Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 00 (2020) 1–14
Journal of Theoretical Biology

The Complexity of Genome Rearrangement Combinatorics under the Infinite Sites Model

Chris D Greenman[a], Luca Penso-Dolfin[a], Taoyang Wu[a]

[a] *School of Computing Sciences, University of East Anglia, Norwich, UK, NR4 7TJ*

Abstract

Rearrangements are discrete processes whereby discrete segments of DNA are deleted, replicated and inserted into novel positions. A sequence of such configurations, termed a rearrangement evolution, results in jumbled DNA arrangements, frequently observed in cancer genomes. We introduce a method that allows us to precisely count these different evolutions for a range of processes including breakage-fusion-bridge-cycles, tandem-duplications, inverted-duplications, reversals, transpositions and deletions, showing that the space of rearrangement evolution is super-exponential in size. These counts assume the infinite sites model of unique breakpoint usage.

Keywords: Rearrangements, Complexity, Combinatorics, Cancer

1. Introduction

Rearrangements are events where a piece of DNA from a genome is moved to a different location, deleted from the cell, duplicated, replaced in reversed orientation or a combination thereof. These arise through a range of different mechanisms, resulting in complicated configurations across many biological entities, for example salmonella [1], grass [2] and cancer [3] all exhibit rearrangements.

The use of rearrangement algorithms to interpret such configurations has a rich history. Notably, the evolution of reversals have been thoroughly studied [4, 5, 6, 7], along with tandem duplications [8, 9, 10, 11, 12, 13], breakage-fusion-bridge processes have been studied more recently [14, 15, 16], and more general processes such as double-cut-and-join and combinations thereof have also been analyzed both graph theoretically [17, 18, 19] and with group theoretic techniques [20, 21]. A general overview of these and other problems found in rearrangement algorithms, as well as a plethora of other references can be found in [22].

There is one feature we focus on for which published results differ significantly. This relates to the infinite sites model [23], which assumes that there are an infinite number of sites on DNA that can be mutated, and that the chance that the same site is mutated twice is effectively zero. In the context of rearrangements, a site is a DNA position that is implicated in a rearrangement, and is often referred to as a *breakpoint*. These can be single positions (in breakage-fusion-bridge events, for example, as detailed below), pairs of sites (in tandem duplications), or indeed a multitude of positions (chromothripsis [24]). For most rearrangement studies, the infinite sites model is not assumed and breakpoints are taken to be reusable. For example, of the references above, only [13, 16, 18, 19] take an infinite sites approach, naturally raising the issue of which methodology is appropriate.

Rearrangement methodology was originally constructed to account for different orders of genes between pairs of genomes. Breakpoints were thus implicitly interpreted as gaps between genes, meaning reuse of breakpoints is perfectly reasonable. An examination of breakpoint reuse first appeared when the assumption that rearrangements occur uniformly across the genome was considered [25], revealing the presence of rearrangement hotspots, or breakpoint regions. Such regions can be the result of highly mutable DNA, such as fragile sites, or regions undergoing positive selection, such as deleted tumour suppressor genes in cancer [26]. Furthermore, these regions are known to rearrange at different rates [27]. Conversely, the regions between such hotspots can exhibit a lack of breakpoints. This can be the result of selection against mutation. Highly conserved synteny blocks are known to be unaltered in bacterial rearrangement processes [28], for example. In cancer the contrast is less extreme; although some regions certainly contain a higher density of breakpoints, rearrangements are found all over the genome.

More recent sequencing technologies allow breakpoints to be examined in very high resolution, meaning breakpoints and their reuse can even be examined down to the level of base nucleotides in some cases. This can be important for the reconstruction of rearrangement evolution; missing information, even from small DNA blocks, can result in erroneous predictions [29]. In cancer genomes, even within fragile sites [26], breakpoints occur at unique positions meaning breakpoint reuse is somewhat unlikely and an infinite sites model is more appropriate. The decision of whether to use an infinite sites model or not then comes down to resolution, that is, whether breakpoints or breakpoint regions are being considered. In the present work, we adopt an infinite sites approach, and building on [13, 16, 18], assume breakpoints occur at unique DNA positions.

The classical problem considered in most rearrangement studies is to estimate the sequence of rearrangements that connect two distinct, known genomes. For example, in cancer genomics, one is comparing a (healthy) reference genome to a rearranged cancer genome. Modern sequencing methods mean that high-resolution rearrangement information is available, although this is generally incomplete; the number of copies of DNA segments (the rearranged blocks of DNA) and pairwise segment adjacency information may be available, but the entire sequence of segments that constitute a full genome is unknown [18]. An exploration of the space of possible configurations that fit the data then becomes desirable, which motivated the following combinatorial problem; specifically, we instead fix one (reference) genome and are interested in the number of different DNA configurations that can subsequently arise from a given sequence of rearrangements. The rearrangements involved can be taken from a range of rearrangement classes.

For sequences of rearrangements taken from a single class, the number of configurations has previously been found exactly in some cases. For breakage-fusion-bridge-cycles, the number is the super-exponential power $2^{\frac{1}{2}n(n-1)}$, where n counts the number of rearrangements [16], and for tandem-duplications, the number is $\prod_{k=1}^{n}(4^k - (2k+1))$ [13]. These counts were achieved by similar graph theoretic constructs, suggesting the methods or results may be applied across a wider class of rearrangement classes. The present work offers a simpler approach, utilizing geometric arguments and algebraic properties of certain sequences of words. Whilst this technique does not describe the rearrangement process in as much detail as [13, 16], it is sufficient to navigate the rearrangement spaces involved and demonstrate super-exponential size. Furthermore, this is achieved for different combinations of rearrangement classes.

In the next section we review the different admissible rearrangement mechanisms and provide an algebraic representation. The third section describes the counting technique. The fourth section provides technical results needed to verify the counts are correct, and a conclusions section completes the work.

2. Rearrangement Mechanics

There are several different types of rearrangement that can occur in DNA evolution, and many of them can occur in combination in a single cancer genome [18]. These include tandem-duplications, reversals, breakage-fusion-bridge-cycles, inverted-duplications and deletions, which are the classes that we consider. Whilst these methods are likely extendable to inter-chromosomal operations such as translocations and double-cut-and-join operations, we stick to intra-chromosomal events for the present study. We note that there are other more complex rearrangements such as chromothripsis that may also be going on in genomes [24], along with unknown processes, which we do not consider.

2.1. Rearrangement Classes

We now describe the rearrangement classes in more detail, utilizing the examples given in Fig. 1.

Breakage-fusion-bridge cycles (BFB) arise when two pieces of duplicated DNA are erroneously joined together resulting in palindromic DNA. A piece of DNA, initially represented as a sequence of two *segments*, 12, loses one of the segments, 2. The remaining segment, 1, is duplicated and the two pieces erroneously attached to each other at the same end, resulting in configuration $1\overline{1}$, where the overline $\overline{1}$ indicates the second copy of 1 is in reversed direction compared to the first. Collectively, this is represented algebraically as $12 \to 1\overline{1}$. Note that if the duplication occurs from the other end with segment 1 being lost, we get $12 \to \overline{2}2$. In both cases the resultant product is palindromic and any subsequent BFBs can be implemented equivalently from either end of the product. A more detailed description of this behaviour can be found in [14, 15, 16].

Tandem-duplications (TD) occur when a piece of DNA is copied and placed adjacent to the first copy. Algebraically, we can consider this as an operation $123 \to 1223$, where the middle segment, labelled 2, has been copied. These processes have been implicated in many duplication processes [30], including cancer [31], and the combinatorics have been well characterized [8, 9, 10, 11, 12].

Deletions (DEL) arise when segments of DNA are excised from a genome. Algebraically these can be represented as $123 \to 13$, where segment 2 has been deleted. Such processes arise naturally as part of genomic evolution, inactivating tumour suppressor genes, for example [26].

Reversals (REV) occur when a piece of DNA is excised and reinserted in the backwards orientation. Algebraically, this can be represented as $123 \to 1\overline{2}3$, where the segment labelled 2 is inverted. These have been implicated in various situations [32]. The combinatorics of these processes are also well characterized [4, 5, 6, 7], as well as for the more general double cut and join process [17, 20]. The term reversal is usually adopted by algebraic descriptions of the process. However, this can also be referred to with the more biological term inversion [18]. Note that the term reversal applies more generally to permutations (such as $12345 \to 14325$, for example) that reverse the order of (but do not invert, or track the orientation of, individual) segments, whereas the term inversion explicitly requires the DNA to be placed in inverted orientation.

Transpositions (TRA) arise when two pieces of DNA are interchanged, or equivalently when one piece is excised and inserted into a different position in the chromosome [22, 33, 34]. We can represent this algebraically as $1234 \to 1324$, where the segment 3 is excised and implanted between segments 1 and 2. This process can be referred to as an insertion, a more general biological term in which the original segment may be excised

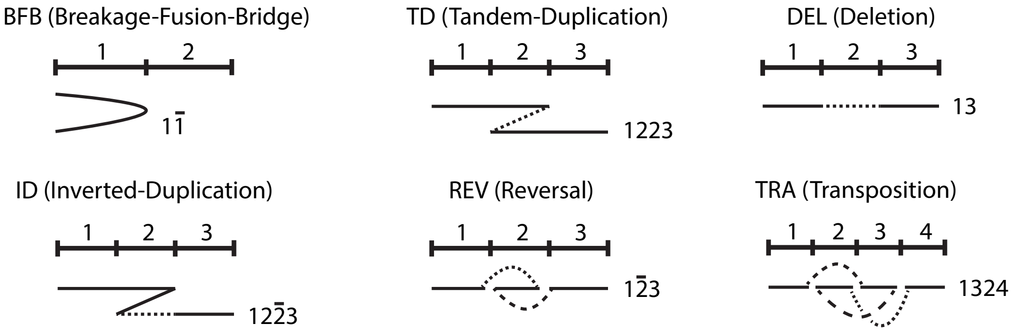


Figure 1. The rearrangement classes analysed, including breakage-fusion-bridge, tandem-duplication, deletion, inverted-duplication, reversal and transposition. The horizontal tick-marked bar indicates the original reference DNA configuration, each number denoting a segment of DNA (123 for inverted-duplication reference, for example). The configurations below each horizontal bar are the rearranged DNA, lined up against the reference. Solid lines are segments, dotted and dashed lines denote connections. Final algebraic configurations are given word representations under the rearrangement names ($12\overline{2}3$ for inverted-duplication, for example, where an overline indicates a segment in reversed configuration).

within-chromosome as above (preserving the number of segment copies) [18], or the inserted segment may be copied or excised from an external source (from other chromosomes or cells, for example) [35, 36]. Note, however, that the term transposition is strictly reserved for intra-chromosomal insertions.

