with $\rho(i) = n+1-i$ which we call the *reverse-index* operation.

Given a set $\Delta = \{M_1, M_2, \ldots, M_n\}$ a *micronuclear rearrangement map* is a sequence

$$\alpha = M_{i_1}^{\epsilon_1} M_{i_2}^{\epsilon_2} \cdots M_{i_{n-1}}^{\epsilon_{n-1}} M_{i_n}^{\epsilon_n}$$

where $(i_1, i_2, \ldots, i_n)$ is a permutation of indices $\{1, 2, \ldots, n\}$ and $\epsilon_i \in \{1, -1\}$ for $i = 1, \ldots, n$. Each symbol M_i indicates an MDS in a MAC contig. A rearrangement map is a reading of the MDS segments in a micronuclear contig annotated with respect to a fixed strand in the MAC contig and a fixed strand in the MIC contig. If $\epsilon_j = -1$ then we say that M_{i_j} is *inverted* and we also write $\overline{M}_{i_j}$. The *inverse* of rearrangement map α is $\alpha^I = M_{i_n}^{-\epsilon_n} M_{i_{n-1}}^{-\epsilon_{n-1}} \cdots M_{i_2}^{-\epsilon_2} M_{i_1}^{-\epsilon_1}$ which corresponds to reading of the MDSs with respect to the other strand of the MIC contig. The *reverse-index* rearrangement map of α is $\alpha^A = M_{\rho(i_1)}^{-\epsilon_1} M_{\rho(i_2)}^{-\epsilon_2} \cdots M_{\rho(i_n)}^{-\epsilon_n}$ which corresponds to ordering of the MDSs with respect to the complementary strand of the MAC contig. Finally, the fourth possible rearrangement map corresponding to the same MIC-MAC contig is the inverse reverse-index rearrangement map $\alpha^{AI} = M_{\rho(i_n)}^{\epsilon_n} M_{\rho(i_{n-1})}^{\epsilon_{n-1}} \cdots M_{\rho(i_1)}^{\epsilon_1}$. We say that two rearrangement maps α and β, which describe unrelated MAC loci, are *equivalent* and write $\alpha \equiv \beta$ if $\beta \in \{\alpha, \alpha^I, \alpha^A, \alpha^{AI}\}$. A *rearrangement pattern* (or just a *pattern*) is an equivalence class of rearrangement maps. A rearrangement map α is *scrambled* if it is not equivalent to $M_1M_2 \cdots M_n$, and we say that the pattern represented by α is also scrambled.

Since each pattern may be represented by up to four equivalent rearrangement maps, for each pattern we choose a *pattern representative* that (1) has minimal number of inverted MDSs in the rearrangement map, (2) has the lowest numerical lexicographic order for the MDS positions, and (3) delays the first inversion as late as possible, applied in order of precedence 1, 2 and 3. Thus, when two equivalent rearrangement maps do not have the same number of inverted MDSs, priority is given to the rearrangement map with fewer non-inverted segments. If there are two such rearrangement maps, the priority is given to the one with smaller lexicographic order of the MDSs. For example, in Fig. 1 both α and α^{AI} minimize the number of inversions, but α^{AI} has a lower lexicographic order so it is chosen to represent the scrambled pattern.

1.3. Reducing MDS Rearrangement maps

MDSs for nonscrambled patterns appear as consecutive in the MIC.

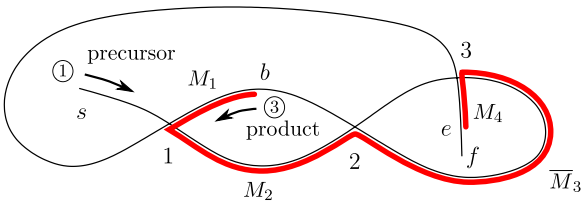


Fig. 2. The assembly graph 123213 that corresponds to the MIC and MAC contigs of Fig. 1. The thin black path, $s123213f$, that transverses the assembly graph represents the (precursor) locus on the MIC contig, and the bold red path, $b123e$, indicates the (product) MAC contig. The vertices of the graph represent the recombination junctions, i.e., the overlapping pointers. The segments between the vertices represent the MDSs of the MAC contig. Because the paths related to the MIC and MAC contigs follow the directions of the strands ① and ③ in Fig. 1, respectively, the MDSs encountered when following the path from s to f are expressed through the rearrangement map $\alpha = M_2\overline{M}_3M_1M_4$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Experimental data for a small number of loci suggests that the excision of IESs separating consecutive MDSs occurs early in macronuclear development (before DNA unscrambling) (Möllenbeck et al., 2008). In terms of rearrangement maps, excision of IESs flanked by consecutive MDSs M_j, M_{j+1}, …, M_{j+k}, can be rewritten as

$$\cdots M_i\big(M_jM_{j+1}\cdots M_{j+k}\big)M_l\cdots \mapsto \cdots M_i\big(M_j\big)M_l\cdots$$

and

$$\cdots M_i\big(\overline{M}_{j+k}\cdots\overline{M}_{j+1}\overline{M}_j\big)M_l\cdots \mapsto \cdots M_i\big(\overline{M}_j\big)M_l\cdots.$$

For convenience, we relabel such a reduced rearrangement map so that there are no breaks in consecutive MDS positions. In these cases, we say that the MDSs $M_j\cdots M_{j+k}$ are *merged into* M_j (or $\overline{M}_j$, respectively). For example, the consecutive MDSs in the rearrangement map $\alpha = M_2M_3M_1\overline{M}_6\overline{M}_5\overline{M}_4M_7$ are (M_2M_3) and $(\overline{M}_6\overline{M}_5\overline{M}_4)$, so α reduces to $(M_2)M_1(\overline{M}_4)M_7$ which we relabel to $\alpha^{red} = M_2M_1\overline{M}_3M_4$. With this convention, all nonscrambled rearrangement maps have a reduced rearrangement map of either $\alpha^{red} = M_1$ or $\overline{M}_1$ because all MDSs merge into M_1.

1.4. Double occurrence words and assembly graphs

MDSs in the MIC begin and end with short sequences, denoted "pointers," that provide some of the information for linking consecutive MDSs. The ith MDS in a MAC contig ends with a short (typically 2–20 bp) nucleotide repeat that also occurs at the beginning of the $(i+1)$ st MDS. These pointers annotate recombination junctions that form the MAC contig (dark blue in Fig. 1). After recombination, these pointer sequences appear as only one copy in the MAC contig. We model the rearrangement process with graphs and their corresponding words. For detailed explanation of the model, we refer the reader to Angeleska et al. (2007), Angeleska et al. (2009).

The pointers at the beginning and end of every MDS M_i are represented by a pair of integers $i-1$ and i, respectively, and for brevity we call them pointers $i-1$ and i. The first MDS has pointers b and 1, with b indicating the beginning of a MAC gene (5′ telomere sequence), and the last (nth) MDS of the gene has pointers $n-1$ and e, with e indicating the end of the gene (3′ telomere sequence). A word associated with a rearrangement map is the list of pointers, in order, as one traces an annotated MIC contig, reading 5′ to 3′. If a MAC contig consists of a single MDS in the MIC then it corresponds to the word be, i.e., beginning and end, because there are no pointers. For the MIC contig depicted in Fig. 1 and the choice of MDS order according to strands ① and ③, we obtain the (ordered) pointer list 1232b13e. Since the 5′ telomere addition site, b, of the first MDS M_1 and the 3′ telomere addition site, e, of the last MDS M_4 do not participate in the MDS reordering process, we can delete them in our representation. In the remaining word, 123213, every symbol, denoting an authentic pointer, appears twice. A *double occurrence word* (DOW for short) is a word in which every symbol appears exactly twice. For every pointer list obtained from a MIC contig, there is a corresponding DOW as described above. In the case of a single MDS MAC contig with a pointer list be, the corresponding DOW is the *empty* word $\varnothing$.

We associate a spatial graph with every DOW in the following way. Place vertices (points) on the plane labeled with pointer symbols i, a starting point s, and a final point f. Starting from the vertex (point) s (indicating the 5′ end of a MIC locus corresponding to strand ①), connect the vertices with line segments (edges) in the order of the symbols that appear in the given DOW. Since each symbol in the DOW appears twice, we visit every vertex (points labeled with symbols from the DOW) twice. When we visit a vertex for the second time, we go straight through it, intersecting the existing line, to the opposite side forming a cross at the vertex. The path terminates at vertex f. A graph constructed in this manner is called an *assembly graph* and corresponds to the MIC loci. The vertices labeled with DOW symbols represent the pointers. Note that s and f, representing the MIC loci ends, can differ from b and e which denote the 5′ and 3′ ends of the MAC loci – except when the terminal MDSs and hence the telomere addition sites, flank the MIC locus.

