A Maximum Parsimony Principle for Multichromosomal Complex Genome Rearrangements

Pijus Simonaitis
Department of Computer Science, Princeton University, Princeton, NJ, USA

Benjamin J. Raphael
Department of Computer Science, Princeton University, Princeton, NJ, USA

Abstract

Motivation. Complex genome rearrangements, such as chromothripsis and chromoplexy, are common in cancer and have also been reported in individuals with various developmental and neurological disorders. These mutations are proposed to involve simultaneous breakage of the genome at many loci and rejoining of these breaks that produce highly rearranged genomes. Since genome sequencing measures only the novel adjacencies present at the time of sequencing, determining whether a collection of novel adjacencies resulted from a complex rearrangement is a complicated and ill-posed problem. Current heuristics for this problem often result in the inference of complex rearrangements that affect many chromosomes.

Results. We introduce a model for complex rearrangements that builds upon the methods developed for analyzing simple genome rearrangements such as inversions and translocations. While nearly all of these existing methods use a maximum parsimony assumption of minimizing the number of rearrangements, we propose an alternative maximum parsimony principle based on minimizing the number of chromosomes involved in a rearrangement scenario. We show that our model leads to inference of more plausible sequences of rearrangements that better explain a complex congenital rearrangement in a human genome and chromothripsis events in 22 cancer genomes.

2012 ACM Subject Classification Applied computing → Bioinformatics

Keywords and phrases Genome rearrangements, maximum parsimony, cancer evolution, chromothripsis, structural variation, affected chromosomes

Digital Object Identifier 10.4230/LIPIcs.WABI.2022.21

Supplementary Material *Software*: `https://github.com/raphael-group/MICRO`

Funding This work was supported by grants U24CA248453 from the US National Cancer Institute (NCI).

1 Introduction

Genome rearrangements transform a genome by breaking two or more genomic loci and joining the resulting chromosomal segments in a different order. The most common and most well-studied genome rearrangements are *simple* rearrangements such as translocations and inversions (reversals) that break a genome at two locations and join the resulting free ends. More recently, *complex* rearrangements [28] that involve simultaneous breaking and joining at several loci – sometimes up to hundreds of loci – have been reported. These complex rearrangements include chromothripsis [33] and chromoplexy [5], which were reported in more than 25% of the cancer patients in recent pan-cancer genome sequencing studies [36, 12, 16]. Complex rearrangements have also been reported in patients harboring congenital and developmental disorders [31, 40] as well as seemingly healthy individuals [13]. A number of

22nd International Workshop on Algorithms in Bioinformatics (WABI 2022).
Editors: Christina Boucher and Sven Rahmann; Article No. 21; pp. 21:1–21:22
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

putative mechanisms for complex rearrangements have been proposed and experimentally induced in cell lines [35, 41, 22], but the precise mechanisms that lead to these mutations occurring in human cells remain largely unknown [27].

Inferring complex rearrangements from genome sequencing data is a difficult problem, since sequencing data measures only *novel adjacencies* – pairs of loci that are adjacent in the sequenced genome but distant in the human reference genome – and determining which combination of these novel adjacencies constitute a complex rearrangement vs. multiple simple rearrangements is a complicated and ill-posed problem. Multiple methods have been developed to predict complex rearrangements from sequencing data [5, 12, 7, 23, 16, 21]. These methods use various heuristics to identify clusters of novel adjacencies whose ends, or *extremities*, are close together on the human reference genome. However, the sensitivity and specificity of these methods in distinguishing one-off complex rearrangements from progressive sequences of simple rearrangements remains a source of debate [24, 19, 26].

There is an extensive literature studying sequences of genome rearrangements, and finding the *minimum* number of genome rearrangements that transform one genome into another. Most of this work focuses on simple rearrangements and analyses sequences of inversions [18], sequences of inversions and translocations [17, 29], or sequences of *double-cut-and-join* operations [37, 32, 39], also called 2-breaks, that break a genome and two loci and join the resulting free ends. A complex genome rearrangement can be modeled by a *k-break* that breaks a genome at k loci and joins back the resulting chromosomal strands thus introducing k novel adjacencies [4, 19, 10], a generalization of a *double-cut-and-join.*

Computing the minimum number of rearrangements that transform one genome into another follows the principle of maximum parsimony, or finding the simplest explanation for the data. However, once complex rearrangements are allowed, it is unclear how to define *simplest.* For example, it may be possible to transform one genome into another with a *single* k-break, but using an extremely large value of k. Indeed complex rearrangements involving more than a hundred simultaneous breaks in cancer genomes have been proposed [12, 16]. However, explaining data with a single arbitrary complex rearrangement is in a sense trivial: this explanation is parsimonious in one criterion (minimizing number of rearrangements) but not parsimonious under another criterion (minimizing k, or complexity of allowed operations). It is generally unknown what values of k are reasonable to analyze complex rearrangements using k-breaks. Thus, there is a gap between the parsimonious k-break scenarios studied in the genome rearrangement literature – where the values of k are relatively small and a minimum sequence of rearrangements is computed – and the arbitrary complex rearrangements described in the cancer genomics literature – where the value of k is unbounded leading to explanations of the data with a single, arbitrarily complicated rearrangement. These two approaches are in a sense two extremes and a natural question is whether there is an intermediate between these extremes.

In this paper, we propose an alternative maximum parsimony principle for studying complex rearrangements that involve multiple chromosomes. Specifically, based on biological knowledge of the mechanisms of complex rearrangements, we propose that a complex rearrangement might be unlikely to simultaneously break a large number of chromosomes. Supporting this approach are two non-exclusive cellular mechanisms that have been proposed to explain a shattering of one or a few chromosomes followed by a random joining of the resulting chromosomal segments [28]. First, defects in chromosomal segregation or formation of acentric chromosomes might lead to a physical isolation and rearrangement of one or a few chromosomes in an aberrant nuclear structure called micronucleus [41]. Second, a dicentric chromosome formed after an end-to-end fusion or a translocation between two chromosomes might get shattered during mitosis [35]. Following these observations, we propose that the *chromosome number*, the maximum number of chromosomes broken by a k-break in a sequence of rearrangements, is a useful statistic for studying complex rearrangements. Further, we propose that minimizing the chromosome number provides an alternative maximum parsimony criterion for evaluating sequences of simple and complex rearrangements that transform one genome into another.

We derive two algorithms to compute the minimum chromosome number of rearrangement scenarios under the *infinite sites assumption* (ISA) [25, 15, 2], also known as the *constraint of no breakpoint reuse* [3], where a genomic locus is involved in at most one genome rearrangement in a sequence of genome rearrangements transforming one genome into another. The first algorithm computes the minimum chromosome number for rearrangement scenarios between two genomes. Unfortunately, current DNA sequencing technologies do not yield complete genomes, but rather measure a set of novel adjacencies that are present in this genome. This measured set is often missing novel adjacencies that are present in the sequenced genome [8], and for sequencing data from bulk tumor this set might be a superposition of novel adjacencies from multiple cancer clones in the tumor [2]. Thus, the second algorithm computes the minimum chromosome number between a genome and a set of novel adjacencies.

We apply our first algorithm to five human genomes that were proposed to harbor congenital complex rearrangements [14, 11, 13, 30] and our second algorithm to 252 cancer genomes that were identified to harbor chromothripsis events [16]. For one of the human genomes and for 22 of the cancer genomes, we derive alternative sequences of rearrangements with lower chromosome number demonstrating that multichromosomal complex rearrangements may be less complicated than previously described.

2 Methods

2.1 Multi-breaks and rearrangement scenarios

Let A be a reference genome and let B be a rearranged genome that is derived from A by a sequence of simple and complex rearrangements. Here, a *genome* is defined as a set of linear and circular DNA molecules called *chromosomes*, each chromosome is partitioned into a sequence of unique directed *synteny blocks*, and pairs of consecutive blocks are separated by *breakpoint regions*. The two *endpoints* of a synteny block are its *extremities*, and an *adjacency* is an unordered pair of extremities separated by a breakpoint region. A *telomere* is an extremity incident to an end of a linear chromosome, and two genomes are *co-tailed* if their sets of telomeres are equal.

A *rearrangement* is a k-break for $k \geq 2$ that breaks a genome at k breakpoint regions and joins the resulting chromosomal fragments back thus forming k new breakpoint regions and modifying the order of the synteny blocks [4]. Specifically, let A be a genome, let $\alpha = \big\{\{u_1, u_2\}, \dots, \{u_{2k-1}, u_{2k}\}\big\}$ be a subset of its adjacencies, and let β be a set $\big\{\{u_{\sigma(1)}, u_{\sigma(2)}\}, \dots, \{u_{\sigma(2k-1)}, u_{\sigma(2k)}\}\big\}$ disjoint from α with σ being a permutation of a set $\{1, \dots, 2k\}$. An ordered pair $\tau = (\alpha, \beta)$ is a *k-break* and we say that it *transforms* A into a genome in which adjacencies α are replaced with adjacencies β. A *multi-break* is a k-break for $k \geq 2$. See Figure 1 for an example of two co-tailed genomes that contain the same synteny blocks.

Given a pair A and B of co-tailed genomes with the same synteny blocks, there is a single multi-break that transforms A into B, formalized in the following lemma.

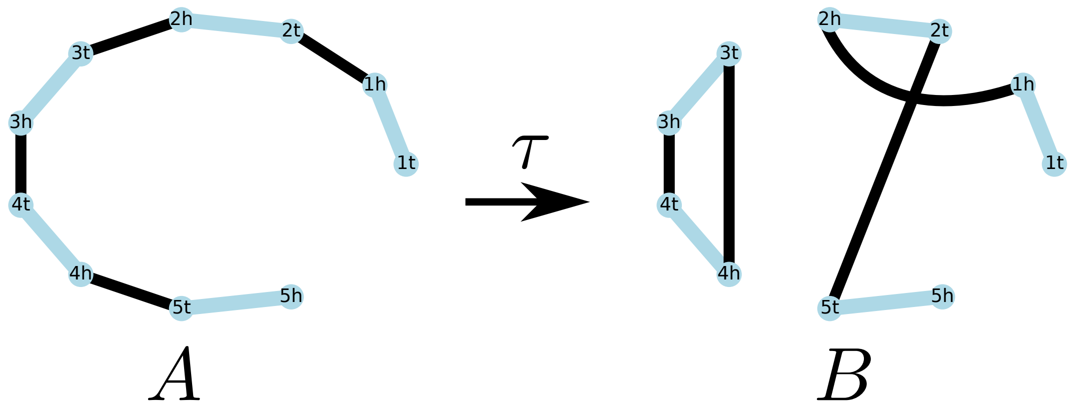

Figure 1 Two co-tailed genomes, A and B, partitioned into five synteny blocks (blue lines) and a 3-break τ transforming A into B. Genome A consists of a single linear chromosome with adjacencies (black lines) $\{\{1_h, 2_t\}, \{2_h, 3_t\}, \{3_h, 4_t\}, \{4_h, 5_t\}\}$. Genome B consists of a circular and a linear chromosomes with adjacencies $\{\{1_h, 2_h\}, \{2_t, 5_t\}, \{3_h, 4_t\}, \{4_h, 3_t\}\}$. A 3-break $\tau = (\alpha, \beta)$ with $\alpha = \{\{1_h, 2_t\}, \{2_h, 3_t\}, \{4_h, 5_t\}\}$ and $\beta = \{\{1_h, 2_h\}, \{2_t, 5_t\}, \{4_h, 3_t\}\}$ transforms A into B.