2.2. *Rearrangement Evolution*

All these processes can be implicated in genomic evolution, where checkpoint machinery fails and aberrant DNA structures are allowed to propagate, cancer genomes exhibiting perhaps the most prominent combinations of these processes, resulting in very scrambled genomes [18]. We consider a combinatorial problem arising from such a mixed process, where we are required to count the number of different discrete structures that can arise from such processes (the discrete nature of the structure is made more precise below).

Consider the rearrangements in Fig. 2, for example, where we can see a combination of three such operations, a tandem-duplication (TD), a breakage-fusion-bridge (BFB), followed by a deletion (DEL). This choice of *rearrangement sequence* $TD \to BFB \to DEL$ is now fixed and we are interested in the number of possible 'configurations' that may arise, where we need to say what we mean by configurations.

Now, any instance of this rearrangement sequence will implicate five breakpoints, delineating the six segments that are initially contiguous, represented as 123456. For this example, after the first event, a TD, the region 234 is duplicated, resulting in sequence 1234.23456. We represent the structure in Fig. 2 by aligning it against the original sequence 123456, lining up the same segments vertically. Novel *somatic* connections between initially disconnected segments are indicated by a dashed line in the structures, and by a decimal point algebraically. Note that the start point of the duplicated region 234 implicates the position between segments 1 and 2, represented as $[12]_1$, the subscript indicating it is the first copy of this breakpoint in the word. Similarly, the duplication termination point between segments 4 and 5 is represented as $[45]_1$. Such positions implicated in rearrangement are examples of the aforementioned *breakpoints*. The rearrangement operation is then represented algebraically with the term $TD([12]_1, [45]_1)$. The next rearrangement, a BFB, duplicates the subsequence 1234.2345, joining both copies at the single breakpoint between segments 5 and 6, resulting in the word 1234.2345.$\overline{5432}.\overline{4321}$, segment 6 being lost. This is represented as $BFB([56]_1)$. The last operation, $DEL([23]_3, [34]_4)$, implicates the third and fourth copy of the remaining two reference positions, those at each end of segment 3, deleting the subsequence $\overline{24}$, resulting in the final sequence 1234.2345.$\overline{543}.\overline{321}$.

Now, the last operation $DEL([23]_3, [34]_4)$ is acting on distinct positions [23] and [34]. It would seem that (following genomic duplication by the BFB) operations utilising distinct copies of the same breakpoint, such as $DEL([23]_2, [23]_3)$, may be possible. However, this would require both ends of the deletion to occur at precisely the same genomic sequence, something that is excluded by the infinite sites assumption, and arguably unlikely to occur in practice, for the rearrangements presently considered at least. There may plausibly be other rearrangement processes (based on sequence homology) that implicate the same genomic sequence twice, but such processes are not considered further, and would require extensions to the methods detailed below.

We thus find that all five breakpoint positions [12], [23], $\ldots$, [56] that demarcate the six original segments $1, 2, \ldots, 6$ are implicated. In general each breakpoint $[mn]_p$ represents the p^{th} copy of the reference position between segments m and n in a word. Because breakpoints are only implicated once, we find that implicated breakpoints are always in their original order with $n = m + 1$.

We can represent this chain of events as an algebraic 'configuration' termed an *evolution*:

$$\begin{aligned} 123456 \to 1234.23456 \to & 1234.2345.\overline{5432}.\overline{4321} \\ & \to 1234.2345.\overline{543}.\overline{321} \end{aligned}$$

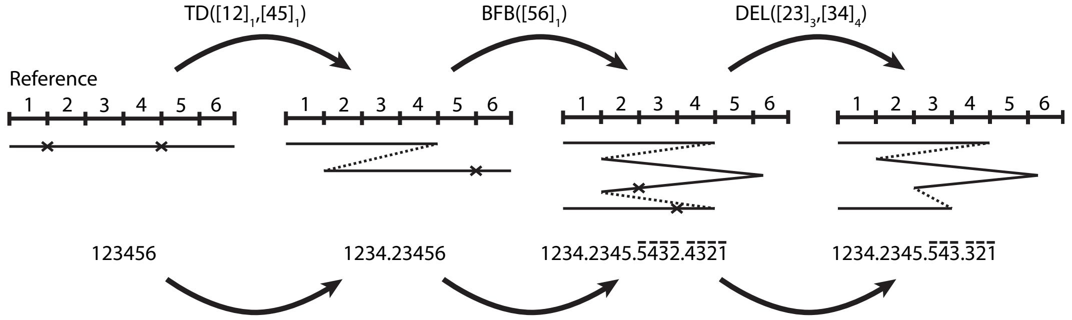

TD([12]1,[45]1)
BFB([56]1)
DEL([23]3,[34]4)
Reference
1 2 3 4 5 6
123456
1234.23456
1234.2345.$\overline{5}\overline{4}\overline{3}\overline{2}$.$\overline{4}\overline{3}\overline{2}\overline{1}$
1234.2345.$\overline{5}\overline{4}\overline{3}$.$\overline{3}\overline{2}\overline{1}$

Figure 2. One possible instance of the *rearrangement sequence* involving a tandem duplication (TD), a breakage-fusion-bridge-cycle (BFB), and lastly a deletion (DEL). The corresponding *evolution* of words based upon a *reference* of six initially contiguous DNA segments numbered 1, 2, ... , 6 is given. Breakpoints acted upon are labelled. For example, the deletion acting on 1234.2345.$\overline{5432}$.$\overline{4321}$ starts from $[23]_3$; the third copy (represented by subscript) of the breakpoint between segments labelled 2 and 3. The resulting structures are aligned relative to the reference; solid lines indicate DNA segments and dashed lines indicate connections. The crosses indicate the breakpoints implicated by the next rearrangement. The sequences of numbers are constructed by walking through the structure from the top left end and reading off the segments. The decimal points indicate *somatic* connections between segments not adjacent in the reference. An overline indicates the segment is in reversed orientation.

This is a sequence of words based upon the reference sequence. However, note from Fig. 2 and the construction described above that if we have (i) the initial word, (ii) the rearrangement sequence, and (iii) the breakpoints implicated in each rearrangement, then the evolution can be reconstituted.

Evolutions are considered distinct if they differ at any subsequence in the chain. For example, in Fig. 3 we see the eleven possible evolutions arising from a sequence of two IDs. The first and last evolution have the same initial and final words (12345 and 12.$\overline{2}$.34.$\overline{4}$.5) but have distinct intermediate structures (with words 1234.$\overline{4}$.5 and 12.$\overline{2}$.345), so are considered distinct. Note that this observation is related to the classical rearrangement problem of connecting two fixed genomes by a path of rearrangements; the existence of distinct paths indicates multiple solutions exist for the corresponding classical problem.

Our primary combinatorial question then is to count the number of distinct evolutions corresponding to a given rearrangement sequence. Note that we are only interested in the different discrete, algebraic structures that arise from such processes; the exact genomic positions of the breakpoints are not of concern.

Note that we have assumed each subsequent rearrangement implicates new breakpoints. Once a rearrangement has occurred, any exposed breakpoints are repaired and the chance that the next rearrangement will occur at the same position is unlikely, and the infinite sites model is adopted.

Now for a chain of n rearrangements of the same type, asking how many different evolutions are possible is a problem we have considered elsewhere, where we have shown that there are $2^{\frac{1}{2}n(n-1)}$ possible evolutions from breakage-fusion-bridge-processes [16] and $\prod_{k=1}^{n}(4^k-(2k+1))$ are possible from tandem-duplications [13]. The proofs rely on an induction that instead of asking what happens if we perform an n^{th} rearrangement after $n-1$ previous events (at the end of the evolution), considers what happens to the structure if a new rearrangement is placed at the start of the evolution. We adapt this approach, obtaining the following.

Theorem 1. *The number of distinct rearrangement evolutions corresponding to the rearrangement sequence* $R_1 \to R_2 \to R_3 \to \cdots \to R_n$ *is given by* $\prod_{k=1}^{n}\Phi_{R_k}(\beta_k)$ *where* R_k *denotes the class of the* k^{th} *rearrangement in the sequence (selected from Table 1),* $\beta_k = \sum_{j>k}\beta(R_j)$ *denotes the number of breakpoints subsequently implicated by later rearrangements, and* $\beta(R_j)$ *counts breakpoints implicated by rearrangement* R_j*. The final number of segments in the reference is one greater than the total number of breakpoints,* $\beta_0 + 1$.

This can then be used to count rearrangement evolutions. For example, consider counting the number of possible evolutions of the form $TD \to BFB \to DEL$, one example of which was given in Fig. 2. We start with the final rearrangement, the DEL, which has no following rearrangements, and so subsequent breakpoints. Then using $\beta_3 = 0$ for DELs in Table 1, we find the number of ways of doing this is $\Phi_{DEL}(\beta_3) = 0 + 1 = 1$. Note that the DEL introduces two breakpoints. We then introduce a new first rearrangement, the BFB, for which there are $\Phi_{BFB}(\beta_2) = 2^2$ configurations (using $\beta_2 = \beta(DEL) = 2$ in Table 1). This give four possible evolutions, all with $\beta_1 = \beta(BFB) + \beta(DEL) = 3$ breakpoints. We finally introduce the first event, the TD, for which there are $\Phi_{TD}(\beta_1) = 2^{3+2} - (3 + 3) = 26$ choices, giving $26\times4\times1 = 104$ possible evolutions in total. For all such evolutions we have $\beta_0 = \beta(TD)+\beta(BFB)+\beta(DEL) = 5$ breakpoints in total, resulting in the $\beta_0 + 1 = 6$ reference segments given in Fig. 2.

We see from the third column in Table 1 that the rearrangement classes duplicating DNA (BFB, TD and ID) result in larger evolution spaces. Specifically we will have the following.