In Fig. 2, the graph corresponds to the DOW 123213, representing the MIC contig, and its pointer list, shown in Fig. 1. The MDSs are indicated with bold red, and their order is labeled according to choice of strands ① and ③, as described above for the DOW 123213. The MAC contig associated with this description represents a specific path through the graph (red portion) with all segments in the order $M_1M_2M_3M_4$. Since the graph represents a specific MIC contig, reading the path of the graph from s to f produces the rearrangement map $\alpha = M_2\overline{M}_3M_1M_4$. Reading the MDSs from f to s (reading strand ② from 5′ to 3′ in Fig. 1) produces the equivalent rearrangement map $\alpha^I = \overline{M}_4\overline{M}_1M_3\overline{M}_2$. Similarly, relabeling the MDSs in Fig. 2 as $M_1M_2M_3M_4$ from e to b (reading strand ④ from 5′ to 3′ in Fig. 1) produces the equivalent rearrangement map $\alpha^A = \overline{M}_3M_2\overline{M}_4\overline{M}_1$.

We note that there may be several patterns that correspond to a given assembly graph. In particular, the map $M_1M_3\overline{M}_2\overline{M}_4$ has the same pointer sequence 123213 as $\alpha = M_2\overline{M}_3M_1M_4$ but its pattern (represented by the path in red in Fig. 3) is distinct from the pattern of α. Both correspond to the same assembly graph depicted in Fig. 2.

The *reverse* of a word $w = a_1\cdots a_k$ is $w^R = a_k\cdots a_1$. Two double occurrence words are *equivalent* if, after renaming of the symbols, either they are identical or one is the reverse of the other. For example, w=123213 is equivalent to its reverse w^R=312321 and $w' = 123132$ is equal to w^R after renaming in w': 1 with 3, 2 with 1, and 3 with 2. Therefore w, w^R, and w' are equivalent. Note that this does not distinguish between the sense or antisense strand order, but is an independent pattern. Two assembly graphs are

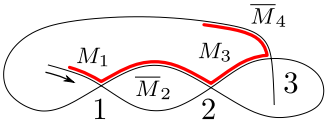


Fig. 3. Another rearrangement pattern represented by the same assembly graph as in Fig. 2. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

isomorphic if and only if they correspond to equivalent DOWs (Angeleska et al., 2009).

For convenience we choose a DOW representative where 1 is the letter assigned to the first vertex that is encountered in the graph starting from s, 2 is the next vertex, and so on. A DOW is in *ascending order* if the first occurrence of the numbers is in the order 1, 2, 3, …. For example, $w=123213$ is an assembly word in ascending order. Its reverse, with the same letters, is $w^R=312321$, which is not in ascending order, but by renaming the letters, we obtain $w' = 123132$ in ascending order, which represents the same graph ordering the vertices from f to s.

Consecutive MDSs in a rearrangement map correspond to a consecutive repetition of symbols. *Reduction* of a rearrangement map removes the consecutive symbols that are equal, corresponding to a DNA recombination event between the two matching pointers, or simplifying the DNA sequence by eliminating simple IESs separating consecutive, nonscrambled MDSs. A similar operation on DOWs was considered in Ehrenfeucht et al. (2004). In the example from Section 1.3, the reduction of $\alpha = M_2M_3M_1\overline{M}_6\overline{M}_5\overline{M}_4M_7$ to $\alpha^{red} = M_2M_1\overline{M}_3M_4$ corresponds to the removal of the repeated letters 2, 4, and 5 from the DOW $w=122316554436$, which becomes 131636 and is relabeled in ascending order to $w^{red} = 121323$. Biologically, during macronuclear development, evidence (Möllenbeck et al., 2008) suggests that simple IESs between nonscrambled MDSs may be eliminated first, before translocations or inversions that descramble the remaining MDS segments. This lends support to the proposed model.

1.5. Repeat and return words

We observe that the particular patterns

$$\cdots M_1M_3M_5\cdots M_2M_4M_6\cdots \quad \text{and} \quad \cdots M_1M_3M_5\cdots \overline{M}_6\overline{M}_4\overline{M}_2\cdots$$

occur highly among the scrambled MIC loci, where the portions indicated "…" can contain any other MDS sequence in the MAC contig (see Fig. 4). These MDS rearrangement maps correspond to 12345…12345, which is a *repeat word*, as in Fig. 4(A) and 12345…54321, which is a *return word*, as in Fig. 4(B) in the DOW. To analyze all general appearances of these patterns in the genome, we define the following formal notions.

- A word u, where $u = i(i+1)\cdots(i+j)$ for some positive integers i, j, is called a *repeat of* w, if w can be written as $w_0uw_1uw_2$. This means that u has at least two symbols and occurs twice in w.
- A word u, where $u = i(i+1)\cdots(i+j)$ for some positive integer i, j, is called a *return of* w, if w can be written as $w_0uw_1u^Rw_2$. Here both uu^R and $w_0w_1w_2$ are DOWs with no letters in common.
- A word u is a subword of w if w is written as $w = v_1uv_2$, where v_1 and v_2 are also words. A repeat or return u of w is called *maximal* if it is maximal with respect to the inclusion (of subwords). That is, u is not a proper subword of another repeat or return word.

In a broader sense, the maximal repeat and return words capture the longest run of consecutive odd–even interleaving MDS segments in a scrambled MIC contig. For example, in the word $w=12345677651234$, the words 12 and 123 are both repeat words of w and subwords of the maximal repeat word 1234 of w. Similarly 67 is a return word of w and a subword or the maximal return word 567 of w. A special case of repeat and return words are studied in Arredondo (2013), which analyzed the nested appearances of these words in a given DOW. Here we perform similar analysis of nested repeat and return words for the annotated data $\mathcal{D}$ to estimate all appearances of odd–even interleaved MDS patterns in *Oxytricha trifallax*. An example of a contig whose graph contains two step nested repeat and return words is depicted in Fig. 5.

Repeat–return iterations: In order to analyze the nested appearances of repeat and return words we iteratively remove their appearances in the DOWs. For each scrambled rearrangement map α in $\mathcal{D}$, we perform the following:

1. Merge consecutive MDSs in α to reduce it as described in Section 1.2. Construct the corresponding DOW w in ascending order. This is iteration 0.
2. Remove all maximal repeat and return words from w to obtain a new word w'.
3. If $w' \neq w$, then increase the count of the number of iterations by one. If w' is equal to w or the empty word, then return the number of iterations. Otherwise go to step 2.

We say that the word w *stabilizes* after the obtained number of iterations. Observe that a reduced word 1 1 can be regarded as a trivial repeat or return word, but the iterative steps stop at the moment w becomes 1 1. Details of the above algorithm are included in the Supplementary Materials under *Nested Repeat–Return Algorithm* section.

2. Results

We examine the rearrangement maps corresponding to $\mathcal{D}$ according to their respective reduced patterns and assembly graphs. In Section 2.1 we analyze the number of consecutive, nonscrambled MDSs contained in the micronuclear loci in $\mathcal{D}$, and describe the most frequent patterns that occur upon reducing their rearrangement maps. In Section 2.2 we show the assembly graphs corresponding to the most frequent reduced patterns in $\mathcal{D}$, and investigate their graph structures in terms of nested repeat–return patterns.

2.1. Frequency of rearrangement patterns

In order to characterize the scrambled patterns that occur in the *O. trifallax* genome, we investigated all micronuclear rearrangement maps in $\mathcal{D}$. Most (87.2%) MIC loci in $\mathcal{D}$ are nonscrambled, i.e., they simply contain consecutive MDSs interrupted by IESs (note that this conclusion differs from Chen et al. (2014) because $\mathcal{D}$ is a subset of the data). The histogram in Fig. 6 represents the number of consecutive MDSs merged (IESs removed) during the rearrangement reducing process, as described in Section 1.2. The blue histogram represents the total number of MDSs merged in each nonscrambled rearrangement map in $\mathcal{D}$. The red portion of the histogram shows the number of consecutive, nonscrambled MDSs present in all scrambled rearrangement maps in $\mathcal{D}$. These nonscrambled MDSs can likewise be merged during the reduction. Fig. 6 shows that most of the rearrangement maps

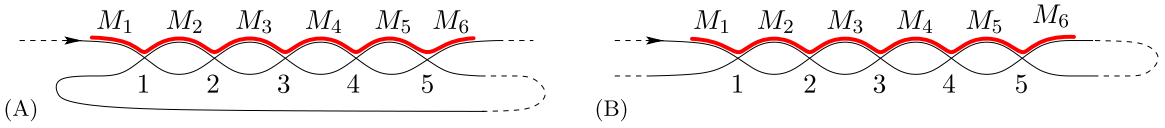


Fig. 4. Assembly graphs corresponding to (A) repeat and (B) return words, associated with odd–even patterns.