▶ **Lemma 1.** *Suppose that A and B are co-tailed genomes that contain the same synteny blocks, $E(A)$ are the adjacencies of A that are absent from B, and $E(B)$ are the adjacencies of B that are absent from A. The multi-break $\tau_t = (E(A), E(B))$ is the unique multi-break that transforms A into B.*

The proof of Lemma 1 along with all the other proofs is to be found in Appendix A. We call $\tau_t = (E(A), E(B))$ the *trivial multi-break* for A and B. For now we assume that genomes A and B are co-tailed and contain the same synteny blocks, however this assumption will be relaxed in Section 2.4.

A multi-break *scenario* $\mathbf{T}$ for genomes A and B is a sequence $(\tau_1, \ldots, \tau_l)$ of multi-breaks transforming A into B. Following previous work [4], we say that $\mathbf{T}$ is a *k-break scenario* if all the multi-breaks in $\mathbf{T}$ break at most k adjacencies; i.e., $\tau_i = (\alpha_i, \beta_i)$ where $|\alpha_i| \leq k$ for $i \in \{1, \ldots, l\}$. Appealing to the principle of parsimony, a common problem studied in the genome rearrangement literature is to find the *most parsimonious* rearrangement scenario, or the rearrangement scenario with the fewest number of mutations. Along these lines, a dynamic programming algorithm that is polynomial in k was previously proposed for finding a parsimonious k-break scenario for A and B, and applied to human and mouse genomes with $k = 3$ [4, 3]. However, when examining complex rearrangements, it is unclear what values of k to allow in a k-break scenario. If the number of breaks is unbounded, then, as shown in Lemma 1, the trivial multi-break τ_t transforms A into B and comprises the *trivial multi-break scenario* $\mathbf{T}_t = (\tau_t)$ for A and B. Thus, there is a gap between the parsimonious k-break scenarios considered in the genome rearrangement literature where the number of breaks k is supposed to be ≤ 3 for all the practical purposes, and the multi-break scenarios assumed in the cancer genomics literature, where the number of breaks is unbounded. We are interested in what biologically motivated constraints might replace the number k of breaks in the study of complex rearrangements.

We build upon existing work in cancer genomics and genome rearrangement literature and suppose that evolution by genome rearrangements respects the *Infinite Sites Assumption (ISA)* [25, 2], also known as the *constraint of no breakpoint reuse* [4]. A multi-break scenario $\mathbf{T}$ is said to be a *ISA multi-break scenario* if an adjacency joined by a multi-break in $\mathbf{T}$ is not broken, or *reused*, by any of the subsequent rearrangements in that scenario. Let $\mathcal{I}(A, B)$ be the set of ISA multi-break scenarios transforming A into B. Note that the trivial multi-break scenario $\mathbf{T}_t = (\tau_t)$ is a ISA multi-break scenario for A and B thus ensuring that $\mathcal{I}(A,B)$ is not empty. We say that a multi-break τ is a *ISA multi-break* for A and B if τ appears in some ISA multi-break scenario that transforms A into B. Let $\mathcal{T}(A,B)$ be the set of ISA multi-breaks for A and B. The ISA is based on two underlying assumptions. First, that the probability of a region to be broken by a rearrangement is proportional to the number of base pairs spanned by this region or its *length* [9]. And second, that breakpoint regions are so short that they are unlikely to be reused.

2.2 The chromosome number of a multi-break scenario

We propose to evaluate multi-break scenarios according to the number of chromosomes broken by the multi-breaks. More specifically, given a genome A and a multi-break $\tau = (\alpha, \beta)$ that transforms A, we say that a chromosome of a genome A is *broken* by $\tau = (\alpha, \beta)$ if α includes an adjacency from that chromosome. Let $c(A, \tau)$ be the number of chromosomes broken by τ in A. Let $\mathbf{T} = (\tau_1, \dots, \tau_l)$ be a multi-break scenario transforming genome $A = A_1$ into genome $B = A_{l+1}$, where the multi-break τ_i transforms genome A_i into genome A_{i+1} for $i = 1, \dots, l$. Let $c(\mathbf{T}, \tau_i) = c(A_i, \tau_i)$ be the number of chromosomes broken by τ_i *in* $\mathbf{T}$. We define $c^*(\mathbf{T})$, the *chromosome number* of $\mathbf{T}$, to be the maximum number of chromosomes broken by any multi-break in $\mathbf{T}$; i.e., $c^*(\mathbf{T}) = \max_{1 \le i \le l} c(\mathbf{T}, \tau_i)$. Let $\mathcal{I}(A,B)$ be the set of ISA multi-break scenarios that transform A into B. We aim to find the minimum chromosome number $c(A,B) = \min_{\mathbf{T} \in \mathcal{I}(A,B)} c^*(\mathbf{T})$ of a ISA multi-break scenario transforming A into B, described formally in the following problem.

▶ **Problem 1** (The MINIMUM CHROMOSOME NUMBER or MCN problem). *Given genomes A and B find the minimum chromosome number $c(A,B) = \min_{\mathbf{T} \in \mathcal{I}(A,B)} c^*(\mathbf{T})$, over all ISA multi-break scenarios that transform A into B.*

We say that the MCN problem is *trivial* for genomes A and B if $c(A,B) = c(\mathbf{T}_t)$, where $\mathbf{T}_t$ is the trivial ISA multi-break scenario containing a single multi-break. The simplest non-trivial examples of the MCN problem are for genomes A and B that admit exactly three ISA multi-break scenarios: $\mathbf{T}_t$, $\mathbf{T} = (\tau, \tau')$ and $\mathbf{T}' = (\tau', \tau)$ (See Section 2.3). In this case it can happen that both $\mathbf{T}$ and $\mathbf{T}'$ but not $\mathbf{T}_t$ have the minimum chromosome number $c(A,B)$ (Figure 2, top), it can also happen that only $\mathbf{T}$ but not $\mathbf{T}'$ and $\mathbf{T}_t$ has the minimum chromosome number $c(A,B)$ (Figure 2, bottom).

To solve the MCN problem we first partition the chromosomes of A and B into subsets $\{A_1, \dots, A_m\}$ and $\{B_1, \dots, B_m\}$ such that $c(A,B) = \max_{i \le m} c(A_i, B_i)$. We perform this partition with a help of the *chromosome graph* $C(A,B)$ whose vertices are the block extremities, and edges are the adjacencies of A, the adjacencies of B and the synteny blocks. Note that for a chromosome h of A or B all the synteny blocks and adjacencies of h belong to the same connected component H of the chromosome graph $C(A,B)$ (Figure 3). We say that chromosome h is *included* in component H.

▶ **Theorem 1.** *Let $\{H_1, \dots, H_m\}$ be the connected components of the chromosomes graph $C(A,B)$. Let A_i (resp. B_i) be the genome consisting of the chromosomes of A (resp. B) that are in H_i. Then A_i and B_i are co-tailed and contain the same synteny blocks. Further, the minimum chromosome number $c(A,B) = \max_{i \le m} c(A_i, B_i)$.*

We find the minimum chromosome number $c(A_i, B_i) = \min_{\mathbf{T} \in \mathcal{I}(A_i, B_i)} c^*(\mathbf{T})$ by applying to A_i one by one all the ISA multi-break scenarios for A_i and B_i. We iterate over these scenarios with a help of a bijection introduced in Section 2.3 between $\mathcal{I}(A_i, B_i)$ and a set that we can enumerate.