Corollary 1. *Let* $N(n)$ *denote the number of evolutions for a particular sequence* $R_1 \to R_2 \to \cdots \to R_n$ *of rearrangements.*

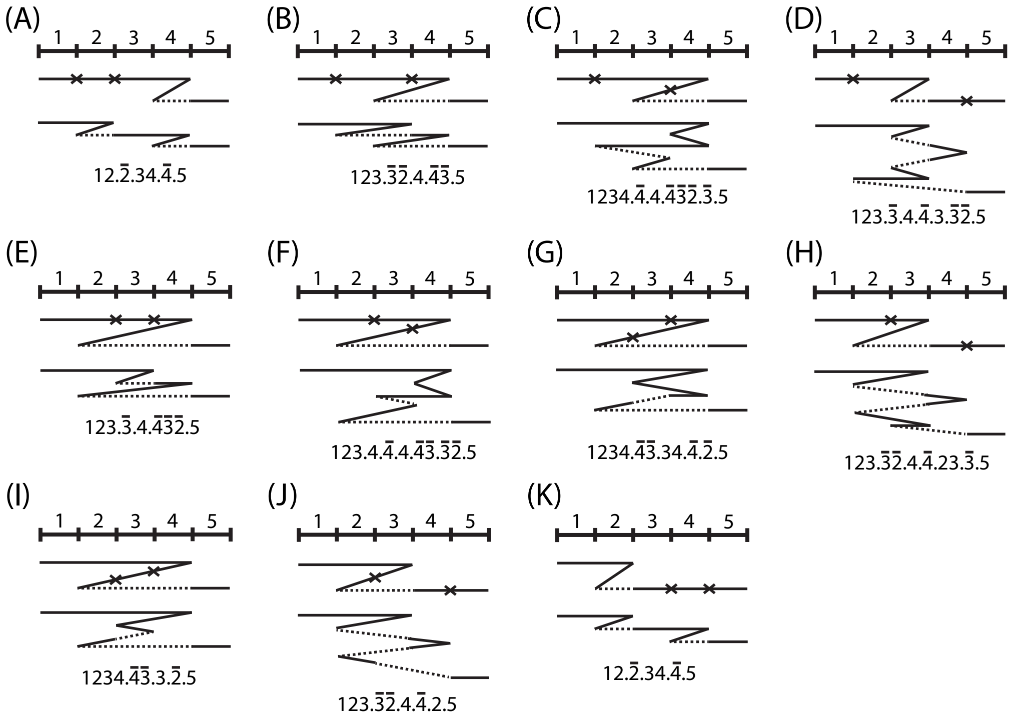

(A)
1 2 3 4 5
$12.\bar{2}.34.\bar{4}.5$
(B)
1 2 3 4 5
$123.\bar{3}\bar{2}.4.\bar{4}\bar{3}.5$
(C)
1 2 3 4 5
$1234.\bar{4}.4.\bar{4}\bar{3}\bar{2}.\bar{3}.5$
(D)
1 2 3 4 5
$123.\bar{3}.4.\bar{4}.3.\bar{3}\bar{2}.5$
(E)
1 2 3 4 5
$123.\bar{3}.4.\bar{4}\bar{3}\bar{2}.5$
(F)
1 2 3 4 5
$123.4.\bar{4}.4.\bar{4}\bar{3}.\bar{3}\bar{2}.5$
(G)
1 2 3 4 5
$1234.\bar{4}\bar{3}.34.\bar{4}.\bar{2}.5$
(H)
1 2 3 4 5
$123.\bar{3}\bar{2}.4.\bar{4}.23.\bar{3}.5$
(I)
1 2 3 4 5
$1234.\bar{4}\bar{3}.3.\bar{2}.5$
(J)
1 2 3 4 5
$123.\bar{3}\bar{2}.4.\bar{4}.2.5$
(K)
1 2 3 4 5
$12.\bar{2}.34.\bar{4}.5$

Figure 3. Eleven possible configurations (A-K) arising from two inverted-duplications (IDs). Each configuration contains three structures and a sequence. The first (upper) structure is the reference, labelled $1, 2, \ldots, 5$. The second indicates the structure after the first ID, the crosses indicated the two breakpoints implicated in the second ID. The third structure indicates the final structure, along with the numerical sequence of associated reference segments (overlined segments are inverted and decimal points denote somatic connections).

Then for duplication sequences (i.e. $R_i \in \{BFB,\ TD,\ ID\}$ *for* $i = 1, 2, \ldots, n$*) we have* $\log N(n) = O(n^2)$*. For sequences without duplication, we find* $\log N(n) = O(n \log n)$*.*

We now develop the machinery to establish these results.

3. Updating the Reference

First we describe combinatorics associated with updating the reference. Instead of taking an evolution and asking how many new evolutions are possible when another rearrangement is applied (at the end of the evolutionary process), we instead suppose the reference is a product of a rearrangement (i.e. we introduce a new rearrangement at the beginning of the process) and count the number of consistent possibilities. We consider this for the range of rearrangement types in Table 1.

3.1. Breakage-Fusion-Bridge-Cycles

Consider a stretch of DNA implicated in a sequence of various rearrangements. For example, consider the case that initially we have a contiguous sequence of five segments, the reference $ABCDE$, demarcated by four breakpoints that will be implicated by subsequent rearrangements (two TDs or four BFBs would implicate four breakpoints, for example). We now assume the reference is not in fact the original genome, but the result of a BFB which will implicate an additional breakpoint, five in total. Then the updated reference can be represented as a sequence of six segments 123456 separated by five breakpoints. We can then ask how the segments $A, B, \ldots, E$ line up against the updated reference of six segments. One possibility is the BFB *mapping structure* given in Fig. 4, where as we walk from point I to II to III we see the contiguous segments A to E, with the upper part I-II containing three of the original four breakpoints and the lower part II-III the remaining one. The newly introduced fifth breakpoint is the fold at position II.

Note that as we walk through the updated structure, the original order of the segments A through to E must be preserved, because we are not changing the configuration $ABCDE$, rather just assuming it is the product of an earlier event. This means that the order of the four original breakpoints is also preserved along the structure, a feature that applies more generally than this example.

This change of reference can be represented as a *reference mapping* M from the original reference to the updated one. For the BFB example in Fig. 4, we see that $A \to 1$ covers one segment from the new reference, whereas $D \to 5\overline{543}$ covers four, one of them twice. The full mapping is given to the right of the mapping structure in Fig. 4. Formally, we can represent the mapping of segments, such as $M(A) = 1$, $M(D) = 5\overline{543}$, for example.

When lined up against the new reference, the lower breakpoint on II-III is positioned between the first two breakpoints on

Rearrangement	Breakpoints	Counting	Number of Evolutions				
Class	Implicated	Factor					
R	$\beta(R)$	$\Phi_R(k)$	$N(n) = \prod_{k=1}^{n} \Phi_R(n_k)$	$N(1)$	$N(2)$	$N(3)$	$N(4)$
BFB	1	2^k	$2^{\frac{1}{2}n(n-1)}$	1	2	8	64
TD	2	$2^{k+2} - (k+3)$	$\prod_{k=1}^{n}(4^k - (2k+1))$	1	11	627	156123
ID	2	$2^{k+2} - (k+3)$	$\prod_{k=1}^{n}(4^k - (2k+1))$	1	11	627	156123
REV	2	$\binom{k+2}{2}$	$\frac{(2n)!}{2^n}$	1	6	90	2520
DEL	2	$k+1$	$(2n-1)!!$	1	3	15	105
TRA	3	$\binom{k+3}{3}$	$\frac{(3n)!}{6^n}$	1	20	1680	369600

Table 1. Size of evolution spaces for six rearrangement classes. The first column is the class of rearrangement (see Fig. 1). The second column contains the number of breakpoints utilized by the rearrangement. The third column is the combinatorial factor $\Phi_R(k)$ contributed by rearrangement R in a sequence of rearrangements, where k denotes the number of breakpoints subsequently formed in the genome after rearrangement R. The fourth column is the size of the rearrangement space if only a single class of rearrangement is in operation. The next few columns highlight the growth of this space with the number of rearrangements.

I-II. However, the lower breakpoint could have been positioned in four different locations relative to the three breakpoints on the arm I-II. These would all result in distinct reference mappings. This choice is counted by the factor $\binom{4}{3}$.

Now, more generally we have k breakpoints to realign rather than four. Along I-II we can position r of them, the remainder $k - r$ along II-III. For any given r, the number of different ways of interleaving the upper and lower breakpoints against the updated reference (whilst preserving their order along the structure) will be $\binom{k}{r}$. If we sum this over the possible values of $r \in \{0, 1, \ldots, k\}$ we find 2^k possibilities. This is the counting factor given in the third column of Table 1.

Furthermore, if we now consider a rearrangement sequence consisting entirely of n BFBs, then each one will introduce a single breakpoint and a factor of the form 2^k where k counts the number of subsequent breakpoints in the process. This suggests the total number of possible structures is given by $\prod_{k=0}^{n-1} 2^k = 2^{\frac{1}{2}n(n-1)}$, giving the total number of configurations found from a sequence of BFBs, as given in the fourth column of Table 1, a result derived by other means in [16]. Note furthermore that the exponent of this count is of order $O(n^2)$, exhibiting super-exponential growth of the space of rearrangements with n.

As remarked earlier, a BFB can duplicate DNA either from the left or right, resulting in two possible palindromic chromosomes. This symmetry means that any BFB events immediately following the first BFB can act from either end with the same result. Thus when using Theorem 1 to count the number of evolutions for a sequence of rearrangements, we must specify the direction of the first BFB for any chain of consecutive BFBs in the sequence (or double the count for each chain of BFBs).

By construction, we have counted the number of different reference mappings we get for the different relative positions of the breakpoints on the mapping structure. This is repeated in the following subsections for other rearrangement classes. We still need to verify that distinct reference mappings result in distinct evolutions and the factors can be combined in the manner of Theorem 1. This is deferred until Section 4.

3.2. *Tandem-Duplications*

We next assume that a reference genome with k breakpoints is actually the product of a TD. We thus need to count the number of different ways of lining up the k breakpoints along a tandem duplication structure such that their original order is preserved. An example of a suitable mapping structure is given in Fig. 4. In general, we can place r breakpoints along the top portion (labelled I to II in Fig. 4) and $k - r$ breakpoints along the bottom portion (labelled III to IV). Including the two breakpoints for the TD (at positions II and III) we now have $r + 1$ along the top portion and $k - r + 1$ along the bottom, $k + 2$ in total. Now these two sets of breakpoints can have any relative order except one; the upper TD breakpoint (II) cannot be to the left of the lower TD breakpoint (III). Thus we find $\binom{k+2}{r+1} - 1$ orders are possible. Then summing over the possible values of r we find $\sum_{r=0}^{k}\left(\binom{k+2}{r+1} - 1\right) = 2^{k+2} - (k+3)$ possible orders.

If we consider a sequence of n TDs, each introduced a pair of breakpoints, then combining these factors together results in the factor $\prod_{k=1}^{n}(4^k - (2k+1))$ given in the fourth column of Table 1. This is precisely the result found by other means in [13]. Note that like BFBs, the size of this super-exponential space also has exponent $O(n^2)$ for n TDs.