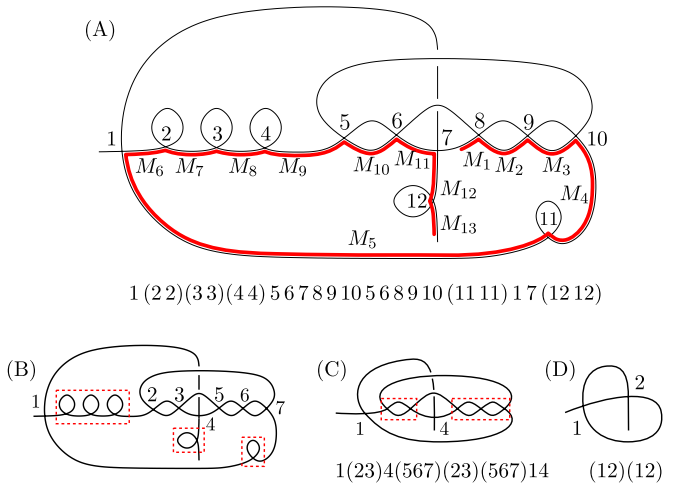


Fig. 5. (A) Assembly graph representation of Contig19801.0, highlighting nested sequences of repeat words and the corresponding DOW. Vertices are labeled in ascending order, and those that correspond to pointers of consecutive MDSs (incident to loops) are in parenthesis. (B) Reducing the rearrangement map (as described in Section 1.2) corresponds to removal of the loops (red boxes). The reduced DOW 156789(10)5689(10)17 relabeled in ascending order is 12345672356714. (C) The repeats 23 and 567 (indicated with parentheses in the DOW and red boxes in the graph) are removed in the first iteration to yield the word 1414, which is 1212 (D) in ascending order and a repeat. Figure is not drawn to scale. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

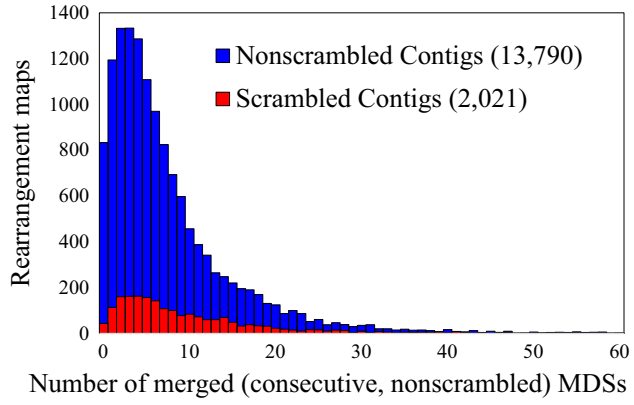


Fig. 6. Number of consecutive, nonscrambled MDSs present in nonscrambled loci (blue) and scrambled loci (red). These can be respectively merged in the rearrangement map reduction process. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

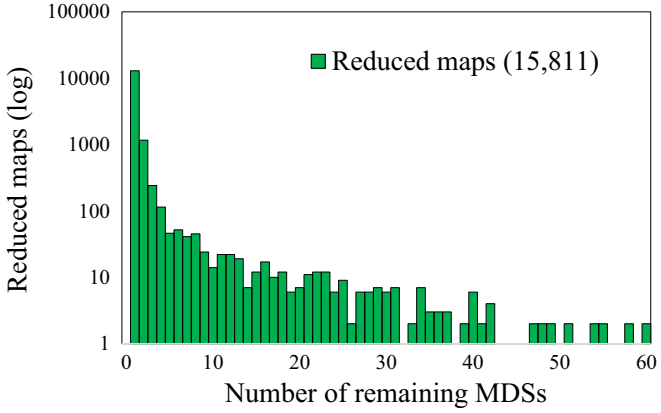


Fig. 7. Number of MDSs remaining in the reduced rearrangement maps. All nonscrambled loci contain only one MDS in their reduced form (hence the spike at 1), whereas the scrambled rearrangement maps have multiple MDSs.

have either 2 (1332 nonscrambled and 160 scrambled loci) or 3 (1333 nonscrambled and 162 scrambled loci) consecutive, nonscrambled MDSs, representing 18.9% of the rearrangement maps in $\mathcal{D}$. For nonscrambled rearrangement maps, the number of MDSs merged during the reduction process equals the number of MDSs in the original rearrangement map, and therefore this histogram reflects the total number of MDSs in these cases (similar to Fig. 2A in Chen et al. (2014)). In contrast, the number of MDSs that can be reduced in the scrambled rearrangement maps is sometimes zero or slightly fewer than the number of MDSs annotated for this locus in the MIC genome. 51.5% of the MAC contigs contain between 1-6 nonscrambled (consecutive) MDSs that can be merged in this process. The maximal number of MDSs that can be merged is 149, and belongs to the scrambled gene Contig13198.0 which has total of 154 MDSs (including a total of 149 nonscrambled MDSs). Fig. 7 lists the number of MDSs remaining in the reduced rearrangement maps. The most complex case is the scrambled Contig16618.0.1, which has 89 scrambled segments, none of which can be reduced.

The dataset $\mathcal{D}$ contains 2021 scrambled rearrangement maps for MAC contigs. Here we analyze their reduced forms. We follow the convention to describe pattern representatives as described at the end of Section 1.2. The patterns appear as five basic types that describe over 90% of the scrambled rearrangement maps. *Translocations* are scrambled patterns containing a block of consecutive MDSs repositioned elsewhere in the MIC genome but in the same orientation. For instance, M_2M_1, $M_1M_3M_2$, and $M_1M_3M_2M_4$ are translocations in Fig. 8(A). *Inversions* are scrambled patterns with reduced representation of the form $M_1\overline{M}_2$ or $M_1\overline{M}_2M_3$ (Fig. 8(B)). A *translocation-inversion* pattern refers to the situation in which a block of MDSs are repositioned in the MIC contig in the opposite orientation (Fig. 8(C)). An example of a reduced pattern representative with translocation-inversion is $M_1M_3\overline{M}_2$.

The fourth type of pattern frequently observed is a sequence of increasing odd numbered MDSs followed by increasing even numbered MDSs, or an *odd–even repeat pattern*, for example $M_1M_3M_5M_2M_4M_6$. A schematic representation of this pattern is shown in Fig. 8(D). The fifth type of pattern observed is a sequence of increasing odd numbered MDSs followed by a decreasing and inverted even numbered MDSs, which we call an *odd–even return pattern*, for example $M_1M_3M_5\overline{M}_6\overline{M}_4\overline{M}_2$. The schematics of this pattern is represented in Fig. 8(E). Return and repeat pattern types correspond to return and repeat DOWs, and their associated assembly graphs are depicted in Fig. 4.

Table 2 lists the frequencies of each scrambled pattern that occurs in the *Oxtricha trifallax* genome annotations (Chen et al., 2014). Among the 2021 scrambled loci, 854 (42.3% of the scrambled rearrangement maps) are represented by the pattern $M_1\overline{M}_2$, i.e., consist of a single inversion at one end of the map. By contrast, a single inversion in the interior of the MIC loci, $M_1\overline{M}_2M_3$, occurs only 27 times (1.3% of the scrambled rearrangement maps). After

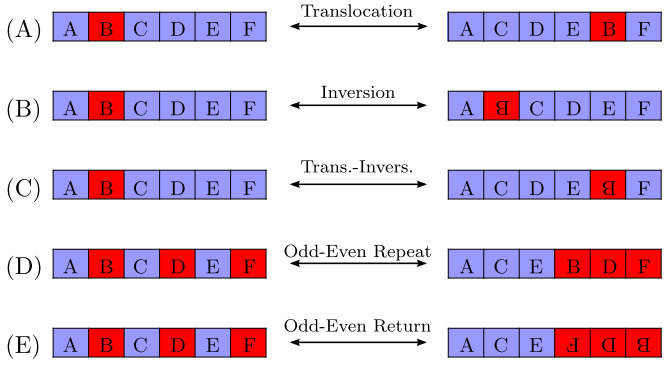


Fig. 8. Schematic representation of the most frequently occurring types of scrambled patterns in *O. trifallax*.

Table 2
Frequency of different patterns in the MIC loci in *Oxytricha trifallax*. The table includes pattern representatives as described in Section 1.2. For example, the rearrangement map $M_2\,M_4\ldots M_{2k}\,M_1\,M_3\ldots M_{2k+1}$ is not listed in the table because its pattern is equivalent to $M_1\,M_3\ldots M_{2k+1}\,M_2\,M_4\ldots M_{2k}$.