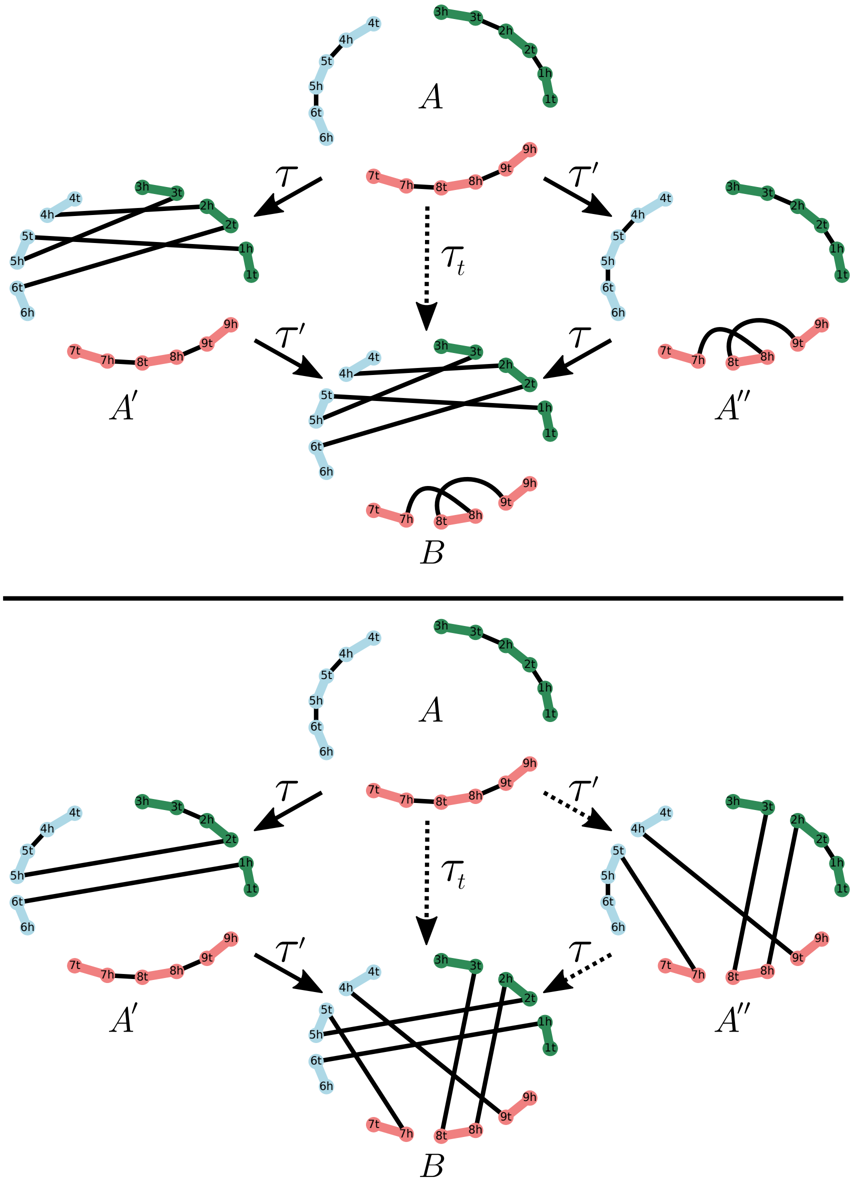

Figure 2 Two examples of genomes A and B for which the Minimum Chromosome Number problem is non-trivial. In both cases there exist three ISA multi-break scenarios $\mathbf{T}_t = (\tau_t)$, $\mathbf{T} = (\tau, \tau')$ and $\mathbf{T}' = (\tau', \tau)$. Solid arrows indicate scenarios with the minimum chromosome number $c(A, B)$, while dashed arrows indicate scenarios with chromosome number greater than $c(A, B)$. (Top) Genomes A and B for which both $\mathbf{T}$ and $\mathbf{T}'$ have the minimum chromosome number; i.e., $c^*(\mathbf{T}) = c^*(\mathbf{T}') = c(A, B) = 2$, while $c^*(\mathbf{T}_t) = 3$. (Bottom) Genomes A and B for which only $\mathbf{T}$ has the chromosome number equal to the minimum chromosome number; i.e., $c^*(\mathbf{T}) = c(A, B) = 2$, while $c^*(\mathbf{T}_t) = c^*(\mathbf{T}') = 3$. Note that in this case τ' breaks two chromosomes in $\mathbf{T}$ but three in $\mathbf{T}'$. Thus, the number of chromosomes broken by a multi-break can vary across the ISA multi-break scenarios.

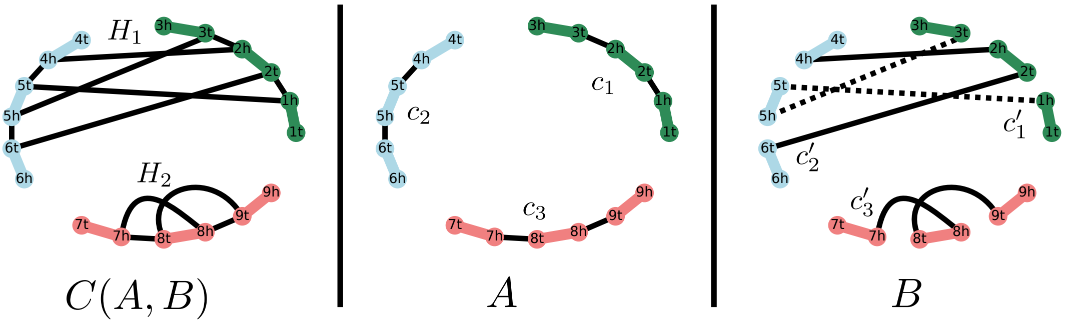

Figure 3 (Left) The chromosome graph $C(A, B)$ has two connected components H_1 and H_2 that respectively include subsets of chromosomes $A_1 = \{c_1, c_2\}$ and $A_2 = \{c_3\}$ of genome A, and $B_1 = \{c'_1, c'_2\}$ and $B_2 = \{c'_3\}$ of genome B. (Right) The adjacencies of a chromosome c'_1 in genome B are dashed in order to distinguish them from the adjacencies of a chromosome c'_2. Note that genomes A_1 and B_1 are co-tailed and contain the same synteny blocks, and similarly for A_2 and B_2.

2.3 Enumeration of ISA multi-break scenarios

In this section we describe a bijection involving ISA multi-break scenarios and the *breakpoint graph* [6, 4], a data structure that is routinely used for a pairwise comparison of genomes, and is defined as follows. The *breakpoint graph* $G(A, B)$ is a 2-edge-colored graph whose vertices are the block extremities, and black and gray edges are respectively the adjacencies of the genomes A and B. Every vertex in $G(A, B)$ is incident to either one black and one gray edge, or to no edges at all. This means that all the non-empty connected components of $G(A, B)$ are alternating cycles whose edges alternate between black and gray. In what follows we say that a *cycle* of $G(A, B)$ is an alternating cycle with at least four edges. It turns out that the ISA multi-break scenarios are related to the set $\mathcal{C}(A, B)$ of the subsets of the cycles of the breakpoint graph $G(A, B)$. Let P be an element of $\mathcal{C}(A, B)$, and let α and β be respectively the black and the gray edges in P. In the proof of Lemma 2 we show that $\tau(P) = (\alpha, \beta)$ is a multi-break. Similarly, let $\mathcal{P}(A, B)$ be the set of the ordered partitions of the cycles of the breakpoint graph, and let $\mathbf{P} = (P_1, \dots, P_l)$ be an element in $\mathcal{P}(A, B)$. In Lemma 2 we establish that a sequence of multi-breaks $\mathbf{T}(\mathbf{P}) = (\tau(P_1), \dots, \tau(P_l))$ is a ISA multi-break scenario for A and B (Figure 4).

▶ **Lemma 2.** *For genomes A and B, the function $\mathbf{T} : \mathcal{P}(A, B) \to \mathcal{I}(A, B)$ between the set $\mathcal{P}(A, B)$ of the ordered partitions of the cycles of the breakpoint graph $G(A, B)$ and the set $\mathcal{I}(A, B)$ of the ISA multi-break scenarios for A and B is a bijection.*

Due to Lemma 2, all the ISA multi-break scenarios for A and B perform the same total number of breaks and contain at most m multi-breaks, where m is the number of the cycles of the breakpoint graph. Given these observations one might expect that there always exists a ISA multi-break scenario for A and B with the minimum chromosome number that contains exactly m multi-breaks, however note that this is not the case for the genomes presented in Figure 4.

2.4 The chromosome number of a complex rearrangement

Current high-throughput DNA sequencing technologies do not measure the rearranged genome B, but rather measure only a set $\mathcal{B}$ of novel adjacencies derived from a sequencing sample. This set $\mathcal{B}$ of novel adjacencies may not correspond to a set of novel adjacencies of a

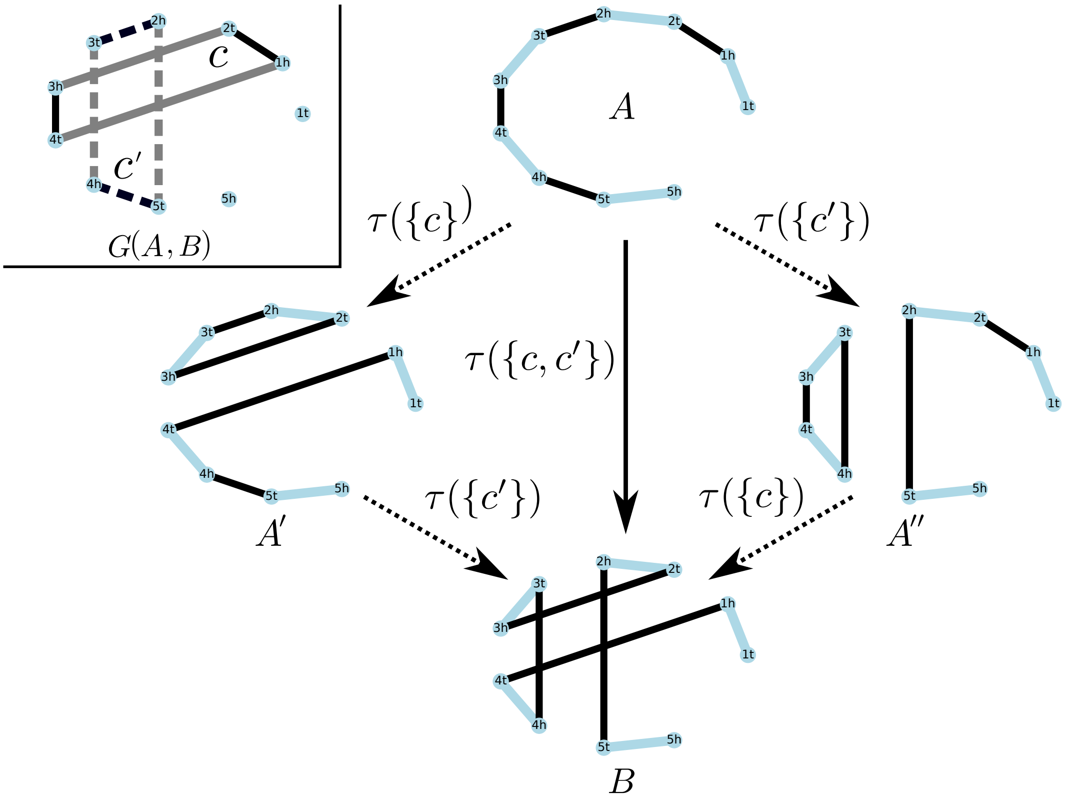

Figure 4 (Left) The breakpoint graph $G(A,B)$ has two cycles c (solid lines) and c' (dashed lines) that admit three ordered partitions $\mathbf{P} = (\{c\}, \{c'\})$ $\mathbf{P}' = (\{c'\}, \{c\})$ and $\mathbf{P}_t = (\{c, c'\})$. (Middle) ISA multi-break scenarios $\mathbf{T}(\mathbf{P}) = (\tau(\{c\}), \tau(\{c'\}))$, $\mathbf{T}(\mathbf{P}') = (\tau(\{c'\}), \tau(\{c\}))$ and $\mathbf{T}_t = \mathbf{T}(\mathbf{P}_t) = (\tau(\{c, c'\}))$ for A and B. The minimum chromosome number is $c(A,B) = 1 = c^*(\mathbf{T}_t)$, while $c^*(\mathbf{T}(\mathbf{P})) = c^*(\mathbf{T}(\mathbf{P}')) = 2$.

unique genome B; for example, $\mathcal{B}$ might be missing some adjacencies or include erroneous adjacencies [1]. The set $\mathcal{B}$ might also include adjacencies from multiple different genomes present in the sample; for example, DNA sequencing data from a bulk tumor is often a mixture of the genomes of different subclones [2]. What is more, a complex rearrangement might not be a multi-break; for example, chromothripsis can delete synteny blocks and chromoanasynthesis can amplify synteny blocks [28]. Finally, even if we were to obtain a rearranged genome B, it might not be co-tailed with the reference genome A. These observations above limit the scope of the Minimum Chromosome Number (MCN) problem.