3.3. *Inverted-Duplications*

IDs produce different structures to TDs, although the final combination count proves to be the same. We again position k breakpoints along a structure in their original order and count the number of ways they align relative to the new reference. From the mapping structure in Fig. 4 we see if the reference is actually the product of an ID there are three regions in which we can place these breakpoints that are later implicated, the stretches I-II, II-III and IV-V. Now we can place r breakpoints along IV-V. These are to the right of any breakpoints placed

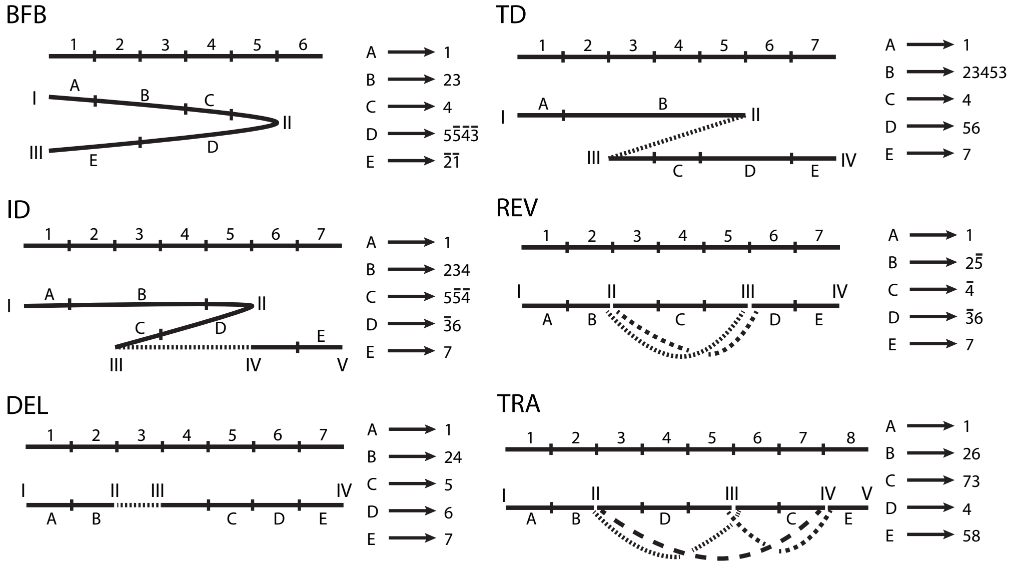


Figure 4. Redefining the reference with mapping structures. A reference sequence $ABCDE$ of five segments is updated to be the product of an earlier rearrangement for six classes of event. The original structure is then aligned against the new reference $123\ldots n$ where n is the number of segments in the updated reference. Solid lines indicate DNA and dotted lines indicate where segments are joined together. Mappings from the original reference to the updated reference are also provided.

along I-II or II-III and only one choice is possible as their order along IV-V is fixed. If we have s breakpoints placed along II-III then the remaining $k-r-s$ are placed along I-II. Now there are $\binom{k+1-r}{s}$ ways to interleave the breakpoints along II-III (including the new breakpoint at III) amongst those on I-II (relative to the new reference). We then sum this over the possibilities of $s \in \{1, 2, \ldots, k+1-r\}$ and $r \in \{0, 1, \ldots, k+1\}$, which results in the same factor $2^{k+2} - (k+3)$ seen for TDs. Note that the space of n IDs therefore also has growth exponent $O(n^2)$.

3.4. Reversals

For this case we suppose that a reference containing k breakpoints in subsequent rearrangements is actually the product of a reversal. Then the different possible evolutions correspond to the different choices of positioning the k breakpoints along the three regions I-II, II-III and III-IV of the REV mapping structure in Fig. 4, where II-III is the inverted region, and the original order of the breakpoints running through the updated structure is the same. If we place x of them in I-II and y of them in II-III, so $k-x-y$ in III-IV, the number of choices is $\sum_{x=0}^{k} \sum_{y=0}^{k-x} 1 = \binom{k+2}{2}$ giving the expression in Table 1.

If we have a sequence of n reversals (each producing two breakpoints), the number of evolutions is a product of such factors resulting in the total count $\frac{(2n)!}{2^n}$. If Stirling's formula is used, one can see that the exponent grows as $O(n \log n)$ rather than $O(n^2)$. This reduction in the growth of the corresponding evolution space is likely due to the fact that unlike reversals, BFB, TD and ID processes are duplicating DNA and provide more combinatorial opportunities for different evolutions.

3.5. Deletions

If a reference (containing k breakpoints) is this time the product of a deletion, then the deleted region cannot contain any of these breakpoints (otherwise they could not be present in the reference after the deletion). As the order of these breakpoints is preserved along the corresponding mapping structure, the only choice is how many of them are left of the deleted region. The number of choices is thus simply $k+1$.

For a sequence of n deletions implicating n breakpoints in total, the number of evolutions is thus $(2n-1)!!$. This process also has growth exponent of order $O(n \log n)$ and like reversal is a process that does not duplicate DNA.

We note that we can also consider arm loss as an additional rearrangement process, where an entire end of the chromosome is lost (rather than an internal segment). In this case there is only one way to position the breakpoints (on the undeleted region) and the factor is simply unity. For any sequence of rearrangements including arm loss we can thus just ignore such events when calculating combinatorics.

3.6. Transpositions

The final rearrangement class considered is transposition. If a reference sequence containing k breakpoints is actually the product of a transposition, then we have three regions to position the breakpoints, the regions I-II, II-III, III-IV and IV-V portrayed in Fig. 4. The number of choices is counted by $\binom{k}{3}$.

For a sequence of n transpositions, this results in $\frac{(3n)!}{6^n}$ evolutions. This is a process without DNA duplication, which again has growth order exponent of order $O(n \log n)$.

4. Evolution Uniqueness

So far, we have taken a rearrangement sequence $R_1 \to R_2 \to \cdots \to R_n$ and an associated evolution S of the form $W_1 \to W_2 \to \cdots \to W_{n+1}$ containing words W_i based upon an *original reference* sequence such as $ABCDE$ in Fig. 4. We have then introduced a new first rearrangement, resulting in updated rearrangement sequence $R_0 \rightarrow R_1 \rightarrow \cdots \rightarrow R_n$ along with *updated reference* $12\ldots m$. Section 3 counted the number of different ways of mapping the original reference $ABCDE$ against the updated reference $12\ldots m$, taking the geometry of the new first rearrangement R_0 into account. However, for our primary combinatorial problem, we also need to consider how the evolution S is affected, including whether we get a well defined updated evolution $E = M(S)$ of the form $V_0 \rightarrow V_1 \rightarrow \cdots \rightarrow V_{n+1}$ with words V_i based upon the updated reference.

Recall that a pair of evolutions are distinct if they contain sequences of words such that the i^{th} word in each sequence differs at some segment for at least one occurrence i. To demonstrate the validity of Theorem 1 will require three features:

Property A: Any mapping M can be applied to an original evolution S to give a well defined updated evolution $E = M(S)$.

Property B: Applying two distinct reference mappings $M_i \neq M_j$ to a single original evolution S results in two distinct updated evolutions. That is, $M_i(S) \neq M_j(S)$ for $i \neq j$.

Property C: Applying reference mappings to two distinct original evolutions $S_1 \neq S_2$ results in two distinct updated evolutions. The pair of mappings utilized can be distinct or identical. That is; $M_i(S_1) \neq M_j(S_2)$ for $S_1 \neq S_2$.

Note that together these properties tell us that if $M_i(S_1) = M_j(S_2)$ we must have $i = j$ and $S_1 = S_2$. The upshot is that all mappings produce unique updated evolutions across all existing original evolutions. This will later be seen as sufficient to establish Theorem 1 (see Section 4.4).

Examples of the situation considered are given in Fig. 5(B,C), where we have two original evolutions S_1 and S_2 (using original reference $ABCD$) corresponding to the rearrangement sequence $BFB \rightarrow DEL$, both resulting in three breakpoints. The introduction of a new first TD rearrangement adds two extra breakpoints, giving five in total, so we have segments $12\ldots 6$ in an updated reference. From Section 3.2, there are $\Phi_{TD}(\beta_1) = 2^{3+2} - (3+3) = 26$ possible reference mappings, where $\beta_1 = 3$ counts the number of breakpoints formed after the TD. In Fig. 5(A) we have three examples from these 26 possibilities, labelled M_1, M_2 and M_3. Applying these three mappings to the two original evolutions results in six updated evolutions E_1-E_6 in terms of new reference 123456 (the example modifying S_1 with M_1 gives the evolution portrayed in Fig. 2). All six cases give a well defined updated evolution in accordance with Property A. Furthermore, if we take a single evolution, say S_1 and a pair of distinct mappings, say M_1 and M_2, the updated evolutions are distinct (note the updated evolutions $E_1 = M_1(S_1)$ and $E_2 = M_2(S_1)$ differ at the final fourth word). This is an example of Property B, where a single evolution produces distinct evolutions when two distinct mappings are applied. Note also that if we take two distinct evolutions (S_1, S_2) and any two mappings (distinct or otherwise), the updated evolutions differ (any evolution from E_1-E_3 is distinct from any evolution from E_4-E_6 in at least one word, that is, $M_i(S_1) \neq M_j(S_2)$ for all i, j). This is the third Property C we need to explain; distinctness of evolutions is preserved under the application of (distinct or identical) mappings.

In order to explain these three properties, we need to consider certain characteristics of the mappings.

4.1. Reference Update Mapping Characteristics

There are four characteristics we require from the mappings. The first is concerned with the mapping of breakpoints, the second with the uniqueness of breakpoint usage, the third with contiguity of segments between implicated breakpoints, and the fourth relates to the chronology of rearrangement events.