Pattern type	Locus	Reduced MIC pattern representative	Count
Non-scrambled (13,790)	–	M_1	13,790
Inversion (881)	End	$M_1\,\overline{M}_2$	854
	Inner	$M_1\,\overline{M}_2\,M_3$	27
Translocation (421)	End-End	$M_2\,M_1$	307
	End-Inner	$M_1\,M_3\,M_2$	79
	Inner-Inner	$M_1\,M_3\,M_2\,M_4$	35
Translocation and Inversion (59)	Inner-End	$M_1\,M_3\,\overline{M}_2$	34
	End-Inner	$M_1\,\overline{M}_3\,M_2$	21
	Inner-Inner	$M_1\,M_3\,\overline{M}_2\,M_4$	4
Odd–Even Global Patterns (290)	Repeat	$M_1\,M_3\ldots M_{2k-1}\,M_2\,M_4\ldots M_{2k}$ $(k>2)$	99
		$M_1\,M_3\ldots M_{2k+1}\,M_2\,M_4\ldots M_{2k}$ $(k>2)$	96
		$M_2\,M_4\ldots M_{2k}\,M_1\,M_3\ldots M_{2k-1}$ $(k>1)$	29
	Return	$M_1\,M_3\ldots M_{2k-1}\,\overline{M}_{2k}\ldots\,\overline{M}_4\,\overline{M}_2$ $(k>1)$	40
		$M_1\,M_3\ldots M_{2k+1}\,\overline{M}_{2k}\ldots\,\overline{M}_4\,\overline{M}_2$ $(k>1)$	26
Other (370)	–	See Supplementary Materials	370

inversion, the next most common rearrangement event is translocation, among which the most common reduced map is the circular permutation M_2M_1 (307 occurrences). Odd–even patterns appear quite often (290 times) while translocation-inversion occurs least often (only 59 occurrences). These five types of patterns account for 81.7% of the scrambled loci. Of the remaining 370 types of scrambled loci, 243 are composites of two of the five recurrent patterns and these account for 93.7% of all scrambled reduced loci. For example, the pattern $M_1\overline{M}_4\overline{M}_2M_3$ can be explained as the odd–even return pattern $M_1M_3\overline{M}_4\overline{M}_2$ together with a translocation of M_3 to the end of the sequence. A complete list of all patterns that occur at least twice is included in the Supplementary Materials.

2.2. Graph patterns of scrambled rearrangement maps

We now consider the DOWs and graph structures corresponding to all reduced rearrangement maps in $\mathcal{D}$. The set of DOWs (written in ascending order) obtained from MDS lists in $\mathcal{D}$ consists of 15,811 words. As in Section 1.2 we reduce each DOW by removing the symbols that correspond to pointers of consecutive, nonscrambled MDSs (M_aM_{a+1}) in the original rearrangement map, i.e., that appear as *aa* in the DOW. We denote with $\mathcal{W}$ the list of all reduced double occurrence words obtained from $\mathcal{D}$. The list $\mathcal{W}$ contains DOWs of length 0 through 88, and the words of up to three symbols are summarized in Table 3.

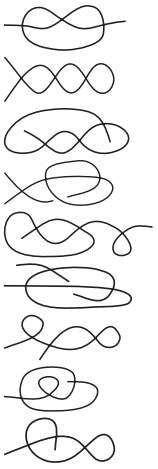
834 MAC contigs (5.3%) lack IESs and thus consist of a single MDS. Their rearrangement map corresponds to the empty DOW. 12,956 nonscrambled MAC contigs contain more than one MDS (maximum 120) and these contigs correspond (before reduction) to DOWs consisting of a list of consecutive identical letters, e.g., 1122… which also reduce to the empty word. Hence, 87.2% of all MAC contigs in $\mathcal{D}$ represent nonscrambled MAC contigs (first row of Table 2). We concentrate below on the remaining 2021 (or 11.8%) scrambled MAC contigs.

Over half of all scrambled loci, or 1161 MAC contigs have a reduced representation as the DOW *aa* for a single letter *a*. This means that the original DOW sequence can be written as $w = w_1aw_2a$, where both w_1 and w_2 comprise a string of nonscrambled MDSs, represented as a sequence of consecutive pairs of identical letters that reduces to the empty word. For example, 1122344553 has this property, where $a=3$, $w_1 = 1122$, and $w_2 = 4455$. There are only two nonequivalent patterns that correspond to this situation: $M_1\overline{M}_2$ and M_2M_1 (compare with Table 2). The distributions of occurrences for each of the two situations are given in Fig. 9, and their assembly graphs are depicted in Fig. 10. The rearrangement map $M_1\cdots M_k\overline{M}_n\overline{M}_{n-1}\cdots\overline{M}_{k+1}$ for some k in Fig. 10 (A) corresponds to the pattern $M_1\overline{M}_2$, and $M_{k+1}M_{k+2}\cdots M_nM_1\cdots M_k$ (Fig. 10(B)), corresponds to the pattern M_2M_1.

There are three possibilities for DOWs with two symbols: *aabb*, *abab*, and *abba*. Table 3 shows the numbers of appearances of words for each case among the scrambled contigs. These graphs correspond to patterns with three segments (after merging MDSs). For three letters, the DOWs and their occurrences are listed in Table 3. Curiously, the repeat word 123123 appears in the list of reduced DOWs far more often (60 times) than the others (15 or less). The third column of the table indicates the number of patterns detected versus the number of possible patterns reflected by the same word. For example, out of 60 occurrences of the repeat word 123123, 53 are represented by the odd–even repeat pattern: $M_1M_3M_2M_4$, suggesting that it is highly overrepresented. We investigate this trend further by subjecting the data to nested repeat–return pattern analysis as described at the end of Section 1.5.

Table 3
Double occurrence words and assembly graphs corresponding to reduced rearrangement maps that contain four or fewer MDSs. The remaining 505 scrambled MAC contigs, with reduced rearrangement maps of five or more MDSs, can be found in the Supplementary Materials.

DOW	Contigs	Patterns
-[a]	834	1
∅[b]	12,956	1
Total DOW	13,790 (87.2%) Contigs	1 Patterns
11	1161	2
Total DOW	1161 (7.3%) Contigs	2 Patterns
1212	177	7
1221	47	2
1122	17	1
Total DOW	241 (1.5%) Contigs	10 Patterns

DOW	Contigs	Patterns	Graphs
123123	60	5 of 10	
123321	15	2 of 2	
123213	12	6 of 20	
123231	9	5 of 11	
121233	6	1 of 4	
121323	4	1 of 14	
122133	4	1 of 1	
122313	3	3 of 6	
123312	1	1 of 2	
Total	114 (0.7%)		

[a] Original word is ∅ which reduces to ∅.
[b] Original word is 1122… which also reduces to ∅.

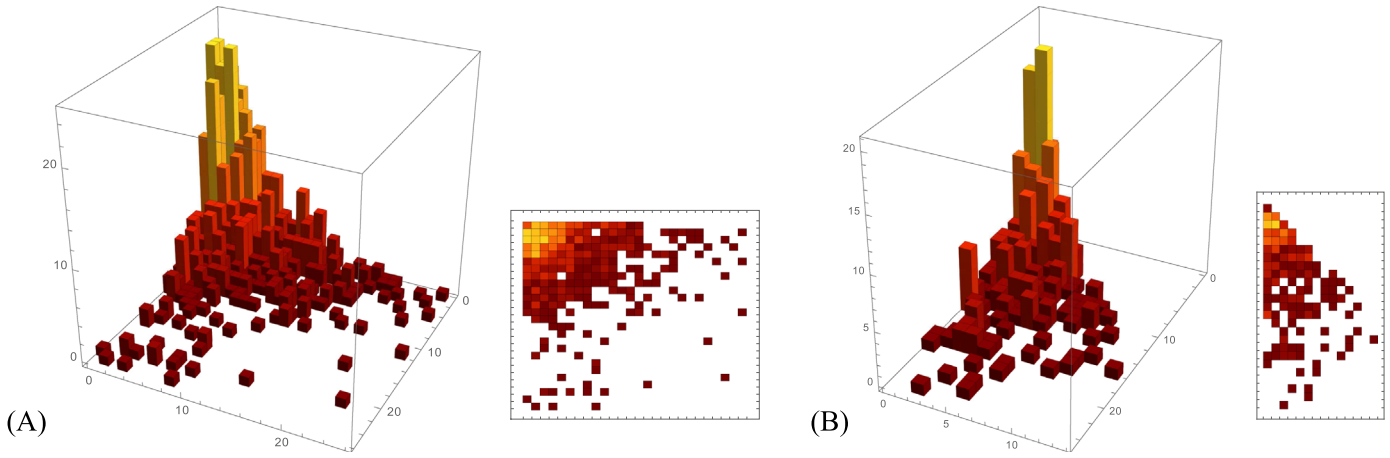


Fig. 9. Contigs that reduce to the word "11" corresponding to pattern (A) $M_1\overline{M}_2$ and (B) M_2M_1. The x-axis and y-axis represent the number of consecutive MDSs merged into M_1 and M_2 respectively, the z-axis shows the number of MAC contigs. In (B), due to the symmetry of the pattern M_2M_1, we chose M_1 to represent the reduced arrangement with more consecutive MDSs merged than M_2.