Below, we introduce the Minimum Chromosome Number of a Complex Rearrangement (MCNR) problem that overcomes the limitations of the Minimum Chromosome Number (MCN) problem. In the MCNR problem our input no longer consists of co-tailed genomes A and B with the same synteny blocks, but of a reference genome A, a set $\mathcal{B}$ of novel adjacencies and a subset $\beta \subseteq \mathcal{B}$ of novel adjacencies that are proposed to result from a complex rearrangement. This input is motivated by the cancer genomics literature which which identifies such subsets $\beta \subseteq \mathcal{B}$ [5, 23, 12, 21, 16]. We propose to evaluate $(A, \mathcal{B}, \beta)$ according to the number of chromosomes broken in an intermediate genome by the complex rearrangement that introduced novel adjacencies β.

A widely reported measure of the "complexity" of a complex rearrangement is the number of reference chromosomes that are *affected by* or *involved in* the adjacencies β introduced by this rearrangement [33, 5, 36, 12, 7]. A chromosome is said to be affected by β if that chromosome includes a block extremity incident to an adjacency in β. Unlike in Section 2.1, in the cancer genomics literature the complex rearrangement is not supposed to be a multi-break; however, note that if there exists a subset α of the adjacencies of the reference genome A such that $\tau = (\alpha, \beta)$ is a multi-break, then the number of chromosomes broken by τ in the reference genome A is equal to the number of chromosomes affected by α in A, and by β in A. In the previous work only the number of chromosomes affected by β in the reference genome was analyzed, however, as it was briefly mentioned by Cortés-Ciriano et al. [12], the complex rearrangement could have rearranged an intermediate genome. Here, we aim to find $c(A, B, \beta)$, the *chromosome number* of β, that is the minimum number of chromosomes affected by β in an intermediate genome under the infinite sites assumption.

First, we formally define the notion of a ISA intermediate genome. As above, the *breakpoint graph* $G(A, \mathcal{B})$ is the 2-edge-colored graph whose black and gray edges are respectively the adjacencies of A and $\mathcal{B}$.

▶ **Definition 1** (ISA intermediate genome). *A multi-break $\tau = (\alpha, \beta)$ is a* ISA multi-break *for a genome A and a set of adjacencies $\mathcal{B}$, if α and β are respectively black and gray edges of a subset of the cycles of the breakpoint graph $G(A, \mathcal{B})$. A genome A' is a* ISA intermediate genome *for A, $\mathcal{B}$ and a subset $\beta \subseteq \mathcal{B}$, if $A = A'$ or if it can be obtained from A by a ISA multi-break $\tau = (\alpha', \beta')$ for A and $\mathcal{B}$ that satisfies $\beta \cap \beta' = \emptyset$.*

Using this definition, we define the following problem.

▶ **Problem 2** (The Minimum Chromosome Number of a Complex Rearrangement or MCNR problem). *Given a genome A, a set of adjacencies $\mathcal{B}$, and a subset $\beta \subseteq \mathcal{B}$, find the chromosome number $c(A, \mathcal{B}, \beta)$ defined as the minimum number of chromosomes affected by β over all ISA intermediate genomes for A, $\mathcal{B}$ and β.*

We say that the MCNR problem is *trivial* if $c(A, \mathcal{B}, \beta)$ is equal to the number of chromosomes affected by β in the reference genome A. The simplest non-trivial example of the MCNR problem is for a triplet $(A, \mathcal{B}, \beta)$ that admits two ISA intermediate genomes A and A' where β affects fewer chromosomes in A' than in A (Figure 5).

To solve the MCNR problem we first partition the chromosomes of A and the adjacencies in $\mathcal{B}$ into subsets $\{A_1, \ldots, A_m\}$ and $\{\mathcal{B}_1, \ldots, \mathcal{B}_m\}$ such that $c(A, \mathcal{B}, \beta) = \Sigma_{i=1}^{m} c(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$. We perform this partition with a help of the *chromosome graph* $C(A, \mathcal{B})$ whose vertices are the block extremities, and edges are the adjacencies of A, the adjacencies in $\mathcal{B}$ and the synteny blocks.

▶ **Theorem 2.** *Let $\{H_1, \ldots, H_m\}$ be the connected components of the chromosomes graph $C(A, \mathcal{B})$. Let A_i be the genome consisting of the chromosomes of A that are in H_i, and let $\mathcal{B}_i$ be the adjacencies in $\mathcal{B}$ that are in H_i. Then the chromosome number $c(A, \mathcal{B}, \beta)$ is equal to $\Sigma_{i=1}^{m} c(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$.*

We find the chromosome number $c(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$ by iterating over all the ISA intermediate genomes for $(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$. We perform this step via a bijection that, by definition, exists between the ISA intermediate genomes for $(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$ and the subsets of the cycles of the breakpoint graph $G(A_i, \mathcal{B}_i)$ that do not contain gray edges from $\mathcal{B}_i \cap \beta$.

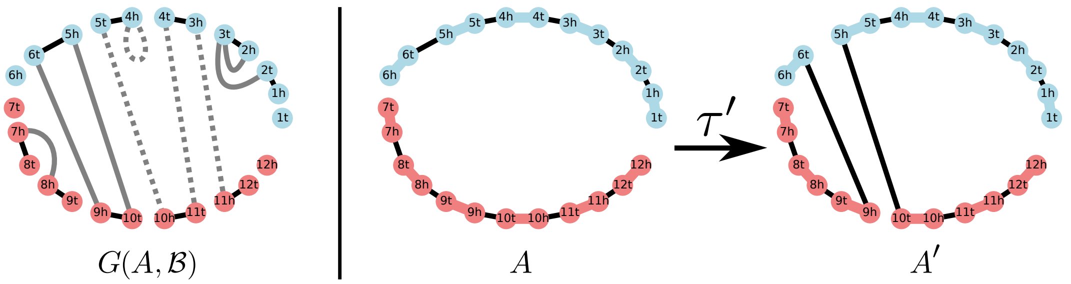

Figure 5 A triplet $(A, \mathcal{B}, \beta)$ for which the Minimum Chromosome Number of a Complex Rearrangement problem is non-trivial. (Left) The breakpoint graph $G(A, \mathcal{B})$ with a subset $\beta \subset \mathcal{B}$ of novel adjacencies (dashed lines) that affects both reference chromosomes. Note that multiple novel adjacencies sharing the same block extremity (3_t) and self-loops (4_h) might occur in both $\mathcal{B}$ and β. (Right) A ISA intermediate genome A' in which β affects a single chromosome. Note that the breakpoint graph $G(A, \mathcal{B})$ contains a single cycle, and A together with A' are the only ISA intermediate genomes for A and $\mathcal{B}$.

3 Results

3.1 The minimum chromosome number of a congenital complex rearrangement

We solved the Minimum Chromosome Number problem for five human genomes harboring congenital complex rearrangements that affect at least three chromosomes [14, 11, 13, 30]. All five genomes are co-tailed with the reference genome. For one of these genomes, the human genome labeled *Case 1* in Eisfeldt et al. [14], we identified a ISA multi-break scenario with a lower chromosome number than previously suggested. Specifically, the published analysis suggested that this genome resulted from a single complex rearrangement that broke chromosomes 2, 8 and 15. In contrast, we derived a ISA multi-break scenario consisting of a multi-break breaking chromosomes 2 and 8 followed by a multi-break breaking chromosome 15 and one of the previously rearranged chromosomes (Figure 6).

For the other four genomes – specifically *Case TL010* and *Case UTR22* from Collins et al. [11], the genome in Eisfeldt et al. [13] and the genome in Plesser et al. [30] – we computed the minimum chromosome number of a ISA multi-break scenario to be equal to 4, which is the same as in the published analysis. Note that to analyze the genome described in Eisfeldt et al. [13], we enumerate 75 ISA multi-break scenarios, the most of the five genomes analyzed.

3.2 The minimum chromosome number of a chromothripsis in cancer

Chromothripsis, is one the most studied types of complex rearrangements in cancer genomes, and is defined as a shattering of *one or a few chromosomes* followed by a random joining of the resulting chromosomal fragments [33, 28]. Some analyses have reported chromothripsis events that include novel adjacencies containing extremities from as many as eighteen reference chromosomes [16, 12, 36]. This seems like an extremely large number of chromosomes to be involved in a simultaneous event, and the current understanding of the molecular mechanisms of genome rearrangements do not indicate simultaneous rearrangements involving more than three chromosomes [41, 35].

To evaluate this discrepancy, we analyzed 252 cancer genomes identified by Hadi et al. [16] as harboring a chromothripsis event that affects at least two reference chromosomes. These genomes form a subset of the 2778 cancer genomes from multiple cancer types with