4.1.1. Mapping Breakpoints

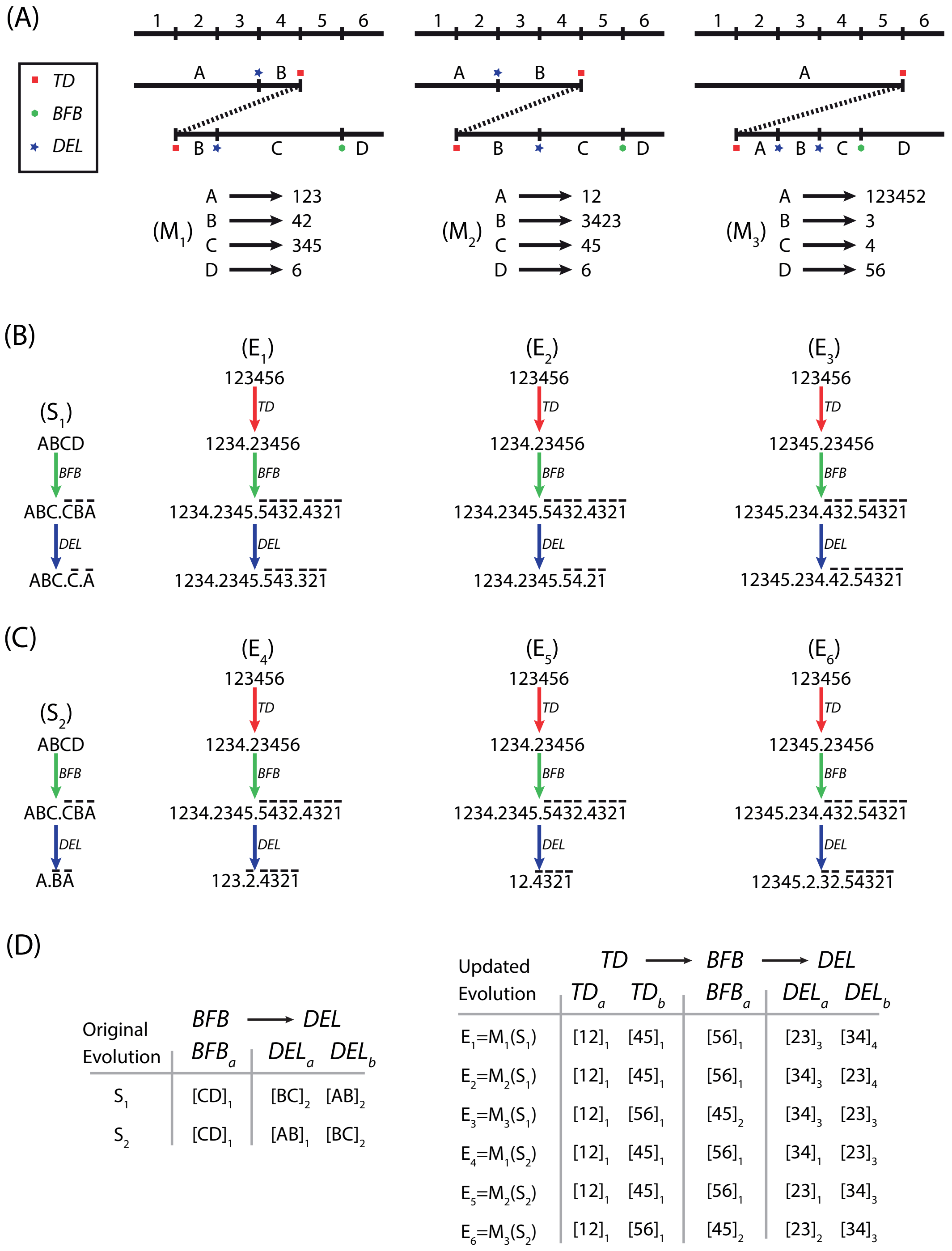
Firstly then, we consider breakpoint usage. In Fig. 5(D) the left table gives the breakpoints implicated for the two original evolutions (S_1, S_2) corresponding to $BFB \rightarrow DEL$ in Fig. 5(B,C). For example, take the entry $[AB]_2$ for the end point of the deleted region in evolution S_1. This indicates the second adjacency of AB in the word $ABC.\overline{CB} \cdot \overline{A}$ is implicated (the position with symbol $\cdot$). Now a mapping, such as M_1, takes $A \rightarrow 123$, $B \rightarrow 42$ and $C \rightarrow 345$, meaning the word is updated to $1234.2345.\overline{5432}.\overline{4} \cdot \overline{321}$. The breakpoint position can still be identified, but the breakpoint label becomes $[34]_4$ with the new reference (as given in the first row, last entry of the right table in Fig. 5(D)). This applies in general; the action of a mapping on breakpoints is well defined for a given evolution; in this case we have $M_1([AB]_2) = [34]_4$. Note that the mapped position $[34]$ does not depend upon the evolution involved, but the copy (subscript 4) does. We thus have (i) an updated rearrangement sequence (with new first rearrangement), (ii) an updated list of breakpoint positions, and (iii) a new first word $12\ldots n$. As noted in Section 2.2, these three things are all that is needed to construct an evolution and we have Property A; the updated evolution $E = M(S)$ is well defined for any mapping M and original evolution S.

4.1.2. Uniqueness

Secondly, note that each breakpoint is implicated uniquely by the infinite sites assumption. For example, each row of the left table in Fig. 5(D) uses original breakpoint positions $[AB], [BC], [CD]$ exactly once, and each row of the right table in Fig. 5(D) uses updated breakpoint positions [12], [23], ..., [56] exactly once. Now, consider two distinct breakpoints $[A_iA_{i+1}]_p$ and $[A_jA_{j+1}]_q$ from the original reference ($A_i, A_j \in \{A, B, C, D\}$). Under any mapping M, these will remain distinct. Either $i \neq j$ in which case the reference positions differ in both the original and updated reference by the infinite sites assumption, or we have $i = j$, $p \neq q$ and the copy of the same breakpoint is different, something that must remain true after the mapping has relabelled segments. Thus we have the observation that distinct breakpoints are mapped to distinct breakpoints. That is, $M(B) \neq M(B')$ for any breakpoints $B \neq B'$.

4.1.3. Segments

Thirdly, consider the consecutive nature of segments. The reference contains consecutive segments, such as 123456, for example. If we have a rearrangement, such as $BFB([56]_1)$ acting at breakpoint $[56]_1$, for example, we end up with modified word $12345.\overline{54321}$. The end product now has a new novel somatic connection $[5\overline{5}]_1$ (at the decimal point). Note that the



Original Evolution	BFB →	DEL	
	BFB_a	DEL_a	DEL_b
S_1	$[CD]_1$	$[BC]_2$	$[AB]_2$
S_2	$[CD]_1$	$[AB]_1$	$[BC]_2$

Updated Evolution	TD →		BFB →	DEL	
	TD_a	TD_b	BFB_a	DEL_a	DEL_b
E_1=$M_1(S_1)$	$[12]_1$	$[45]_1$	$[56]_1$	$[23]_3$	$[34]_4$
E_2=$M_2(S_1)$	$[12]_1$	$[45]_1$	$[56]_1$	$[34]_3$	$[23]_4$
E_3=$M_3(S_1)$	$[12]_1$	$[56]_1$	$[45]_2$	$[34]_3$	$[23]_3$
E_4=$M_1(S_2)$	$[12]_1$	$[45]_1$	$[56]_1$	$[34]_1$	$[23]_3$
E_5=$M_2(S_2)$	$[12]_1$	$[45]_1$	$[56]_1$	$[23]_1$	$[34]_3$
E_6=$M_3(S_2)$	$[12]_1$	$[56]_1$	$[45]_2$	$[23]_2$	$[34]_3$

Figure 5. A sample of updated evolutions. (A) Three sample reference mappings (M_1-M_3) corresponding to the update of evolution $BFB \rightarrow DEL$ to $TD \rightarrow BFB \rightarrow DEL$. The original reference $ABCD$ is mapped onto updated reference 123456. (B,C) Respectively describe how evolutions S_1, S_2 are updated under the mappings of (A). (D) Provides breakpoint reference positions for original and updated evolutions (e.g., $[34]_3$ for the end of deleted region in evolution $E_5 = M_2(E_2)$ indicates the third copy (subscript) of the position between segments 3 and 4 is utilized).

segments either side of the connection remain in (their original) consecutive order, albeit reversed on the right side of the connection. This applies in general; segments between somatic connections run in consecutive order.

4.1.4. Chronology

Lastly, we consider the chronology of events. To do this we again consider the action of the mappings on breakpoints (rather than segments). In Fig. 5(D) the second table (with updated evolutions) has six rows corresponding to the action of three example mappings $M_1 - M_3$ to two original evolutions

(S_1 and S_2) for the updated rearrangement sequence $TD \rightarrow BFB \rightarrow DEL$. Each row contains the positions of breakpoints using the updated reference. For example, updated evolution $E_1 = M_1(S_1)$ has a final deletion from breakpoint $[23]_3$ to $[34]_4$. This is applied to the penultimate word $1234.2345.\overline{5432}.\overline{4321}$, removing $\overline{24}$ to give $1234.2345.\overline{543}.\overline{321}$. These breakpoints can be lifted to positions on the structure representing the mapping; the blue stars in the structure for M_1 in Figure 5(A).

Note that if we compare the mappings M_1 and M_2 that are applied to original evolution S_1 to give updated evolutions $E_1 = M_1(S_1)$ and $E_2 = M_2(S_2)$ (the first two rows of the second table in Fig. 5(D)), the only differences in mapped breakpoints arise in the final DEL event. More specifically, the first rearrangement (TD) uses breakpoints $([12]_1, [45]_1)$ in both cases, the second rearrangement (BFB) uses $[56]_1$ in both cases, whereas in the last rearrangement (DEL), evolution E_1 uses $([23]_3, [34]_4)$, whereas E_2 uses $([34]_3, [23]_4)$. This is also apparent in Fig. 5(A), where we see the blue stars for the DEL are positioned differently when comparing the first two diagrams (for M_1 and M_2), whereas the red squares (TD) and green circle (BFB) are similarly positioned. Conversely, when comparing M_2 and M_3, we see that all three rearrangements are acting at different positions (e.g. the red squares for TD_b take distinct positions between the second and third diagrams). This is reflected in the second table in Fig. 5(D), where all three breakpoint sets in the rows for $E_2 = M_2(S_1)$ and $E_3 = M_3(S_1)$ differ. Thus chronologically, M_1 and M_2 differ in breakpoint usage at the last rearrangement, whereas M_2 and M_3 differ from the first event onwards.

More generally, for two mappings M and M' updating an evolutions S to $E = M(S)$ and $E' = M'(S)$, respectively, we can compare the positions of breakpoints in the corresponding mapping structures. The first rearrangement (chronologically speaking, that is, the first in the sequence $R_0 \rightarrow R_1 \rightarrow R_2 \dots$) with different breakpoint positions in the mapping structures will be the one that first cause the corresponding updated evolutions E and E' to differ.

We next utilize these observations to validate Properties B and C.

4.2. Single Evolutions with Distinct Mappings

Firstly then, to consider Property B, suppose we have a single evolution (such as S_1 in Fig. 5B, for example), and two distinct reference mappings (say M_1 and M_2). Then we have two corresponding updated evolutions (E_1 and E_2). Now, $E_1 = M_1(S_1)$ and $E_2 = M_2(S_1)$ are identical apart from the last word in their respective sequences. As discussed in the last section, this is because when the breakpoints of the first two rearrangements are considered (chronologically), the TD (red squares in Fig. 5(A)) and BFB (green hexagons) are positioned identically along both mapping structures, whereas the breakpoints for the DEL (blue stars) were in different locations. Thus it is only when the DEL is implemented in the last step that the evolutions differ. Conversely, if we compare the action of M_1 to M_3 (on S_1), the positions of the breakpoints on the corresponding mapping structures for all three rearrangements are distinct, thus the evolutions differ as soon as the first rearrangement is implemented (i.e. the second words in the corresponding evolutions $E_1 = M_1(S_1)$ and $E_3 = M_1(S_1)$ are distinct).