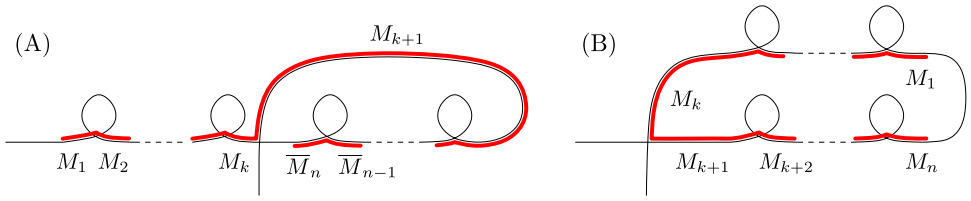


Fig. 10. Two assembly graphs with reduced DOW "11", corresponding to the rearrangement maps (A) $M_1\cdots M_k\overline{M}_n\overline{M}_{n-1}\cdots\overline{M}_{k+1}$ and (B) $M_{k+1}M_{k+2}\cdots M_nM_1\cdots M_k$ with respective patterns $M_1\overline{M}_2$ and M_2M_1.

Table 4
Number of repeat–return iterations necessary for the reduced DOWs with two or more letters, which correspond to rearrangement maps of scrambled genes, to stabilize. (A) Double occurrence words that iterate to the empty word correspond to genes that could arise from multiple iterations of the process that generates repeat–return patterns. On the other hand, words that stabilize before reaching the empty word (B) contain scramble patterns that are more complex than nested instances of the repeat–return patterns.

Repeat–return iterations	DOWs of scrambled maps
1	634
2	37
3	11
4	2
Reduced[a]	176
Total	860
Repeat–return iterations	**DOWs of irreducible maps**
0	6
1	145
2	18
3	5[b]
4	1[c]
5	1[d]
Total	176

[a] Not a nested sequence of the repeat and return patterns.
[b] Contig7885.0, Contig22552.0, Contig15736.0, Contig11366.0.0, and Contig5956.0.
[c] Contig12243.0.
[d] Contig19180.0.

After the first iteration removing maximal repeat and return words, out of 860 contigs corresponding to reduced words with at least two symbols, 634 contigs have words that reduce to the empty word, and the remaining 226 stabilize after 5 iterations. Table 4(A) indicates the number of MAC contigs whose words stabilize to the empty word after 1, 2, 3, and 4 iterations, respectively. All patterns in Table 2, except $M_1\,M_3\,\overline{M}_2\,M_4$ (which corresponds to translocation-inversion of an inner segment within an inner portion of the molecule) are special cases of repeat or return words. The number of contigs whose words do not stabilize to the empty word is 176. Four stabilize to words correspond to pattern $M_1\,M_3\,\overline{M}_2\,M_4$ and all the rest stabilize to patterns included in "other" in Table 2. Out of these 176 words, 103 stabilize to a word *aa*. Considering that *aa* corresponds to the simplest odd–even patterns, M_2M_1 and $M_1\overline{M}_2$, only 73 patterns, or 3.6% of all scrambled loci cannot be explained with nested combinations of repeats and returns. Therefore, approximately 96.4% of all scrambled MIC contigs have patterns that require various combinations of nested repeat and return words for their description.

Computations from the algorithm in Section 1.5 indicate that all words stabilize after 5 iterations. Table 4 lists the 176 words that do not stabilize to the empty word, but stabilize after 0, 1, 2, 3, 4, and 5 iterations. Of those that do not stabilize to *aa*, 32 stabilize to a word *aabb*. Of the remaining 41, 19 have 3 symbols and 22 have 4 or more symbols, and it appears that the MDS rearrangement maps in those 41 contigs have varied scrambled patterns. Most of them have a single occurrence in the set $\mathcal{W}$.

3. Discussion

In this paper, we initiate and introduce a mathematical formulation for analyzing patterns in gene rearrangements, and we apply this method to the completely sequenced, encrypted genome of *O. trifallax*. Rearrangements (as processes) have been studied with string rewriting operations (e.g. Ehrenfeucht et al., 2002; Angeleska et al., 2012) and graph operations (e.g. Ehrenfeucht et al., 2002; Angeleska et al., 2007), but the patterns in which segments appear in the scrambled genome and the extent

of their appearance have not been considered, especially on a genome-wide scale. The notion of a rearrangement map pattern – as defined here – describes the relative sequence rearrangement between genetic precursor and product sequences, and it can be described with double occurrence words and also depicted with spatial graphs. Even the earliest sequencing of scrambled genes Mitcham et al., 1992; Hoffman and Prescott, 1996 revealed the presence of even–odd patterns (the most prevalent patterns by our analysis). The description of patterns in rearrangement maps via assembly graphs offers a tool to classify and characterize the patterns. The recent sequencing of two (precursor and product) *O. trifallax* genomes (Swart et al., 2013; Chen et al., 2014) permits a comprehensive study of the scrambled patterns in the micronuclear genome. Our study describes the patterns that appear at greatest frequency. We suggest that the theoretical models of gene rearrangement should focus on operations that can be applied to those frequent patterns.

The vast majority of the micronuclear genes in *O. trifallax* (~87.2% in data $\mathcal{D}$) are nonscrambled and can therefore be described by simple nonscrambled rearrangement maps, similar to the situation in many other distant species of ciliates, such as *Paramecium*. In addition, many large regions of the scrambled micronuclear gene loci contain high number of consecutive, nonscrambled MDSs separated by IESs. Therefore their rearrangement maps, after compression of all the sequential nonscrambled MDSs, simply reduce to the same five fundamental types as seen before (e.g. Prescott and Greslin, 1992): inversion, translocation, translation-inversion, and various odd–even patterns. However, we note that inversion and translocation-inversion are the simplest instance of an odd–even return pattern (Landweber, 1998), and translocation is the simplest version of the odd–even repeat pattern. From this perspective, the odd–even repeat and return patterns account for 81.7% of the 2021 scrambled MAC contigs in $\mathcal{D}$. Almost all scrambled MIC contigs (96.4%) of the studied annotation $\mathcal{D}$ are nested appearances of repeat-return words. Development of theoretical models that address rearrangement processes of these patterns, as well as experimental findings of intermediate molecules appearing during the rearrangement process may be of interest. The most complex rearrangement maps that do not conform to the iterative odd–even repeat-return patterns will be the subject of another study.

The evolutionary expansion of these recurring patterns in the germline loci must themselves have arisen from germline meitoic or mitotic recombination events that produce inversions and translocations in the MIC. This also suggests that the germline micronuclei may be particularly prone to recombination events that introduce or propagate the recurring odd/even split patterns in the genome, as proposed in Landweber (1998) and described empirically in Chang et al. (2005), Wong and Landweber (2006), Hogan et al. (2001). Chang et al. (2005), in particular, noted the evolutionary trend towards expansion of the number of MDSs within odd–even clusters among later diverging lineages, including *Oxytricha*. More comparisons among orthologous genes in other species with scrambled genomes will reveal whether these patterns are conserved or embellished over evolutionary time. Once a locus has become scrambled, the chromosomal region may continue to become more complex during germline evolution via additional translocations or inversions (Chang et al., 2005). Hence the iterative process that this study describes for reducing the scrambled rearrangement maps to their underlying simpler ones may allow us to estimate how many layers of germline recombination events have historically shaped these genomic regions. Comparative germline genomics among closely related species will be the best way to test these hypotheses and to infer the evolutionary steps through which a genome became so densely scrambled.

Acknowledgments

We wish to thank Kelsi Lindblad, Talya Yerlici, and Leslie Beh for their valuable comments. This study was supported by NIH grant GM109459 and NSF grant CCF-1526485.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2016.08.038.

References

Aeschlimann, S.H., Jönsson, F., Postberg, J., Stover, N.A., Petera, R.L., Lipps, H.-J., Nowacki, M., Swart, E.C., 2014. The draft assembly of the radically organized Stylonychia lemnae macronuclear genome. Genome Biol. Evol. 6 (7), 1707–1723. http://dx.doi.org/10.1093/gbe/evu139.

Angeleska, A., Jonoska, N., Saito, M., Landweber, L.F., 2007. RNA-guided DNA assembly. J. Theor. Biol. 248 (4), 706–720. http://dx.doi.org/10.1016/j.jtbi.2007.06.007.