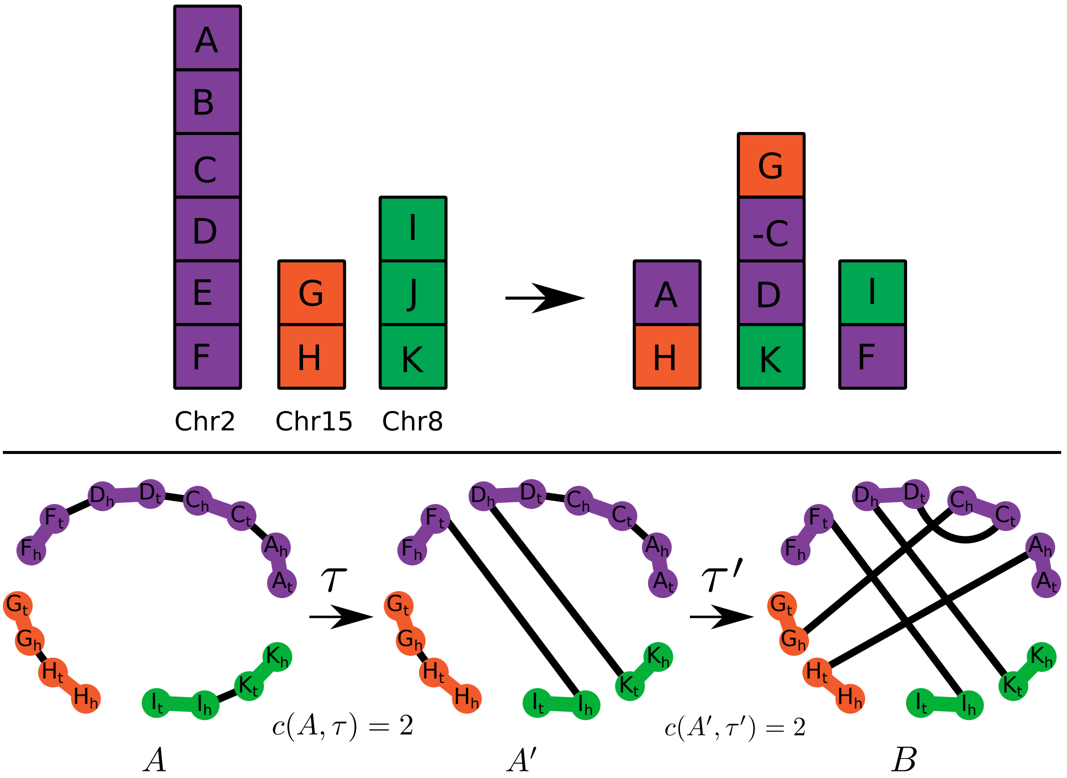
A
B
C
D
E
F
G
H
I
J
K
Chr2
Chr15
Chr8
A
H
G
-C
D
K
I
F
F_h F_t D_h D_t C_h C_t A_h A_t
G_t G_h H_t H_h I_t I_h K_t K_h
τ
τ'
$c(A,\tau) = 2$
$c(A',\tau') = 2$
A
A'
B

Figure 6 Congenital complex rearrangement in human genome *Case 1* from Eisfeldt et al. [14]. (Top left) Reference chromosomes 2 (purple), 8 (green), and 15 (orange) are partitioned into eleven synteny blocks `A` to `K` in accordance with notation published by Eisfeldt et al. [14]. (Top right) Rearranged chromosomes lack synteny blocks `B`, `E` and `J`, and the direction of the block `C` is inverted as indicated by the minus sign. The published analysis suggested that reference chromosomes 2, 8 and 15 were transformed into the rearranged ones by a single complex rearrangement. (Bottom) Genome A is obtained from the reference chromosomes by removing the blocks `B`, `E` and `J`, while genome B corresponds to the rearranged chromosomes. A ISA multi-break scenario (τ, τ') for A and B has the chromosome number equal to two, which contrasts with the published complex rearrangement that simultaneously breaks three chromosomes.

whole-genome sequencing data that are available through gGnome.js portal [16]. In particular, we analyzed a total of 288 triplets $(A, \mathcal{B}, \beta)$ generated from a data structure called the *JaBbA graph* described in [16] (see Section 3.2.3 for further details). These triplets consist of a reference genome A, a set $\mathcal{B}$ of novel adjacencies derived from a cancer sample, and a subset $\beta \subseteq \mathcal{B}$ identified as introduced by a chromothripsis event by Hadi et al. [16]. Since some cancer genomes were identified to harbor multiple chromothripsis events, the number of triplets, 288, is larger than the number 252 of genomes.

3.2.1 Chromothripsis event breaks less chromosomes than it affects

We solved the Minimum Chromosome Number of a Complex Rearrangement problem for all the 288 triplets $(A, \mathcal{B}, \beta)$. For 5 triplets, we identified a ISA intermediate genome in which β affects fewer chromosomes than in the reference genome. This illustrates that a chromothripsis event could have broken fewer chromosomes in an intermediate genome than it affects in the reference.

One such triplet $(A, \mathcal{B}, \beta)$ is from a prostate adenocarcinoma sample `PR-3042` (Figure 7) in which β[1] affects chromosomes 4, 5 and 10. In this sample we found a ISA 3-break $\tau' = (\alpha', \beta')$ that breaks chromosomes 4 and 5, and transforms the reference genome A into a genome in which β affects two chromosomes. This suggests that the chromothripsis event could have broken two rearranged chromosomes instead of the three reference chromosomes affected by β.

3.2.2 The number of affected chromosomes is overestimated

Since identifying a subset $\beta \subseteq \mathcal{B}$ of the novel adjacencies introduced by a chromothripsis event is challenging, we further analyzed the 288 triplets $(A, \mathcal{B}, \beta)$ considering the possibility that false positives in β might lead to an overestimation of the number of the affected reference chromosomes. For 17/288 triplets we found a multi-break $\tau' = (\alpha', \beta')$ such that $\beta \setminus \beta'$ affects fewer reference chromosomes than β and the ratio $\frac{|\beta \cap \beta'|}{|\beta|}$ is less than 0.2. The latter property ensures that only a small fraction of the adjacencies in β are proposed to be false positives, while the former establishes that the multi-break τ' and a chromothripsis event introducing adjacencies $\beta \setminus \beta'$ is a simpler evolutionary explanation than a chromothripsis event introducing adjacencies β.

One such triplet $(A, \mathcal{B}, \beta)$, shown in Figure 8, is from a Barrett's esophagus sample `740_T2_35_24253` in which β affects chromosomes 2, 3, 4, 17 and 22. In this sample we found a ISA 2-break $\tau' = (\alpha', \beta')$ that breaks chromosomes 4 and 17, and a ISA 3-break $\tau'' = (\alpha'', \beta'')$ that breaks chromosomes 2, 3, and 4, such that $\beta \setminus (\beta' \cup \beta'')$ only affects chromosomes 4 and 22. This suggests that the chromothripsis event could have broken two chromosomes (4 and 22) instead of five chromosomes (2, 3, 4, 17 and 22) affected by β.

3.2.3 Processing JaBbA graphs

A JaBbA graph is a data structure that stores synteny blocks (called *intervals*), reference adjacencies (called *REF connections*) and novel adjacencies (called *ALT connections*). See for example the JaBbA graph of a Barrett's esophagus sample `740_T2_35_24253` downloaded from gGnome.js portal [16].

A reference genome A_J and a set of novel adjacencies $\mathcal{B}_J$ can be immediately retrieved from the JaBbA graph, however the breakpoint graph $G(A_J, \mathcal{B}_J)$ thus obtained would have almost no cycles. One technical reason for this is that the synteny blocks are identified with single-nucleotide precision in JaBbA graphs and the breakpoint regions between adjacent synteny blocks are empty. However genome rearrangements oftentimes result in duplications or deletions of the regions surrounding chromosomal breaks [23], and in the genome rearrangement literature [29, 2] such regions are usually interpreted as non-empty breakpoint regions instead of synteny blocks. We thus identify synteny blocks potentially deleted or duplicated by genome rearrangements, and incorporate this information to obtain genome graph A and novel adjacencies $\mathcal{B}$ used in our analysis (see Figure 9 for further details).

3.2.4 Summary

We presented two examples for how a chromothripsis event could have broken fewer chromosomes than it affects in the reference genome. The analyzed chromothripsis events in the gGnome.js portal affect up to seven chromosomes, and in future work it would be of interest to analyze the chromothripsis events identified by Cortés Ciriano et al. [12] that affect up to eighteen chromosomes.

[1] Links to gGnome.js portal [16] for visualizing the JaBbA graphs were tested to work on Google Chrome.

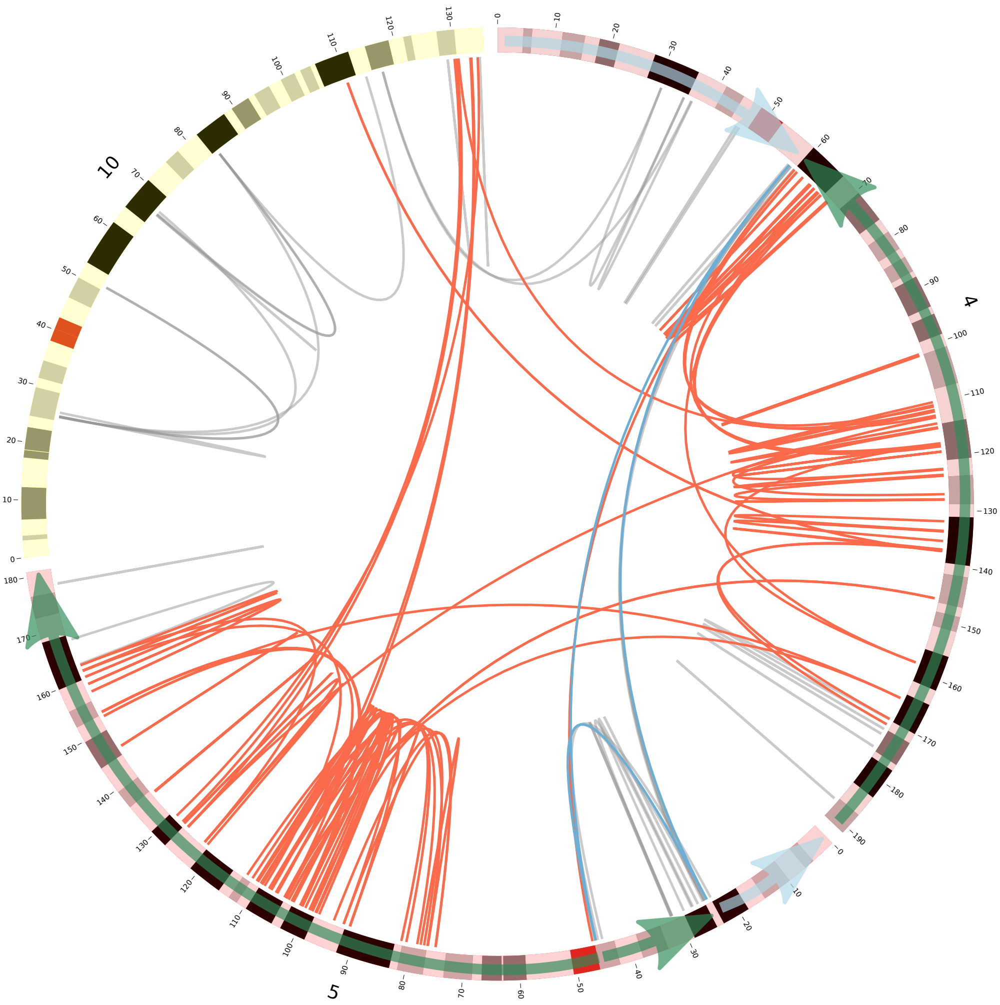

Figure 7 A Circos plot [20] of chromosomes 4, 5, and 10 from prostate adenocarcinoma sample `PR-3042`. Arcs in the plot indicate the novel adjacencies $\mathcal{B}$. Red arcs are the adjacencies β identified as introduced by a chromothripsis event by Hadi et al. [16], the three blue arcs are the adjacencies β' introduced by a ISA 3-break $\tau' = (\alpha', \beta')$ found by our method, and gray arcs are the remaining adjacencies $\mathcal{B} \setminus (\beta \cup \beta')$. The multi-break τ' transforms chromosomes 4 and 5 into the rearranged chromosomes whose segments are indicated by green and blue arrows. Red arcs β affect all three chromosomes (4, 5, and 10) in the reference genome, but only two chromosomes (10 and the green rearranged chromosome) in the rearranged genome. We suggest that the multi-break τ' preceded the chromothripsis event that broke two chromosomes.