Now in general, we have an original sequence of n rearrangements of the form $R_1 \rightarrow R_2 \rightarrow \cdots \rightarrow R_n$. If S is a corresponding original evolution then we can write this as

$$W_1 \xrightarrow{B_1} W_2 \xrightarrow{B_2} \dots \xrightarrow{B_n} W_{n+1} \qquad [S] \tag{1}$$

where W_i is the i^{th} word in the evolution, and B_i represents the breakpoints implicated by rearrangement R_i in the evolution (so B_i is a vector of $\beta(R_i)$ breakpoints; see Table 1).

Now, the introduction of a new first rearrangement results in an updated rearrangement sequence $R_0 \rightarrow R_1 \rightarrow \cdots \rightarrow R_n$. Consider two corresponding updates of the evolution S under distinct mappings M_1 and M_2. Then we obtain two updated evolutions $E_1 = M_1(S)$ and $E_2 = M_2(S)$, along with updated breakpoints $\hat{B}_i = M_1(B_i)$ and $\hat{B}'_i = M_2(B'_i)$. Now if these two mappings are distinct (in the sense of the different mapping structures counted in Section 3), the breakpoint positions on the mapping structure must differ somewhere (between E_1 and E_2). There must therefore be a rearrangement R_i with minimum value i that these differences correspond to (by the chronological properties discussed in Section 4.1.4). That is, $\hat{B}'_k = \hat{B}_k$ for $k < i$ and $\hat{B}'_i \neq \hat{B}_i$. Prior to rearrangement R_i, both mappings are using the same breakpoint positions and the evolutions will contain the same sequence of reference words prior to this event. At the next rearrangement different reference positions are implicated, so different pieces of DNA are moved, copied or deleted, and so the next word in each sequence must therefore differ. Then if V_k and V'_k represent the updated words, we can represent the updated evolutions as follows:

$$\begin{tikzcd}
 & & & & V_{i+1} \arrow[r, "\hat{B}_{i+1}"] & \dots [E_1] \\
V_0 \arrow[r, "\hat{B}_0"] & V_1 \arrow[r, "\hat{B}_1"] & V_2 \dots V_{i-1} \arrow[r, "\hat{B}_{i-1}"] & V_i \arrow[ur, "\hat{B}_i"] \arrow[dr, "\hat{B}'_i"'] & & \\
 & & & & V'_{i+1} \arrow[r, "\hat{B}'_{i+1}"] & \dots [E_2]
\end{tikzcd} \tag{2}$$

The two evolutions thus bifurcate at rearrangement R_i and the two sequences of words are not identical, as required for Property B.

4.3. Distinct Evolutions After Mapping

Secondly then, for Property C we consider two distinct original evolutions corresponding to a particular sequence of rearrangements, and need to show that after mappings are applied, the updated evolutions remain distinct (the two mappings involved do not have to be distinct).

For example, in Fig. 5(B,C), S_1 and S_2 are distinct original evolutions corresponding to the rearrangement sequence $BFB \rightarrow DEL$. We also have three sample mappings M_1-M_3 corresponding to the updated rearrangement sequence $TD \rightarrow BFB \rightarrow DEL$. Then if we apply any of these mappings to S_1 and S_2, the updated evolutions that are obtained remain distinct (any of $E_1 - E_3$ differ to any of $E_4 - E_6$ in Fig. 5(B,C)), as required.

Now, for this example, the two original evolutions are $S_1 : ABCD \rightarrow ABC.\overline{CBA} \rightarrow ABC.\overline{C}.\overline{A}$ along with $S_2 : ABCD \rightarrow ABC.\overline{CBA} \rightarrow A.\overline{BA}$. Thus the two evolutions differ in their last word. The initial rearrangement BFB acts on the same breakpoint $[CD]_1$ in both evolutions (see the first table in Fig. 5D), meaning the same second word $ABC.\overline{CBA}$ occurs. However, the final DEL rearrangement acts on different breakpoints; in S_1 we have operation $DEL([BC]_2, [AB]_2)$ giving final word $ABC.\overline{C}.\overline{A}$, whereas in S_2 we have operation $DEL([AB]_1, [BC]_2)$ giving a different final word $A.\overline{BA}$. More generally then we can diagrammatically represent the two original evolutions as follows.

Suppose that the two original evolutions S_1 and S_2 correspond to a rearrangement sequence $R_1 \to R_2 \to \cdots \to R_n$. As they are distinct, we can assume that they first differ at the $i+1^{\text{th}}$ word in their sequences. Then we represent original evolutions as follows:

$$W_1 \xrightarrow{B_1} W_2 \xrightarrow{B_2} W_3 \dots W_{i-1} \xrightarrow{B_{i-1}} W_i \begin{array}{l} \overset{B_i}{\nearrow} \;\; W_{i+1} \xrightarrow{B_{i+1}} \dots [S_1] \\ \underset{B'_i}{\searrow} \;\; W'_{i+1} \xrightarrow{B'_{i+1}} \dots [S_2] \end{array} \tag{3}$$

Here B_k and B'_k denote the vector of breakpoints that rearrangement R_k acts upon, in evolutions S_1 and S_2, respectively (so $B_k = B'_k$ for $1 \leq k \leq i-1$). The terms W_k and W'_k represent the words in original evolutions S_1 and S_2, respectively ($W_k = W'_k$ for $1 \leq k \leq i$). Thus the words are identical in both evolutions for $k \leq i$, but the breakpoint positions implicated by rearrangement R_i are different ($B_i \neq B'_i$), meaning the next word in the respective evolutions differ ($W_{i+1} \neq W'_{i+1}$), and the evolutions bifurcate at rearrangement R_i.

Next we need to consider the effect of updating the evolution with the introduction of a new first rearrangement R_0. This results in updated rearrangement sequence $R_0 \to R_1 \to R_2 \to \cdots \to R_n$. The evolutions are also updated to $E_1 = M_1(S_1)$ and $E_2 = M_2(S_2)$ via the action of corresponding mappings M_1 and M_2 on the breakpoints, resulting in updated breakpoints $\hat{B}_k = M_1(B_k)$ and $\hat{B}'_k = M_2(B'_k)$ ($k \leq n$). We also obtain corresponding words in the updated evolutions V_k and V'_k ($0 \leq k \leq n+1$). Now suppose we have $\hat{B}_j = M_1(B_j) \neq M_2(B_j) = \hat{B}'_j$ for some minimal value j (so that $\hat{B}_k = \hat{B}'_k$ for $k < j$). Then the first $j+1$ words in the updated evolution are the same ($V_k = V'_k$ for $0 \leq k \leq j$). Furthermore, the distinct action of the mappings on B_j means that the next word in the sequences must be distinct, and the updated evolutions of Eq. 3 can then be represented as

$$V_0 \xrightarrow{\hat{B}_0} V_1 \xrightarrow{\hat{B}_1} V_2 \dots V_{j-1} \xrightarrow{\hat{B}_{j-1}} V_j \begin{array}{l} \overset{\hat{B}_j}{\nearrow} \;\; V_{j+1} \xrightarrow{\hat{B}_{j+1}} \dots [E_1] \\ \underset{\hat{B}'_j}{\searrow} \;\; V'_{j+1} \xrightarrow{\hat{B}'_{j+1}} \dots [E_2] \end{array} \tag{4}$$

Then provided such a value j exists, the evolutions bifurcate at rearrangement R_j and Property C is satisfied. Now, there are three things that can happen when reference mappings are applied as above, depending on how the updated bifurcation event (R_j action in Eq. 4) compares to the original (R_i action in Eq. 3). These cases are treated in turn, corresponding to $j < i$, $j = i$ and $j > i$, respectively.

Case I ($j < i$): The two mappings implicate distinct breakpoints at rearrangement R_j, meaning the updated evolutions diverge at an earlier rearrangement than the original (which differed after the action of R_i). That is, we have $M_1(B_k) = M_2(B_k)$ for $k < j$ but $M_1(B_j) \neq M_2(B_j)$. Subsequently, we find that although $W_j = W'_j$ in the original evolution, the updated evolutions now differ at rearrangement R_j, with $V_j \neq V'_j$, and the bifurcation point occurs earlier. However, the updated evolutions E_1 and E_2 are thus still distinct, as required for Property C.