Angeleska, A., Jonoska, N., Saito, M., 2009. DNA recombination through assembly graphs. Discrete Appl. Math. 157 (14), 3020–3037. http://dx.doi.org/10.1016/j.dam.2009.06.011.

Angeleska, A., Jonoska, N., Saito, M., 2012. Rewriting rule chains modeling DNA rearrangement pathways. Theor. Comput. Sci. 454, 5–22. http://dx.doi.org/10.1016/j.tcs.2012.04.041.

Arredondo, R., 2013. On double occurrence words. Congr. Numer. 218, 43–56.

Bracht, J.R., Fang, W., Goldman, A.D., Dolzhenko, E., Stein, E.M., Landweber, L.F., 2013. Genomes on the edge: programmed genome instability in ciliates. Cell 152 (3), 406–416. http://dx.doi.org/10.1016/j.cell.2013.01.005.

Chang, W.-J., Bryson, P.D., Liang, H., Shin, M.K., Landweber, L.F., 2005. The evolutionary origin of a complex scrambled gene. Proc. Natl. Acad. Sci. U.S.A. 102 (42) 15149–15154, http://dx.doi.org/10.1073/pnas.0507682102.

Chen, X., Bracht, J.R., Goldman, A.D., Dolzhenko, E., Clay, D.M., Swart, E.C., Perlman, D.H., Doak, T.G., Stuart, A., Amemiya, C.T., Sebra, R.P., Landweber, L.F., 2014. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell 158 (5), 1187–1198. http://dx.doi.org/10.1016/j.cell.2014.07.034.

Ehrenfeucht, A., Harju, T., Rozenberg, G., 2002. Gene assembly through cyclic graph decomposition. Theor. Comput. Sci. 281 (1–2), 325–349. http://dx.doi.org/10.1016/S0304-3975(02)00019-1, Selected Papers in Honour of Maurice Nivat.

Ehrenfeucht, A., Harju, T., Petre, I., Prescott, D.M., Rozenberg, G., 2004. In: Computation in Living Cells: Gene Assembly in Ciliates, Natural Computing Series, Springer, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-662-06371-2.

Eisen, J.A., Coyne, R.S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J.R., Badger, J.H., Ren, Q., Amedeo, P., Jones, K.M., Tallon, L.J., Delcher, A.L., Salzberg, S.L., Silva, J.C., Haas, B.J., Majoros, W.H., Farzad, M., Carlton, J.M., Smith Jr., R.K., Garg, J., Pearlman, R.E., Karrer, K.M., Sun, L., Manning, G., Elde, N.C., Turkewitz, A.P., Asai, D.J., Wilkes, D.E., Wang, Y., Cai, H., Collins, K., Stewart, B.A., Lee, S.R., Wilamowska, K., Weinberg, Z., Ruzzo, W.L., Wloga, D., Gaertig, J., Frankel, J., Tsao, C.-C., Gorovsky, M.A., Keeling, P.J., Waller, R.F., Patron, N.J., Cherry, J.M., Stover, N.A., Krieger, C.J., del Toro, C., Ryder, H.F., Williamson, S.C., Barbeau, R.A., Hamilton, E.P., Orias, E., 2006. Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol. 4 (9), e286. http://dx.doi.org/10.1371/journal.pbio.0040286.

Hoffman, D., Prescott, D.M., 1996. The germline gene encoding dna polymerase alpha in the hypotrichous ciliate oxytricha nova is extremely scrambled. Nucl. Acids Res. 24, 3337–3340. http://dx.doi.org/10.1093/nar/24.17.3337.

Hogan, D.J., Hewitt, E.A., Orr, K.E., Prescott, D.M., Müller, K.M., 2001. Evolution of IESs and scrambling in the actin I gene in hypotrichous ciliates. Proc. Natl. Acad. Sci. U.S.A. 98 (26) 15101–15106, http://dx.doi.org/10.1073/pnas.011578598.

Kari, L., Landweber, L.F., 1999. Computational power of gene rearrangement. In: Winfree, E., Gifford, D.K. (Eds.), Proceedings of DNA Based Computers V, DIMACS: Series in Discrete Mathematics and Theoretical Computer Science, vol. 54, American Mathematical Society, Providence, RI, pp. 207–216.

Landweber, L.F., 1998. The evolution of DNA computing: nature's solution to a path problem. In: IEEE Proceedings of Symposia on Intelligence and Systems '98, May 21–23, IEEE Computer Society Press, 1998, pp. 133–139.

Möllenbeck, M., Zhou, Y., Cavalcanti, A.R., Jönsson, F., Higgins, B.P., Chang, W.-J., Juranek, S., Doak, T.G., Rozenberg, G., Lipps, H.J., Landweber, L.F., 2008. The pathway to detangle a scrambled gene. PLoS One 3 (6), e2330. http://dx.doi.org/10.1371/journal.pone.0002330.

Mitcham, J.L., Lynn, A.J., Prescott, D.M., 1992. Analysis of a scrambled gene: the gene encoding α-telomere-binding protein in oxytricha nova. Gene Dev. 6 (5), 788–800. http://dx.doi.org/10.1101/gad.6.5.788.

Nowacki, M., Vijayan, V., Zhou, Y., Schotanus, K., Doak, T.G., Landweber, L.F., 2008. RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature 451 (7175), 153–158. http://dx.doi.org/10.1038/nature06452.

Prescott, D.M., Greslin, A.F., 1992. Scrambled actin I gene in the micronucleus of

Oxytricha nova. Dev. Genet. 13 (1), 66–74. http://dx.doi.org/10.1002/dvg.1020130111.

Prescott, D.M., Ehrenfeucht, A., Rozenberg, G., 2003. Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates. J. Theor. Biol. 222 (3), 323–330. http://dx.doi.org/10.1016/S0022-5193(03)00037-7.

Ricard, G., de Graaf, R., Dutilh, B., Duarte, I., van Alen, T., van Hoek, A., Boxma, B., van der Staay, G., Moon-van der Staay, S., Chang, W.-J., Landweber, L.F., Hackstein, J., Huynen, M., 2008. Macronuclear genome structure of the ciliate Nyctotherus ovalis: single-gene chromosomes and tiny introns. BMC Genom. 9 (1), 587. http://dx.doi.org/10.1186/1471-2164-9-587.

Swart, E.C., Bracht, J.R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J.S., Goldman, A.D., Nowacki, M., Schotanus, K., Jung, S., Fulton, R.S., Ly, A., McGrath, S., Haub, K., Wiggins, J.L., Storton, D., Matese, J.C., Parsons, L., Chang, W.-J., Bowen, M.S., Stover, N.A., Jones, T.A., Eddy, S.R., Herrick, G.A., Doak, T.G., Wilson, R.K., Mardis, E.R., Landweber, L.F., 2013. The Oxytricha trifallax macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol. 11 (1), e1001473. http://dx.doi.org/10.1371/journal.pbio.1001473.

Vinogradov, D., Tsoi, O., Zaika, A., Lobanov, A., Turanov, A., Gladishev, V., Gelfand, M., 2012. Draft macronucleus genome of Euplotes crassus ciliate. Mol. Biol. 46 (2), 328–333. http://dx.doi.org/10.1134/S0026893312020197.

Wong, L.C., Landweber, L.F., 2006. Evolution of programmed DNA rearrangements in a scrambled gene. Mol. Biol. Evol. 23 (4), 756–763. http://dx.doi.org/10.1093/molbev/msj089.

Yerlici, V.T., Landweber, L.F., 2014. Programmed genome rearrangements in the Ciliate Oxytricha. Microbiol. Spectr. 2 (6). http://dx.doi.org/10.1128/microbiolspec.MDNA3-0025-2014, MDNA3–0025–2014.

Supplemental Information

S1 Data Filtering

Filtering of the MDS annotation from Chen et al. "The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development" ***Cell*****, 158-5, (2014) 1187-1198.**

If the intervals $[a, a']$ and $[b, b']$ in a MIC DNA sequence correspond to the MDS annotations M and M' respectively, then M and M' are *intersecting* if $[a, a']$ intersects $[b, b']$ and *adjacent* if $b = a' + 1$ or $a = b' + 1$.

The dataset $\mathcal{D}$ is filtered from Chen et al. (2014) in the following way:

Step 1. Merge any two MDS annotations that are consecutive in a MAC contig and overlap or are adjacent in a MIC contig.

Step 2. Remove any two MDS annotations that are non-consecutive in a MAC contig and overlap or are adjacent in a MIC contig.

Step 3. Remove all duplicate MDSs, e.g., MDSs that correspond to MAC contigs that produced from alternative MIC processing.

Three processed files are included at `http://knot.math.usf.edu/data/scrambled_patterns/`.