4 Discussion

In this work, we introduce a unified model for simple and complex genome rearrangements where each mutation is modeled as a multi-break. Within this model we formulate a novel maximum parsimony principle based on minimizing the number of chromosomes broken by a rearrangement. We formulate the problem of minimizing the chromosome number for the case of a pair of genomes and for the case of a genome and a set of novel adjacencies derived from a sequencing sample. We present exact algorithms to solve both problems under the

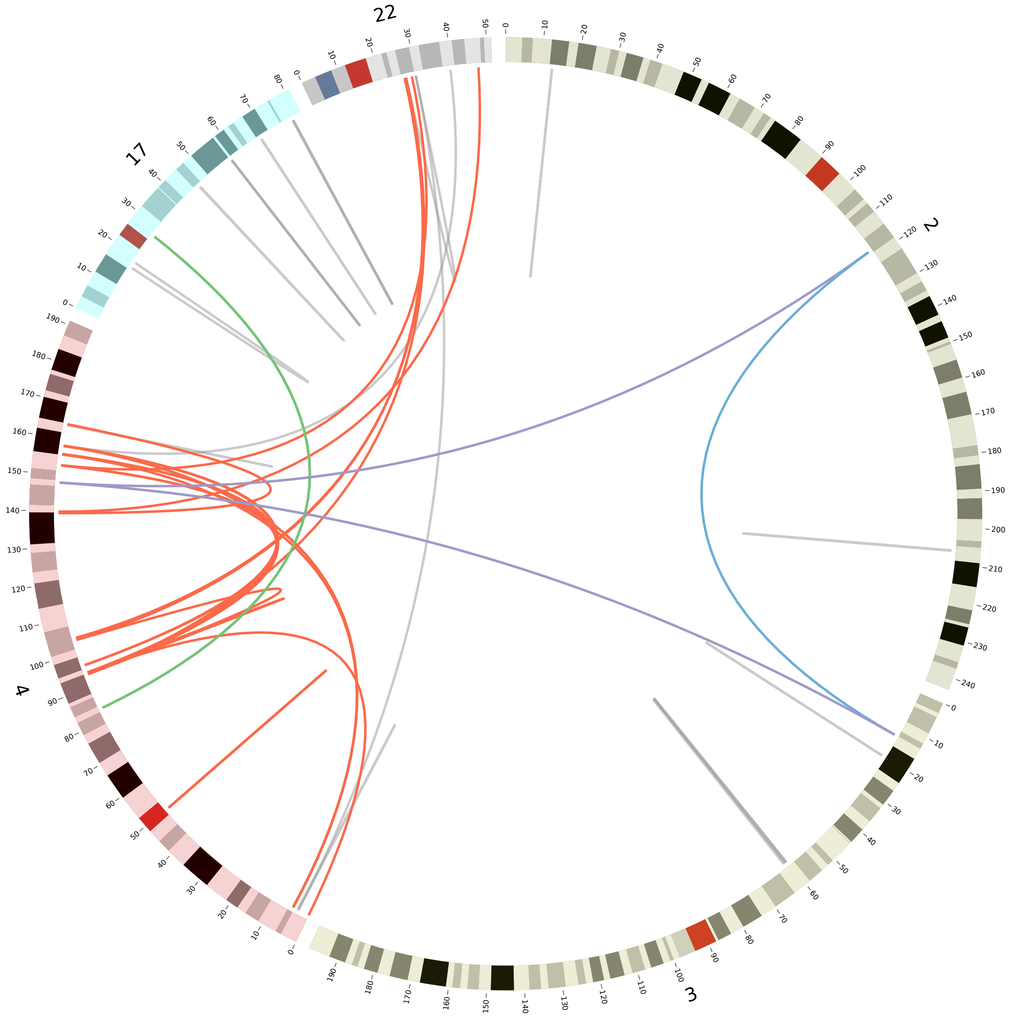

Figure 8 A Circos plot [20] of chromosomes 2, 3, 4, 17, and 22 from a Barrett's esophagus sample `740_T2_35_24253`. Arcs in the plot indicate the novel adjacencies $\mathcal{B}$. A subset $\beta \subseteq \mathcal{B}$, affecting chromosomes 2, 3, 4, 17, and 22, is identified as introduced by a chromothripsis event by Hadi et al. [16], however our analysis partitions β into three subsets: two green arcs β' (indistinguishable in this plot) introduced by a ISA 2-break $\tau' = (\alpha', \beta')$; two purple arcs introduced by a ISA 3-break $\tau'' = (\alpha'', \beta'')$; and nineteen red arcs $\beta \setminus (\beta' \cup \beta'')$. The blue arc is introduced by the 3-break τ'' in addition to the two purple arcs, while gray arcs are the remaining adjacencies $\mathcal{B} \setminus (\beta \cup \beta' \cup \beta'')$. We suggest that the chromothripsis event only introduced the red arcs $\beta \setminus (\beta' \cup \beta'')$ affecting chromosomes 4 and 22, while β' and β'' were introduced instead by the ISA multi-breaks τ' and τ''.

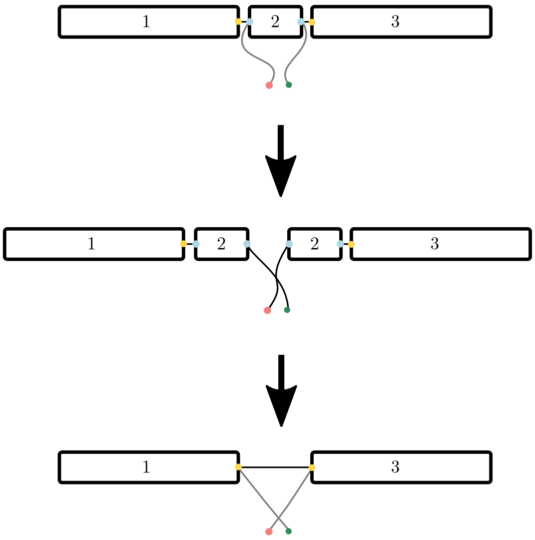

Figure 9 (Top left) A portion of a reference genome (black) and novel adjacencies (gray) retrieved from a JaBbA graph with synteny block 2 being *duplicated by a rearrangement.* We say that a synteny block is duplicated by a rearrangement if it contains less than 1kbp, both of its extremities (blue) are incident to separate novel adjacencies (gray), neither of the adjacent extremities (yellow) is incident to a novel adjacency, and the copy numbers provided in the JaBbA graph of blocks 1 and 3 are lower than the copy number of block 2. (Middle left) We assume that synteny block 2 got duplicated by a rearrangement that resulted in the depicted local organization of a genome. (Bottom left) A portion of the updated reference genome and novel adjacencies that are used in our analysis. (Top right) A portion of a reference genome (black) and novel adjacencies (gray) retrieved from a JaBbA graph with synteny block 2 being *deleted by a rearrangement.* We say that a synteny block is deleted by a rearrangement if it contains less than 100kbp, neither of its extremities (blue) is incident to a novel adjacency, both adjacent extremities (yellow) are incident to separate novel adjacencies (gray), neither of the neighboring synteny blocks (1 and 3) is duplicated by a rearrangement, and the copy numbers provided in the JaBbA graph of blocks 1 and 3 are greater than the copy number of block 2. (Middle right) We assume that synteny block 2 got deleted by a rearrangement that resulted in the depicted local organization of a genome. (Bottom right) A portion of the updated reference genome and novel adjacencies that are used in our analysis.

infinite sites assumption and apply these algorithms to analyze 5 human genomes harboring congenital complex rearrangements and 252 cancer genomes harboring chromothripsis events. For one human genome and 22 cancer genomes we compute multi-break scenarios containing complex rearrangements that affect fewer chromosomes than previously reported.

While multi-breaks have previously been used to model complex genome rearrangements [4, 10], to our knowledge the present work is the first to use the number of chromosomes broken by a rearrangement as a constraint on a rearrangement scenario. For simple rearrangements, Yin et al. [38] briefly mention the problem of prioritizing intra-chromosomal rearrangements (inversions) over inter-chromosomal rearrangements (translocations); however, as far as we are aware, the problem remains open.

We note a number of limitations and directions for future work. First, the time complexities of the Minimum Chromosome Number and the Minimum Chromosome Number of a Complex Rearrangement problems remain unknown. It would be desirable to derive a more efficient algorithm to compute or approximate the minimum chromosome number. Second, our method is sensitive to missing novel adjacencies – if at least one novel adjacency introduced by a multi-break remains unidentified from the sequencing data, then this multi-break is excluded from our analyses. Such missing novel adjacencies are abundant in short-read sequencing data [1, 8]. It would be helpful to further extend our model to address missing and erroneous novel adjacencies present in real data, although such issues will also be reduced from improved identification of novel adjacencies from long-read sequencing data. Third, extending our genome representation and space of allowed rearrangements would yield more realistic reconstructions. In particular, following the standard approach in the genome rearrangement literature, we analyze a haploid representation of the genome. However, the human genome is diploid and assigning novel adjacencies to the correct chromosomal homolog [2] will be useful for analyzing complex rearrangements and counting the distinct homologous chromosomes involved in these rearrangements. Another extension is to incorporate additional events including duplication and loss of chromosomal regions. Finally, our enumeration algorithm could be used to solve other optimization problems for the ISA multi-break scenarios, such as minimizing the number of circular excisions (e.g. from ecDNA [34]) in a ISA multi-break scenario or maximizing the number of inversions.

References

1 Sergey Aganezov, Sara Goodwin, Rachel M Sherman, Fritz J Sedlazeck, Gayatri Arun, Sonam Bhatia, Isac Lee, Melanie Kirsche, Robert Wappel, Melissa Kramer, et al. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome research*, 30(9):1258–1273, 2020.

2 Sergey Aganezov and Benjamin J Raphael. Reconstruction of clone-and haplotype-specific cancer genome karyotypes from bulk tumor samples. *Genome research*, 30(9):1274–1290, 2020.