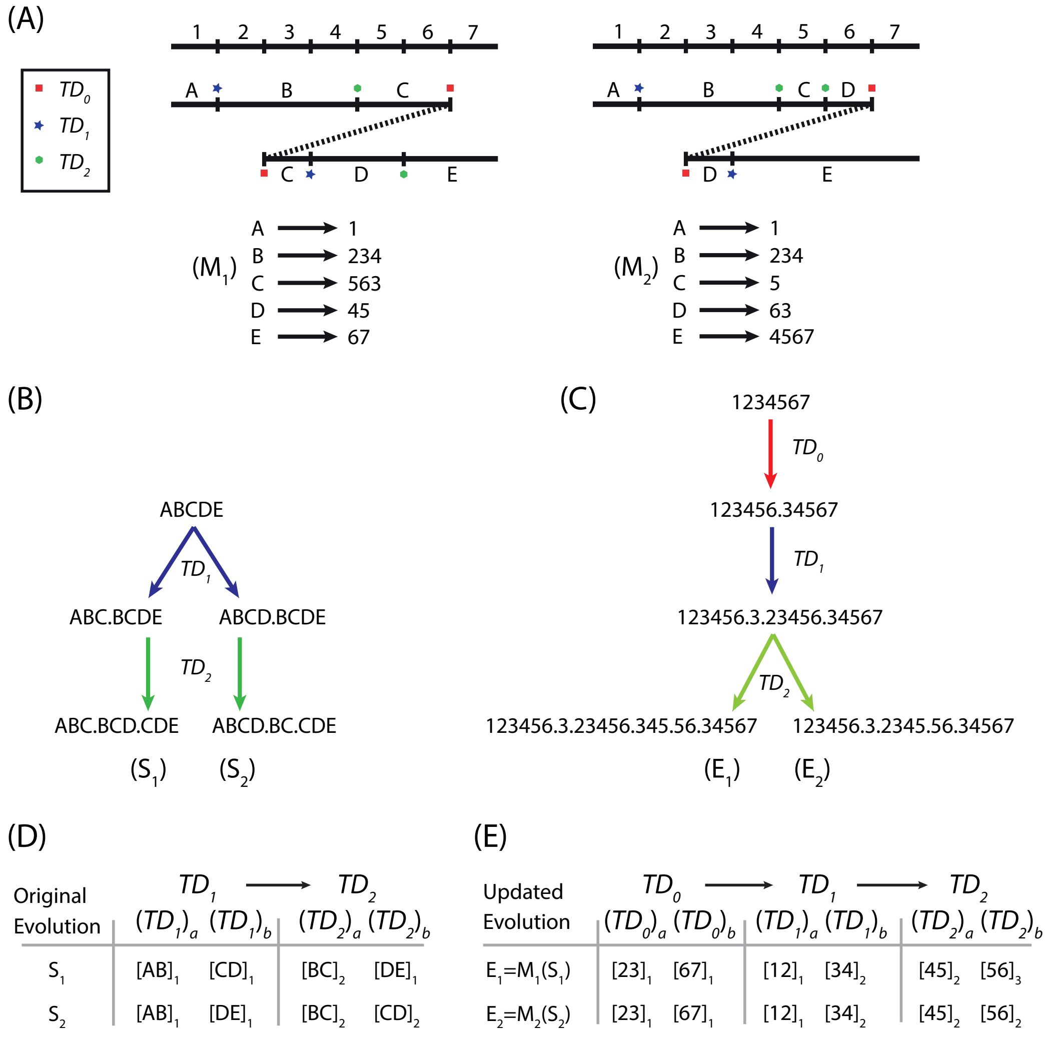
For example, consider the updated evolutions $E_1 = M_1(S_1)$ and $E_6 = M_3(S_2)$ from Fig. 5. Now, S_1 and S_2 diverge at the last rearrangement ($i = 2$). When comparing the breakpoints acted upon by M_1 and M_3 (see the first and last row in the second table of 5(D)) we see that M_1 and M_3 implicate distinct reference positions for breakpoints in all three rearrangements ($j = 0$), so applying M_1 to S_1 and M_3 to S_2 will see a difference occur at the second word $1234.23456 = V_1 \neq V'_1 = 12345.23456$ (after the actions of the first rearrangement $R_0 = TD$) in the updated evolutions. The differences between the two original evolutions (occurring at the final (DEL) rearrangement when comparing S_1 and S_2) are superseded because the mapping now induces an earlier dissimilarity after the first (TD) rearrangement, and the two updated evolutions remain distinct, as required.

Case II ($j = i$): Differences between mapped breakpoint positions occur at the same point as the original evolutions. More specifically, this case assumes M_1 and M_2 have the same action on breakpoint sets $B_1 = B'_1, B_2 = B'_2, \ldots, B_{i-1} = B'_{i-1}$, that is, $\hat{B}_k = M_1(B_k) = M_2(B'_k) = \hat{B}'_k$, for $k \leq i-1$, (otherwise we have Case I). Then the updated evolutions both have the same word $V_i = V'_i$. Now we know that $B_i \neq B'_i$. If we further suppose that $\hat{B}_i = M_1(B_i) \neq M_2(B'_i) = \hat{B}'_i$ then the updated words V_{i+1} and V'_{i+1} will be distinct and so the updated evolutions will bifurcate at the same rearrangement R_i as the original evolution. The updated evolutions are again distinct, as required.

For example, one possibility is that the same mapping is applied to both evolutions. Now, S_1 and S_2 diverge at the last step (Fig. 5(B,C)). This is because the last event is a deletion, with S_1 having action $DEL([BC]_2, [AB]_2)$ (deleting $\overline{B}$ from the word $ABC\overline{CBA}$), whereas S_2 has action $DEL([AB]_1, [BC]_2)$ (deleting $BC\overline{C}$ from $ABC\overline{CBA}$), resulting in distinct final words and so evolutions (with bifurcation at the second rearrangement $i = 2$). Now if the single mapping M_1 is applied to both S_1 and S_2, the implicated breakpoints are distinct. Specifically, $([BC]_2, [AB]_2) = B_2 \neq B'_2 = ([AB]_1, [BC]_2))$, which remains true when a single mapping is applied to both breakpoints ($B_1 \neq B_2 \implies M_1(B_2) \neq M_1(B'_2)$; see Section 4.1.2). The updated evolutions $E_1 = M_1(S_1)$ and $E_4 = M_1(S_2)$ thus remaining distinct, as required, and have the same bifurcation point ($j = 2$).

Case III ($j > i$): We can also have the situation that differences between evolutions S_1 and S_2 and differences between mappings M_1 and M_2 conspire to remove differences in the updated evolutions, resulting in a later bifurcation point. More specifically, we can have distinct breakpoints $B_i \neq B'_i$ (with $B_k = B'_k$ for $k < i$) that are equal after they are updated; $M_1(B_i) = M_2(B'_i)$ (with $M_1(B_k) = M_2(B'_k)$ for $k < i$). The updated evolution no longer bifurcates at rearrangement R_i.

An example of this is given in Fig. 6(B), where we have two original tandem-duplication evolutions S_1 and S_2 corresponding to rearrangement sequence $TD_1 \to TD_2$ that differ



Original Evolution	TD_1 →		TD_2	
	$(TD_1)_a$	$(TD_1)_b$	$(TD_2)_a$	$(TD_2)_b$
S_1	$[AB]_1$	$[CD]_1$	$[BC]_2$	$[DE]_1$
S_2	$[AB]_1$	$[DE]_1$	$[BC]_2$	$[CD]_2$

Updated Evolution	TD_0 →		TD_1 →		TD_2	
	$(TD_0)_a$	$(TD_0)_b$	$(TD_1)_a$	$(TD_1)_b$	$(TD_2)_a$	$(TD_2)_b$
$E_1=M_1(S_1)$	$[23]_1$	$[67]_1$	$[12]_1$	$[34]_2$	$[45]_2$	$[56]_3$
$E_2=M_2(S_2)$	$[23]_1$	$[67]_1$	$[12]_1$	$[34]_2$	$[45]_2$	$[56]_2$

Figure 6. Two original evolutions for tandem-duplication sequence $TD_1 \to TD_2$ that are less distinct after mapping to updated sequence $TD_0 \to TD_1 \to TD_2$. (A) Two mappings M_1 and M_2 from original reference $ABCDE$ to updated reference 1234567. (B) Original evolutions S_1 and S_2 differ after the first rearrangement. (C) Updated evolutions $E_1 = M_1(S_1)$ and $E_2 = M_2(S_2)$ only differ at last word. (D-E) Breakpoint usage for rearrangements under original and updated evolutions.

after TD_1. The rearrangement sequence is updated to $TD_0 \to TD_1 \to TD_2$ with mappings M_1 and M_2 giving updated evolutions $E_1 = M_1(S_1)$ and $E_2 = M_2(S_2)$, respectively. The evolutions E_1 and E_2 now only differ following TD_2 (Fig. 6(C)). In particular, the TD_1 acts on distinct breakpoints between original evolutions S_1 and S_2, with actions $TD_1([AB]_1, [CD]_1)$ and $TD_1([AB]_1, [DE]_1)$ (see Fig. 6D), resulting in distinct words $W_2 = ABC.BCDE$ and $W_2' = ABCD.BCDE$. However, the updated breakpoints $\hat{B}_1 = M_1([AB]_1, [CD]_1) = ([12]_1, [34]_2) = M_2([AB]_1, [DE]_1) = \hat{B}_1'$ are equal, which results in the same updated word $V_2 = V_2' = 123456.3.23456.34567$. Thus the mappings M_1 and M_2 in the updated evolution have vanquished one of the evolutionary differences. However, not all differences are removed and the updated evolutions in Fig. 6(C) are still distinct.

For Property C to apply, we need to ensure that distinct original evolutions and mappings cannot conspire to remove all differences in the updated evolutions, otherwise we have a counterexample to Property C. There are two situations to consider.

Consider first the possibility that B_i and B_i' relate to different copies of the same breakpoint (e.g. $B_i = [AB]_p$ and $B_i' = [AB]_q$ for some $p \neq q$, say $p < q$ without loss of generality). Now the segments in a word between somatic connections are consecutive (Section 4.1.3), increasing by one value per segment (or decreasing if the region is inverted). Thus the breakpoints $[AB]_p$ and $[AB]_q$ (with the same breakpoint value $[AB]$) in the word W_i must be separated by at least one somatic connection. That is we have a word of the form $W_i = W_i' = \ldots[AB]_p \ldots [MN]_r \ldots [AB]_q \ldots$. Note that A and B are consecutive segments from the original reference that will be acted on by rearrangement R_i, whereas M and N are adjacent segments in a somatic connection that has already been formed by rearrangement R_k ($k < i$), and are not consecutive in the reference. Now, under mappings M_1 and M_2, by the assumptions of

Case III, $[AB]_p$ and $[AB]_q$ are mapped to the same value, say $[ij]_k = M_1([AB]_p) = M_2([AB]_q)$. Also, $[MN]_r$ is the product of an earlier rearrangement, and so by assumption is mapped to the same value, say $[rs]_t = M_1([ij]_k) = M_2([ij]_k)$, where in general $s \neq r + 1$. Then we find that word W_i is mapped by M_1 to a word of the form $V_i = \ldots [ij]_k \ldots [rs]_t \ldots$, whereas using mapping M_2 produces word $V_i' = \ldots [rs]_t \ldots [ij]_k \ldots$. Then the number of copies of some breakpoint $[rs]$ that are positioned to the left of $[ij]_k$ must be different in the two words, and so $V_i \neq V_i'$. This contradicts the assumption that $\hat{B}_i = \hat{B}_i'$, and the situation that B_i and B_i' relate to different copies of the same breakpoint cannot arise for Case III.

The second possibility is that B_i and B_i' relate to copies of distinct breakpoints. For example, $B_1 = ([AB]_1, [CD]_1) \neq B_1' = ([AB]_1, [DE]_1)$ (so $i = 1$) in Fig. 6). Now, by assumption the distinct sets B_i and B_i' have the same image under mapping. For this example we have $M_1([AB]_1, [CD]_1) = M_2([AB]_1, [DE]_1) = ([12]_1, [34]_2)$ (see Fig. 6(D,E)). The reason that the distinct breakpoints $[CD]_1$ and $[DE]_1$ have the same image is because $([AB], [CD])$ and $([AB], [DE])$ have the same position on the mapping structures for M_1 and M_2 in Fig. 6(A) (note the blue stars are in the same place in both diagrams). However, positioning breakpoints $[CD]$ in the first mapping structure and $[DE]$ in the second, at the same location on both structures (the lower blue star), along with the fact that $[AB]$, $[BC], \ldots, [DE]$ run consecutively along the structures, means that there must be a difference somewhere else. For this example, the differences manifest in the positions corresponding to rearrangement TD_3 (the green hexagons in Fig. 6(A) are different in the two figures). Consequently, the evolutions differ after TD_3 (Fig. 6(C)), and we cannot remove all differences between the evolutions and an evolutionary difference will manifest at a later stage, and updated evolutions E_1 and E_2 will be distinct, as required for Property C.