1. processed_annotation_of_oxy_tri.gff contains all of the processed data $\mathcal{D}$.
2. list_of_DOWs_related_to_MICMAC_maps.txt contains all rearrangement maps, corresponding double occurrence words, and the reduced double occurrence words.
3. result_of_repeatreturn_pattern_application.txt contains the list of all 176 words that stabilize to non-empty words after performing the iterative repeat/return removal.

S2 Extended Table of Rearrangement Patterns

Reduced Rearrangement Patterns occurring in *O. trifallax*

Count	Reduced MIC pattern representative
854	$M_1\overline{M}_2$
307	M_2M_1
79	$M_1M_3M_2$
35	$M_1M_3M_2M_4$
34	$M_1M_3\overline{M}_2$
27	$M_1\overline{M}_2M_3$
24	$M_1M_3M_5M_7M_2M_4M_6M_8$
22	$M_1M_3M_5M_2M_4M_6$
21	$M_1\overline{M}_3M_2$
18	$M_2M_4M_1M_3$
17	$M_3\overline{M}_2M_1$
16	$M_1M_3M_5M_2M_4$
14	$\overline{M}_1M_3M_2$
14	$\overline{M}_2M_3M_1$
14	$M_1M_3\overline{M}_4\overline{M}_2$
14	$M_1M_3M_5M_7M_2M_4M_6$
13	$M_2\overline{M}_3M_1$
13	$M_3M_2\overline{M}_1$
11	$M_1M_3M_5M_7M_9M_2M_4M_6M_8$
11	$M_1M_3M_5M_7M_9M_{11}M_2M_4M_6M_8M_{10}M_{12}$
10	$M_1M_3M_5M_7M_9M_{11}M_{13}M_2M_4M_6M_8M_{10}M_{12}$
9	$M_3M_2M_1$
9	$M_1M_3M_5\overline{M}_6\overline{M}_4\overline{M}_2$
9	$M_1M_3M_5M_7M_9M_2M_4M_6M_8M_{10}$
7	$M_1M_3M_5\overline{M}_4\overline{M}_2$
7	$M_1M_3M_5M_7M_9M_{11}M_{13}M_{15}M_2M_4M_6M_8M_{10}M_{12}M_{14}M_{16}$
6	$\overline{M}_1M_2M_4M_3$
6	$M_1M_3M_5M_7M_9M_{11}M_2M_4M_6M_8M_{10}$
5	$M_2M_4M_3M_1$
5	$M_3M_2M_4M_1$
5	$M_1M_3M_5M_7M_9\overline{M}_8\overline{M}_6\overline{M}_4\overline{M}_2$
5	$M_1M_3\ldots M_{15}M_{17}M_2M_4\ldots M_{16}M_{18}$
5	$M_1M_3\ldots M_{19}M_{21}M_2M_4\ldots M_{18}M_{20}$
5	$M_1M_3\ldots M_{25}M_{27}M_2M_4\ldots M_{24}M_{26}$
4	$\overline{M}_1M_2M_4\overline{M}_3$
4	$M_1M_3\overline{M}_2M_4$
4	$M_1\overline{M}_4\overline{M}_2M_3$
4	$M_3M_5M_2M_4M_1$

Continued on next page

Count	Reduced MIC pattern representative
4	$M_1M_3M_5M_7\overline{M}_8\overline{M}_6\overline{M}_4\overline{M}_2$
4	$M_1M_3\ldots M_{13}M_{15}M_2M_4\ldots M_{12}M_{14}$
4	$M_1M_3\ldots M_{19}M_{21}M_2M_4\ldots M_{20}M_{22}$
4	$M_1M_3\ldots M_{21}M_{23}M_2M_4\ldots M_{20}M_{22}$
4	$M_1M_3\ldots M_{23}M_{25}M_2M_4\ldots M_{22}M_{24}$
3	$M_1M_3M_5M_7\overline{M}_6\overline{M}_4\overline{M}_2$
3	$M_2M_4M_6\overline{M}_7\overline{M}_5\overline{M}_3M_1$
3	$M_1M_3M_5M_7M_9M_{11}\overline{M}_{10}\overline{M}_8\overline{M}_6\overline{M}_4\overline{M}_2$
3	$M_1M_3\ldots M_{15}M_{17}M_2M_4\ldots M_{14}M_{16}$
3	$M_1M_3\ldots M_{27}M_{29}M_2M_4\ldots M_{28}M_{30}$
2	$M_1\overline{M}_2\overline{M}_4M_3$
2	$M_1M_3M_5\overline{M}_2\overline{M}_4$
2	$M_2M_4M_3M_5M_1$
2	$M_1\overline{M}_2M_5\overline{M}_3\overline{M}_6M_4$
2	$M_2M_4M_6M_1M_3M_5$
2	$M_3M_5M_2M_4M_6M_1$
2	$M_2M_1M_3\overline{M}_6\overline{M}_7M_5\overline{M}_4$
2	$M_2M_4M_6M_3M_5M_7M_1$
2	$M_4M_6M_2M_5M_7\overline{M}_3\overline{M}_1$
2	$M_2M_4M_6M_8M_1M_3M_5M_7$
2	$M_2M_6M_8M_4M_7M_9\overline{M}_5\overline{M}_3\overline{M}_1$
2	$M_1M_3M_5M_7M_9M_{11}\overline{M}_{12}\overline{M}_{10}\overline{M}_8\overline{M}_6\overline{M}_4\overline{M}_2$
2	$M_2M_4M_6M_8M_{10}M_{12}M_1M_3M_5M_7M_9M_{11}$
2	$M_1M_3M_5M_7M_9M_{11}M_{13}\overline{M}_{12}\overline{M}_{10}\overline{M}_8\overline{M}_6\overline{M}_4\overline{M}_2$
2	$M_1M_3\ldots M_{11}M_{13}M_2M_4\ldots M_{12}M_{14}$
2	$M_1M_3\ldots M_{15}M_{17}\overline{M}_{18}\overline{M}_{16}\ldots\overline{M}_4\overline{M}_2$
2	$M_1M_3\ldots M_{17}M_{19}M_2M_4\ldots M_{16}M_{18}$
2	$M_1M_3\ldots M_{17}M_{19}M_2M_4\ldots M_{18}M_{20}$
2	$M_1M_3M_5M_2M_4\ldots M_{20}M_{22}M_7M_9\ldots M_{21}M_{23}$
2	$M_1M_3\ldots M_{27}M_{29}M_2M_4\ldots M_{26}M_{28}$
2	$M_1M_3\ldots M_{29}M_{31}M_2M_4\ldots M_{28}M_{30}$
2	$\overline{M}_1M_3M_5\ldots M_{19}M_{21}\overline{M}_{37}\overline{M}_{35}\ldots\overline{M}_{25}\overline{M}_{23}M_2M_4\ldots M_{34}M_{36}$
2	$M_1M_3\ldots M_{37}M_{39}\ldots M_2M_4\ldots M_{38}M_{40}$
273	Other

Table 1: Reduced scrambled rearrangement patterns of *O. trifallax* listed in order of frequency.

DOW	Reduced Pattern	Count
123213	$M_1\overline{M}_4\overline{M}_2M_3$	4
	$M_1M_3\overline{M}_2M_4$	4
	$\overline{M}_2M_4M_3M_1$	1
	$\overline{M}_2\overline{M}_4M_3M_1$	1
	$M_1\overline{M}_3M_4M_2$	1
	$\overline{M}_3M_2M_4M_1$	1
	$M_1M_4M_3M_2$	0
	$\overline{M}_1M_4M_3M_2$	0
	$M_1\overline{M}_4M_3M_2$	0
	$\overline{M}_1\overline{M}_4M_3M_2$	0
	$\overline{M}_1M_4\overline{M}_2M_3$	0
	$M_1\overline{M}_3\overline{M}_4M_2$	0
	$M_2M_4M_1\overline{M}_3$	0
	$M_3M_1M_4\overline{M}_2$	0
	$M_2M_4\overline{M}_1\overline{M}_3$	0
	$M_2\overline{M}_4M_1\overline{M}_3$	0
	$M_4M_2\overline{M}_3M_1$	0
	$M_1M_3\overline{M}_2\overline{M}_4$	0
	$M_1\overline{M}_3M_2\overline{M}_4$	0
	$M_1M_4\overline{M}_2M_3$	0
121323	$M_1\overline{M}_2\overline{M}_4M_3$	2
	$\overline{M}_1M_4M_2\overline{M}_3$	1
	$M_2M_1M_4M_3$	1
	$M_1M_4M_2\overline{M}_3$	0
	$\overline{M}_3M_4M_2M_1$	0
	$M_1\overline{M}_2M_4M_3$	0
	$\overline{M}_1\overline{M}_2M_4M_3$	0
	$M_3\overline{M}_2M_4M_1$	0
	$M_1\overline{M}_4M_2\overline{M}_3$	0
	$\overline{M}_1\overline{M}_2M_4M_3$	0
	$M_1\overline{M}_2\overline{M}_4M_3$	0
	$M_1\overline{M}_2\overline{M}_3M_4$	0
	$M_4M_3M_2\overline{M}_1$	0
	$M_4M_3M_2M_1$	0
122133	$\overline{M}_1M_2M_4\overline{M}_3$	4