3 Max A Alekseyev and Pavel A Pevzner. Are there rearrangement hotspots in the human genome? *PLoS Computational Biology*, 3(11):e209, 2007.

4 Max A Alekseyev and Pavel A Pevzner. Multi-break rearrangements and chromosomal evolution. *Theoretical Computer Science*, 395(2-3):193–202, 2008.

5 Sylvan C Baca, Davide Prandi, Michael S Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y MacDonald, Mahmoud Ghandi, et al. Punctuated evolution of prostate cancer genomes. *Cell*, 153(3):666–677, 2013.

6 Vineet Bafna and Pavel A Pevzner. Genome rearrangements and sorting by reversals. *SIAM Journal on Computing*, 25(2):272–289, 1996.

7 Lisui Bao, Xiaoming Zhong, Yang Yang, and Lixing Yang. Mutational signatures of complex genomic rearrangements in human cancer. *bioRxiv*, 2021.

8 Julie M Behr, Xiaotong Yao, Kevin Hadi, Huasong Tian, Aditya Deshpande, Joel Rosiene, Titia de Lange, and Marcin Imielinski. Loose ends in cancer genome structure. *bioRxiv*, 2021.

9 Priscila Biller, Laurent Guéguen, Carole Knibbe, and Eric Tannier. Breaking good: accounting for fragility of genomic regions in rearrangement distance estimation. *Genome Biology and Evolution*, 8(5):1427–1439, 2016.

10 Laurent Bulteau, Guillaume Fertin, Géraldine Jean, and Christian Komusiewicz. Sorting by multi-cut rearrangements. *Algorithms*, 14(6):169, 2021.

11 Ryan L Collins, Harrison Brand, Claire E Redin, Carrie Hanscom, Caroline Antolik, Matthew R Stone, Joseph T Glessner, Tamara Mason, Giulia Pregno, Naghmeh Dorrani, et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome biology*, 18(1):1–21, 2017.

12 Isidro Cortés-Ciriano, Jake June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L Jung, Lixing Yang, Dmitry Gordenin, Leszek J Klimczak, Cheng-Zhong Zhang, David S Pellman, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature genetics*, 52(3):331–341, 2020.

13 Jesper Eisfeldt, Maria Pettersson, Anna Petri, Daniel Nilsson, Lars Feuk, and Anna Lindstrand. Hybrid sequencing resolves two germline ultra-complex chromosomal rearrangements consisting of 137 breakpoint junctions in a single carrier. *Human Genetics*, pages 1–16, 2020.

14 Jesper Eisfeldt, Maria Pettersson, Francesco Vezzi, Josephine Wincent, Max Käller, Joel Gruselius, Daniel Nilsson, Elisabeth Syk Lundberg, Claudia MB Carvalho, and Anna Lindstrand. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. *PLoS genetics*, 15(2):e1007858, 2019.

15 Chris D Greenman, Luca Penso-Dolfin, and Taoyang Wu. The complexity of genome rearrangement combinatorics under the infinite sites model. *Journal of Theoretical Biology*, 501:110335, 2020.

16 Kevin Hadi, Xiaotong Yao, Julie M Behr, Aditya Deshpande, Charalampos Xanthopoulakis, Huasong Tian, Sarah Kudman, Joel Rosiene, Madison Darmofal, Joseph DeRose, et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell*, 183(1):197–210, 2020.

17 Sridhar Hannenhalli and Pavel A Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of IEEE 36th annual foundations of computer science*, pages 581–592. IEEE, 1995.

18 Sridhar Hannenhalli and Pavel A Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM (JACM)*, 46(1):1–27, 1999.

19 Marcus Kinsella, Anand Patel, and Vineet Bafna. The elusive evidence for chromothripsis. *Nucleic acids research*, 42(13):8231–8242, 2014.

20 Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

21 Yeonghun Lee and Hyunju Lee. Integrative reconstruction of cancer genome karyotypes using InfoGenomeR. *Nature communications*, 12(1):1–13, 2021.

22 Mitchell L Leibowitz, Stamatis Papathanasiou, Phillip A Doerfler, Logan J Blaine, Lili Sun, Yu Yao, Cheng-Zhong Zhang, Mitchell J Weiss, and David Pellman. Chromothripsis as an on-target consequence of CRISPR–Cas9 genome editing. *Nature Genetics*, 53(6):895–905, 2021.

23 Yilong Li, Nicola D Roberts, Jeremiah A Wala, Ofer Shapira, Steven E Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O Korbel, James E Haber, et al. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, 2020.

24 Pengfei Liu, Ayelet Erez, Sandesh C Sreenath Nagamani, Shweta U Dhar, Katarzyna E Kołodziejska, Avinash V Dharmadhikari, M Lance Cooper, Joanna Wiszniewska, Feng Zhang, Marjorie A Withers, et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell*, 146(6):889–903, 2011.

25 Jian Ma, Aakrosh Ratan, Brian J Raney, Bernard B Suh, Webb Miller, and David Haussler. The infinite sites model of genome evolution. *Proceedings of the National Academy of Sciences*, 105(38):14254–14261, 2008.

26 Layla Oesper, Simone Dantas, and Benjamin J Raphael. Identifying simultaneous rearrangements in cancer genomes. *Bioinformatics*, 34:346–352, 2018.

27 R Gonzalo Parra, Moritz J Przybilla, Milena Simovic, Hana Susak, Manasi Ratnaparkhe, John KL Wong, Verena Koerber, Philipp Mallm, Martin Sill, Thorsten Kolb, et al. Single cell multi-omics analysis of chromothriptic medulloblastoma highlights genomic and transcriptomic consequences of genome instability. *bioRxiv*, 2021.

28 F Pellestor, JB Gaillard, A Schneider, J Puechberty, and V Gatinois. Chromoanagenesis, the mechanisms of a genomic chaos. In *Seminars in Cell & Developmental Biology*. Elsevier, 2021.

29 Pavel Pevzner and Glenn Tesler. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences*, 100(13):7672–7677, 2003.

30 Morasha Plesser Duvdevani, Maria Pettersson, Jesper Eisfeldt, Ortal Avraham, Judith Dagan, Ayala Frumkin, James R Lupski, Anna Lindstrand, and Tamar Harel. Whole-genome sequencing reveals complex chromosome rearrangement disrupting NIPBL in infant with Cornelia de Lange syndrome. *American Journal of Medical Genetics Part A*, 182(5):1143–1151, 2020.

31 Claire Redin, Harrison Brand, Ryan L Collins, Tammy Kammin, Elyse Mitchell, Jennelle C Hodge, Carrie Hanscom, Vamsee Pillalamarri, Catarina M Seabra, Mary-Alice Abbott, et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature genetics*, 49(1):36–45, 2017.

32 Pijus Simonaitis, Annie Chateau, and Krister Swenson. Weighted minimum-length rearrangement scenarios. In *19th International Workshop on Algorithms in Bioinformatics (WABI)*, pages 13–1, 2019.

33 Philip J Stephens, Chris D Greenman, Beiyuan Fu, Fengtang Yang, Graham R Bignell, Laura J Mudie, Erin D Pleasance, King Wai Lau, David Beare, Lucy A Stebbings, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *cell*, 144(1):27–40, 2011.

34 Kristen M Turner, Viraj Deshpande, Doruk Beyter, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, Karen Arden, Bing Ren, David A Nathanson, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature*, 543(7643):122–125, 2017.

35 Neil T Umbreit, Cheng-Zhong Zhang, Luke D Lynch, Logan J Blaine, Anna M Cheng, Richard Tourdot, Lili Sun, Hannah F Almubarak, Kim Judge, Thomas J Mitchell, et al. Mechanisms generating cancer genome complexity from a single cell division error. *Science*, 368(6488), 2020.

36 Natalia Voronina, John KL Wong, Daniel Hübschmann, Mario Hlevnjak, Sebastian Uhrig, Christoph E Heilig, Peter Horak, Simon Kreutzfeldt, Andreas Mock, Albrecht Stenzinger, et al. The landscape of chromothripsis across adult cancer types. *Nature communications*, 11(1):1–13, 2020.

37 Sophia Yancopoulos, Oliver Attie, and Richard Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.

38 Xiao Yin and Daming Zhu. Sorting genomes by reversals and translocations. In *2009 Asia-Pacific Conference on Information Processing*, volume 2, pages 391–394. IEEE, 2009.

39 Ron Zeira and Ron Shamir. Sorting cancer karyotypes using double-cut-and-joins, duplications and deletions. *Bioinformatics*, 37(11):1489–1496, 2021.

40 Cinthya J Zepeda-Mendoza and Cynthia C Morton. The iceberg under water: unexplored complexity of chromoanagenesis in congenital disorders. *The American Journal of Human Genetics*, 104(4):565–577, 2019.

41 Cheng-Zhong Zhang, Alexander Spektor, Hauke Cornils, Joshua M Francis, Emily K Jackson, Shiwei Liu, Matthew Meyerson, and David Pellman. Chromothripsis from DNA damage in micronuclei. *Nature*, 522(7555):179–184, 2015.

A Proofs

A.1 Lemma 1

▶ **Lemma.** *Suppose that A and B are co-tailed genomes that contain the same synteny blocks, $E(A)$ are the adjacencies of A that are absent from B, and $E(B)$ are the adjacencies of B that are absent from A. The multi-break $\tau_t = (E(A), E(B))$ is the unique multi-break that transforms A into B.*

Proof. If genomes A and B are co-tailed and contain the same synteny blocks, then a block extremity is in some adjacency of A if and only if it is in some adjacency of B, and the same property holds for $E(A)$ and $E(B)$. As $E(A) \cap E(B) = \emptyset$ by construction, we obtain that $\tau_t = (E(A), E(B))$ is a multi-break that transforms A into B.

Now if a multi-break $\tau = (\alpha, \beta)$ transforms A into B, then $E(A) \subseteq \alpha$ and $E(B) \subseteq \beta$ as τ has to break the adjacencies in $E(A)$ and introduce the adjacencies in $E(B)$. What is more, if τ breaks an adjacency $e = \{u, v\}$ then, due to the definition of a multi-break, we obtain that it introduces an adjacency $f \neq e$ of B incident to u. However B contains a single adjacency incident to u, thus B does not contain e, which establishes that $\alpha = E(A)$. As $|\alpha| = |\beta|$ and $|E(A)| = |E(B)|$ we obtain that $\beta = E(B)$ and conclude that $\tau = \tau_t$. ◀

A.2 Theorem 1

▶ **Theorem.** *Let $\{H_1, \ldots, H_m\}$ be the connected components of the chromosomes graph $C(A, B)$. Genomes A_i and B_i consisting respectively of the chromosomes of A and B included in a component H_i are co-tailed and contain the same synteny blocks. Also, the minimum chromosome number $c(A, B)$ is equal to $\max_{i \leq m} c(A_i, B_i)$.*

Proof. Let S_i be the synteny blocks, and let U_i be the telomeres of a genome A included in a connected component H_i of the chromosome graph $C(A, B)$ for $i \leq m$. Due to A and B being co-tailed and containing the same synteny blocks, S_i are the synteny blocks, and U_i are the telomeres of both genomes A_i and B_i, which ensures that A_i and B_i are also co-tailed and contain the same synteny blocks.