4.4. Induction

Thus we find that we have the observations required for Theorem 1. For a rearrangement sequence $R_1 \to R_2 \to \cdots \to R_n$ we start with the last rearrangement to construct the single rearrangement sequence R_n and count the number of evolutions with $\Phi_{R_n}(0)$, the value 0 being used because no breakpoints are formed after the action of rearrangement R_n has been completed. Each of the evolutions formed by R_n now contains $\beta(R_n)$ breakpoints. Now for each of these evolutions, we introduce a new first rearrangement R_{n-1} to produce updated rearrangement sequence $R_{n-1} \to R_n$, resulting in a set of $\Phi_{R_{n-1}}(\beta(R_n))$ distinct evolutions (using Property B from Section 4.2). By Property C from Section 4.3 we also know that each such set of evolutions are mutually distinct across the original $\Phi_{R_n}(0)$ evolutions, giving $\Phi_{R_{n-1}}(\beta(R_n)) \times \Phi_{R_n}(0)$ distinct evolutions in total, all of which now contain $\beta(R_{n-1}) + \beta(R_n)$ breakpoints. We proceed inductively (working backwards chronologically, introducing earlier and earlier rearrangements), resulting in the product formula given in Theorem 1.

5. Conclusions

The sequencing of cancer genomes offers a freeze-frame time snapshot of a genome (at the moment of diagnosis) and understanding the evolution (of rearrangements) that has resulted in the observed genome involves reverse engineering the rearrangement path from the final configuration. This is a path through spaces of rearrangements of the form that we have described (although other classes of rearrangement could be involved). The upshot of this is that the size of the space is too large to fully explore computationally for more than six or seven rearrangement events, even restricting the possibilities to just the classes of rearrangements we have examined. Other sources of information may be useful; longer reads such as those offered by nanopore technologies will help restrict this space, however, even if the entire rearranged chromosome is known, exploring the evolution space to explain the observed chromosome will likely still be a formidable task.

One unexplored issue relates to evolution similarity. Specifically, given a final genomic configuration (and initial reference), finding out how many evolutions start and finish at these loci is desirable. For complicated rearrangement sequences clustering in one genomic region it would seem likely that there will usually be only one evolution that can terminate at a particular genomic configuration. However, there are many cases where this is not so. For example, two overlapping (but not nested) TDs give the same final product irrespective of which TD is done first. Two non-overlapping clusters of rearrangements also have many evolutions with identical final structures. This can also occur for BFB evolutions [16]. Furthermore, pairs of BFB evolutions with distinct final structures, but equal numbers of copies of genomic segments can be found. A more comprehensive understanding of these issues for infinite site rearrangement models, similar in vein to [37], would certainly be of interest.

One possible way to describe the evolution of observed rearranged genomes is to suppose the original contiguous configuration is completely dismantled into pieces (break the genome at all breakpoints initially in one go), perform any required duplication, and then reassembled into the final configuration. Although this can happen with chromothripsis, albeit in a more controlled localized manner, this is an unlikely explanation in general, and furthermore offers little insight into the rearrangement processes that have taken place. Indeed many rearrangement processes are understood as local events occurring during distinct cell cycles. Breakage-fusion-bridge-cycles are known to happen during consecutive cell cycles, for example [2, 16]. Gaining a better understanding of the biological mechanisms that underlie the different rearrangements is obviously a desirable aim, and understanding how corresponding spaces are structured may help.

Finally, the work described pertains to the growth of a single clone. Whilst many cancers are monoclonal (or at least dominated by a single clone), many are polyclonal, with a mixture of competing evolution paths. With more mature development of single cell sequencing, phylogenetic techniques may be possible and may help reduce the complexity of inferring rearrangement evolution.

References

[1] R. A. Helm, A. G. Lee, H. D. Christman, S. Maloy, Genomic rearrangements at rrn operons in salmonella, Genetics 165 (3) (2003) 951–959.

[2] B. McClintock, The stability of broken ends of chromosomes in zea mays, Genetics 26 (2) (1941) 234–282.

[3] G. R. Bignell, T. Santarius, J. C. Pole, A. P. Butler, J. Perry, E. Pleasance, C. Greenman, A. Menzies, S. Taylor, S. Edkins, et al., Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution, Genome Research 17 (9) (2007) 1296–1303.

[4] V. Bafna, P. A. Pevzner, Genome rearrangements and sorting by reversals, SIAM Journal on Computing 25 (2) (1996) 272–289.

[5] S. Hannenhalli, P. A. Pevzner, Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals, Journal of the ACM (JACM) 46 (1) (1999) 1–27.

[6] H. Kaplan, R. Shamir, R. E. Tarjan, A faster and simpler algorithm for sorting signed permutations by reversals, SIAM Journal on Computing 29 (3) (2000) 880–892.

[7] D. A. Bader, B. M. Moret, M. Yan, A linear-time algorithm for computing inversion distance between signed permutations with an experimental study, Journal of Computational Biology 8 (5) (2001) 483–491.

[8] G. Benson, L. Dong, Reconstructing the duplication history of a tandem repeat., in: ISMB, 1999, pp. 44–53.

[9] O. Elemento, O. Gascuel, M.-P. Lefranc, Reconstructing the duplication history of tandemly repeated genes, Molecular Biology and Evolution 19 (3) (2002) 278–288.

[10] O. Gascuel, M. D. Hendy, A. Jean-Marie, R. McLachlan, The combinatorics of tandem duplication trees, Systematic Biology 52 (1) (2003) 110–118.

[11] J. Yang, L. Zhang, On counting tandem duplication trees, Molecular Biology and Evolution 21 (6) (2004) 1160–1163.

[12] D. Bertrand, M. Lajoie, N. El-Mabrouk, Inferring ancestral gene orders for a family of tandemly arrayed genes, Journal of Computational Biology 15 (8) (2008) 1063–1077.

[13] L. Penso-Dolfin, T. Wu, C. D. Greenman, The combinatorics of tandem duplication, Discrete Applied Mathematics 194 (2015) 1–22.

[14] M. Kinsella, V. Bafna, Combinatorics of the breakage-fusion-bridge mechanism, Journal of Computational Biology 19 (6) (2012) 662–678.

[15] S. Zakov, M. Kinsella, V. Bafna, An algorithmic approach for breakage-fusion-bridge detection in tumor genomes, Proceedings of the National Academy of Sciences 110 (14) (2013) 5546–5551.

[16] C. Greenman, S. Cooke, J. Marshall, M. Stratton, P. Campbell, Modeling the evolution space of breakage fusion bridge cycles with a stochastic folding process, Journal of Mathematical Biology 72 (1-2) (2016) 47–86.

[17] S. Yancopoulos, O. Attie, R. Friedberg, Efficient sorting of genomic permutations by translocation, inversion and block interchange, Bioinformatics 21 (16) (2005) 3340–3346.

[18] C. D. Greenman, E. D. Pleasance, S. Newman, F. Yang, B. Fu, S. Nik-Zainal, D. Jones, K. W. Lau, N. Carter, P. A. Edwards, et al., Estimation of rearrangement phylogeny for cancer genomes, Genome Research 22 (2) (2012) 346–361.

[19] J. Ma, A. Ratan, B. J. Raney, B. B. Suh, W. Miller, D. Haussler, The infinite sites model of genome evolution, Proceedings of the National Academy of Sciences 105 (38) (2008) 14254–14261.

[20] S. Bhatia, A. Egri-Nagy, A. R. Francis, Algebraic double cut and join, Journal of Mathematical Biology 71 (5) (2015) 1149–1178.

[21] J. Meidanis, Z. Dias, An alternative algebraic formalism for genome rearrangements, in: Comparative Genomics, Springer, 2000, pp. 213–223.

[22] G. Fertin, A. Labarre, I. Rusu, S. Vialette, E. Tannier, Combinatorics of genome rearrangements, MIT Press, 2009.

[23] M. Kimura, The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, Genetics 61 (4) (1969) 893.

[24] P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, et al., Massive genomic rearrangement acquired in a single catastrophic event during cancer development, Cell 144 (1) (2011) 27–40.

[25] P. Pevzner, G. Tesler, Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution, Proceedings of the National Academy of Sciences 100 (13) (2003) 7672–7677.

[26] G. R. Bignell, C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, et al., Signatures of mutation and selection in the cancer genome, Nature 463 (7283) (2010) 893.

[27] P. Biller, L. Guéguen, C. Knibbe, E. Tannier, Breaking good: accounting for fragility of genomic regions in rearrangement distance estimation, Genome Biology and Evolution 8 (5) (2016) 1427–1439.

[28] A. E. Darling, I. Miklós, M. A. Ragan, Dynamics of genome rearrangement in bacterial populations, PLoS genetics 4 (7) (2008).

[29] O. Attie, A. E. Darling, S. Yancopoulos, The rise and fall of breakpoint reuse depending on genome resolution, in: BMC Bioinformatics, Vol. 12, Springer, 2011, p. S1.

[30] T. M. Nye, Modelling the evolution of multi-gene families, Statistical Methods in Medical Research 18 (5) (2009) 487–504.

[31] D. J. McBride, D. Etemadmoghadam, S. L. Cooke, K. Alsop, J. George, A. Butler, J. Cho, D. Galappaththige, C. Greenman, K. D. Howarth, et al., Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes, The Journal of Pathology 227 (4) (2012) 446–455.

[32] P. Stankiewicz, J. R. Lupski, Genome architecture, rearrangements and genomic disorders, TRENDS in Genetics 18 (2) (2002) 74–82.

[33] V. Bafna, P. Pevzner, Sorting permutations by tanspositions, in: Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms, 1995, pp. 614–623.

[34] V. Bafna, P. A. Pevzner, Sorting by transpositions, SIAM Journal on Discrete Mathematics 11 (2) (1998) 224–240.

[35] N. El-Mabrouk, Genome rearrangement by reversals and insertions/deletions of contiguous segments, in: Annual Symposium on Combinatorial Pattern Matching, Springer, 2000, pp. 222–234.

[36] S. Yancopoulos, R. Friedberg, DCJ path formulation for genome transformations which include insertions, deletions, and duplications, Journal of Computational Biology 16 (10) (2009) 1311–1338.

[37] A. Bergeron, C. Chauve, T. Hartman, K. St-Onge, On the properties of sequences of reversals that sort a signed permutation, in: Proceedings of JOBIM, Vol. 2, Citeseer, 2002, pp. 99–108.