DOW	Reduced Pattern	Count
123231	$M_2M_4M_3M_1$	5
	$M_1\overline{M}_4M_3\overline{M}_2$	1
	$M_2M_4M_3\overline{M}_1$	1
	$M_2\overline{M}_4M_3M_1$	1
	$M_3M_1M_4M_2$	1
	$M_1M_4M_3\overline{M}_2$	0
	$M_1\overline{M}_3M_4\overline{M}_2$	0
	$\overline{M}_1M_4\overline{M}_3M_2$	0
	$M_2\overline{M}_3M_4M_1$	0
	$\overline{M}_2M_4M_1\overline{M}_3$	0
	$M_3M_1\overline{M}_4M_2$	0
123123	$M_1M_3M_2M_4$	35
	$M_2M_4M_1M_3$	18
	$M_3M_2M_4M_1$	5
	$M_3M_2\overline{M}_4M_1$	1
	$M_1M_4\overline{M}_2\overline{M}_3$	1
	$M_1M_3M_2\overline{M}_4$	0
	$\overline{M}_1M_3M_2\overline{M}_4$	0
	$M_2M_4\overline{M}_1M_3$	0
	$M_2\overline{M}_4\overline{M}_1M_3$	0
	$M_3M_2M_4\overline{M}_1$	0
122313	$M_2M_1M_4\overline{M}_3$	1
	$M_2\overline{M}_1M_4\overline{M}_3$	1
	$M_3\overline{M}_4M_2M_1$	1
	$M_1M_4\overline{M}_3M_2$	0
	$M_1\overline{M}_2M_4\overline{M}_3$	0
	$\overline{M}_1M_4\overline{M}_3M_2$	0
121233	$\overline{M}_1M_2M_4M_3$	6
	$\overline{M}_1M_2\overline{M}_4M_3$	0
	$M_4M_3\overline{M}_2M_1$	0
	$M_1\overline{M}_2M_3\overline{M}_4$	0
123321	$M_1M_3\overline{M}_4\overline{M}_2$	14
	$M_2M_4\overline{M}_3M_1$	1
123312	$M_1M_3\overline{M}_4M_2$	1
122331	$\overline{M}_2M_1M_4\overline{M}_3$	0

Table 2: The number of reduced patterns having four MDSs observed in *O. trifallax*, grouped by their associated double occurrence word.

S3 Nested Repeat-Return Removal Algorithm

Get DOW from the contig and store it in $RDOW$ variable
Remove all loops from $RDOW$
Put $RDOW$ in the ascending order
repeat
    $prevWord \leftarrow RDOW$
    $ListOfSets_1 \leftarrow empty\ list$, $ListOfSets_2 \leftarrow empty\ list$
    $i \leftarrow 1$
    while $i < length(RDOW)$ **do**
        if $RDOW[i+1] = RDOW[i] + 1$ **then**
            $letterSet \leftarrow \{RDOW[i]\}$
            $i \leftarrow i + 1$
            $letterSet \leftarrow letterSet \cup \{RDOW[i]\}$
            while $RDOW[i+1] = RDOW[i] + 1$ **do**
                $i \leftarrow i + 1$
                $letterSet \leftarrow letterSet \cup \{RDOW[i]\}$
            end while
            Add $letterSet$ to $ListOfSets_1$
        else if $RDOW[i+1] = RDOW[i] - 1$ **then**
            $letterSet \leftarrow \{RDOW[i]\}$
            $i \leftarrow i + 1$
            $letterSet \leftarrow letterSet \cup \{RDOW[i]\}$
            while $RDOW[i+1] = RDOW[i] - 1$ **do**
                $i \leftarrow i + 1$
                $letterSet \leftarrow letterSet \cup \{RDOW[i]\}$
            end while
            Add $letterSet$ to $ListOfSets_2$
        end if
        $i \leftarrow i + 1$
    end while
    $intersection_1 \leftarrow \emptyset$, $intersection_2 \leftarrow \emptyset$
    for all $letterSet_i, letterSet_j$ in $ListOfSets_1$ with $i \neq j$ **do**
        if $|letterSet_i \cap letterSet_j| \geq 2$ **then**
            $intersection_1 \leftarrow intersection_1 \cup (letterSet_i \cap letterSet_j)$
        end if
    end for
    for all $letterSet_i$ in $ListOfSets_1$ and $letterSet_j$ in $ListOfSets_2$ **do**
        if $|letterSet_i \cap letterSet_j| \geq 2$ **then**
            $intersection_2 \leftarrow intersection_2 \cup (letterSet_i \cap letterSet_j)$
        end if
    end for
    $lettersToRemove \leftarrow intersection_1 \cup intersection_2$
    for all letter a in $RDOW$ **do**

43: **if** a is in $lettersToRemove$ **then**
44: Remove a from $RDOW$
45: **end if**
46: **end for**
47: Remove all loops from $RDOW$
48: Put $RDOW$ in the ascending order
49: **until** $prevWord = RDOW$ or $RDOW = empty\ word$

S4 Proof that the S3 Algorithm Removes Maximal Repeat-Return Words

Let $w = a_1 a_2 \cdots a_{2n}$ be an assembly word. Note that w may be represented uniquely as $w = uv$ where u is a maximal sequence of the form

$$i(i+1)\cdots(i+j-1) \text{ or } i(i-1)\cdots(i-j+1) \qquad (\star)$$

for some $i, j \in [n]$ and v is a word. Consider the recursive construction where $w_0 = w$ and $w_{k-1} = u_k w_k$ where u_k is the maximal length sequence defined above ($\star$), and w_k is eventually an empty set. This results in a finite set $\{u_1, \ldots, u_p\}$. Define the binary operation $\wedge$ such that, given two words x and y,

$$x \wedge y = \begin{cases} \{a \mid a \in x, a \in y\}, & x \text{ and } y \text{ have at least two letters in common} \\ \emptyset, & \text{otherwise} \end{cases}$$

and the set

$$R = \bigcup_{k,l \in [p], k \neq l} (u_k \wedge u_l).$$

Claim 1. *R is the set of letters that form all maximal sub-repeats and sub-returns of w.*

Proof. Let $x \in R$, then $\exists u_i, u_j$, such that $x \in u_i \wedge u_j$, and we note that $u_i \wedge u_j$ consists of consecutive symbols that correspond to a sub-repeat/sub-return of w. Hence, x is a part of some maximal sub-repeat/sub-return of w.

Let x be a letter of some maximal sub-repeat/sub-return v of w, then there is a letter $y \in v$ such that y is next to x in v. So, $y = x + 1$, or $y = x - 1$ and WLOG assume $y = x + 1$. Consider the first occurrence of v inside w. We claim that x and y belong to the same u_k. Assume otherwise, then we have $x \in u_k$ and $y \in u_{k+1}$. Since $y = x + 1$, then u_k can not be a singleton. Also, u_k can not be of the form $i(i+1)\cdots(x-1)x$, since otherwise u_k can be extended by appending $y = (x+1)$ to u_k, but u_k is maximal. So, x can only belong to u_k of the form $(i)(i-1)\cdots(x+1)x$. Hence, $x + 1 = y \in u_k$, and we also have that $y \in u_{k+1}$, so y already appears twice in w and there is yet another occurrence of y in w. Thus, x and y belong to the same u_k during the first occurrence of v inside w. For the second occurrence of v inside w, we have that either x is before y (i.e. v is a sub-repeat), or y is before x (i.e. v is a sub-return). If x is before y, then the similar argument shows that x and y belong to the same u_l. Assume that y is before x and $y \in u_l$, $x \in u_{l+1}$. Then, since $y = x + 1$, we have that u_l can not be a singleton. Also, u_l can not be of the form $i(i-1)\cdots(y+1)y$, since otherwise u_l can be extended by appending $x = y - 1$ to u_l, but u_l is maximal. Thus, u_l can only be of the form $i(i+1)\cdots(y-1)y$. Hence, $y-1 = x \in u_l$, and $x \in u_{l+1}$, so x appears twice in w as a letter of u_l and u_{l+1}. Also, $x \in u_k$ (which is disjoint from u_l), so x appears 3 times in w. Therefore, x and y belong to the same u_l, so $x \in u_k \wedge u_l \neq \emptyset$, since there are at least two letters (x and y) in u_k and $u_l \Rightarrow x \in R$. □