Let $\mathbf{T}_i$ be a ISA multi-break scenario for A_i and B_i with the chromosome number $c^*(\mathbf{T}_i) = c(A_i, B_i)$ for $i \leq m$. A sequence $\mathbf{T}$ of multi-breaks obtained by concatenating the scenarios $\{\mathbf{T}_1, \ldots, \mathbf{T}_m\}$ is a ISA multi-break scenario for A and B with a chromosome number equal to $\max_{i \leq m} c(A_i, B_i)$, which establishes inequality $c(A, B) \leq c^*(T) = \max_{i \leq m} c(A_i, B_i)$.

We say that a multi-break $\tau = (\alpha, \beta)$ is *split* if α includes edges from more than one connected component of the chromosome graph $C(A, B)$. A multi-break scenario is *splitless* if it does not contain a split multi-break. Let $c_s(A, B)$ be the minimum chromosome number of a splitless ISA multi-break scenario for A and B. In what follows we show that $c(A, B) = c_s(A, B)$ and $c_s(A, B) \geq \max_{i \leq m} c(A_i, B_i)$.

First, let $\mathbf{T}$ be a ISA multi-break scenario for A and B with $c^*(\mathbf{T}) = c(A, B)$. If $\mathbf{T}$ is splitless, then $c(A, B) \geq c_s(A, B)$. Otherwise, let $\tau = (\alpha, \beta)$ be a split ISA multi-break in $\mathbf{T}$ with α and β including adjacencies from the connected components $\{H_{\sigma(1)}, \ldots, H_\sigma(l)\}$ of the chromosome graph $C(A, B)$, where $\sigma : \{1, \ldots, l\} \to \{1, \ldots, m\}$ is an injection for $l \leq m$. The scenario $\mathbf{T}$ being a ISA multi-break scenario means that α does not include adjacencies introduced by multi-breaks preceding τ in $\mathbf{T}$, and that β does not include adjacencies broken by the multi-breaks proceeding τ in $\mathbf{T}$, which ensures that α and β are respectively adjacencies of A and B, and thus are included among the edges of the chromosome graph $C(A, B)$. Partition the adjacencies α and β into subsets $\{\alpha_1, \ldots, \alpha_l\}$

and $\{\beta_1, \dots, \beta_l\}$ where α_i and β_i respectively are the adjacencies of α and β included in a component $H_{\sigma(i)}$ of the chromosome graph. This way $\tau_i = (\alpha_i, \beta_i)$ a ISA multi-break for A and B that is not split, and the multi-break τ then can be replaced in the scenario $\mathbf{T}$ with a sequence of multi-breaks $((\alpha_1, \beta_1), \dots, (a_l, \beta_l))$ to obtain a ISA multi-break scenario $\mathbf{T}'$ for A and B that has less split multi-breaks than $\mathbf{T}$. Let A' be a genome transformed by τ during the scenario $\mathbf{T}$. By construction, the multi-breaks $\{(\alpha_1, \beta_1), \dots, (\alpha_l, \beta_l)\}$ break disjoint subsets of chromosomes of A', which ensures that the number of chromosomes broken by a multi-break τ during $\mathbf{T}$ is equal to the sum of the numbers of chromosomes broken by the multi-breaks $\{(\alpha_1, \beta_1), \dots, (\alpha_l, \beta_l)\}$ during $\mathbf{T}'$, and thus $c^*(\mathbf{T}') \leq c^*(\mathbf{T})$. We proceed until a splitless ISA multi-break scenario for A and B with a chromosome number smaller than or equal to $c^*(\mathbf{T})$ is obtained, thus establishing that $c(A, B) = c^*(\mathbf{T}) \geq c_s(A, B)$. A splitless ISA multi-break scenario is also a ISA multi-break scenario, thus we have that $c(A, B) \leq c_s(A, B)$, and conclude that $c(A, B) = c_s(A, B)$.

Finally, let $\mathbf{T}$ be a splitless ISA multi-break scenario for A and B with $c^*(\mathbf{T}) = c_s(A, B)$, and let $\mathbf{T}_i$ be a subsequence of $\mathbf{T}$ that consists of the multi-breaks that break adjacencies in A_i. The subsequence $\mathbf{T}_i$ is a ISA multi-break scenario for A_i and B_i, and every subset of chromosomes broken by a multi-break in $\mathbf{T}_i$ is also broken by the same multi-break in $\mathbf{T}$, which ensures that $c^*(\mathbf{T}) \geq c^*(\mathbf{T}_i)$. Thus we obtain an inequality $c_s(A, B) = c^*(\mathbf{T}) \geq \max_{i \leq m} c^*(\mathbf{T}_i) \geq \max_{i \leq m} c(A_i, B_i)$, which allows us to conclude that $c(A, B) = \max_{i \leq m} c(A_i, B_i)$. ◀

A.3 Lemma 2

▶ **Lemma.** *For genomes A and B, the function $\mathbf{T} : \mathcal{P}(A, B) \to \mathcal{I}(A, B)$ between the set $\mathcal{P}(A, B)$ of the ordered partitions of the cycles of the breakpoint graph $G(A, B)$ and the set $\mathcal{I}(A, B)$ of the ISA multi-break scenarios for A and B is a bijection.*

Proof. Let $\mathbf{P} = (P_1, \dots, P_l)$ be an ordered partition of the connected components of the breakpoint graph $G(A, B)$, and let α_i and β_i be respectively black and gray edges in P_i for $i \in \{1, \dots, l\}$. We start by showing that $\tau(P_i) = (\alpha_i, \beta_i)$ is a multi-break. By construction of the breakpoint graph we have that $\alpha_i \cap \beta_i \subseteq E(A) \cap E(B) = \emptyset$, where $E(A)$ and $E(B)$ are respectively the adjacencies of A that are absent from B and the adjacencies of B that are absent from A. Every vertex in P_i is incident to one black and one gray edge. This ensures that for $\alpha_i = \big\{\{u_1, u_2\}, \dots, \{u_{2k-1}, u_{2k}\}\big\}$ there exists a permutation σ of a set $\{1, \dots, 2k\}$ such that $\beta_i = \big\{\{u_{\sigma(1)}, u_{\sigma(2)}\}, \dots, \{u_{\sigma(2k-1)}, u_{\sigma(2k)}\}\big\}$, which means that $\tau(P_i) = (\alpha_i, \beta_i)$ is a multi-break.

We proceed by showing that $\mathbf{T}(\mathbf{P}) = ((\alpha_1, \beta_1), \dots, (\alpha_l, \beta_l))$ is a ISA multi-break scenario for A and B. The set $\mathbf{P}$ is an ordered partition of the connected components of $G(A, B)$, thus $\{\alpha_1, \dots, \alpha_l\}$ and $\{\beta_1, \dots, \beta_l\}$ respectively partitions $E(A)$ and $E(B)$. This already ensures that $\mathbf{T}(\mathbf{P}) = ((\alpha_1, \beta_1), \dots, (\alpha_l, \beta_l))$ is a multi-break scenario for A and B. In order to establish that $\mathbf{T}(\mathbf{P})$ is a ISA multi-break scenario for A and B we have to show that an adjacency joined by a multi-break in $\mathbf{T}(\mathbf{P})$ is not broken by a subsequent multi-break in $\mathbf{T}(\mathbf{P})$; i.e., that for $1 \leq i < j \leq l$ we have $\beta_i \cap \alpha_j = \emptyset$. Let $i < j$ be two elements from $\{1, \dots, l\}$, and let $e = \{u, v\}$ be an adjacency in α_j. Due to (α_j, β_j) being a multi-break, there exists an adjacency $f \neq e$ in β_j that includes u. By construction, we have that $\beta_j \subseteq E(B)$, which means that f is an adjacency of B and, by definition of a genome, f is the single adjacency of B that includes u. As $\{\beta_1, \dots, \beta_l\}$ partitions $E(B)$ and $i \neq j$, we conclude that β_i does not contain an adjacency that includes u, thus β_i does not include $e = \{u, v\}$. This way we obtain that $\beta_i \cap \alpha_j = \emptyset$, and conclude that $\mathbf{T}(\mathbf{P})$ is a ISA multi-break scenario for A and B.

The function $P \mapsto \tau(P)$ is an injection between the subsets of the connected components of the breakpoint graph and multi-breaks. Let $\mathbf{P} = (P_1, \dots, P_l)$ and $\mathbf{P'} = (P'_1, \dots, P'_m)$ be two ordered partitions of the connected components of the breakpoint graph with $\mathbf{T}(\mathbf{P}) = (\tau(P_1), \dots, \tau(P_l)) = (\tau(P'_1), \dots, \tau(P'_m)) = \mathbf{T}(\mathbf{P'})$. From the equality $\mathbf{T}(\mathbf{P}) = \mathbf{T}(\mathbf{P'})$ and the injectivity of $P \mapsto \tau(P)$ we obtain that $l = m$ and that $P_i = P'_i$ for $i \in \{1, \dots, l\}$. This way we obtain that $\mathbf{P} = \mathbf{P'}$, and conclude that $\mathbf{P} \mapsto \mathbf{T}(\mathbf{P})$ is an injection between the ordered partitions of the connected components of the breakpoint graph $G(A, B)$ and the ISA multi-break scenarios for A and B.

It remains to show that $\mathbf{P} \mapsto \mathbf{T}(\mathbf{P})$ is a surjection between the ordered partitions of the connected components of the breakpoint graph $G(A, B)$ and the ISA multi-break scenarios for A and B. Let $\mathbf{T} = ((\alpha_1, \beta_1), \dots, (\alpha_l, \beta_l))$ be a ISA multi-break scenario for A and B. We start by showing that $\cup_{i=1}^{l} \alpha_i \subseteq E(A)$ and $\cup_{i=1}^{l} \beta_i \subseteq E(B)$. Let i be an element in $\{1, \dots, l\}$, and let $e = \{u, v\}$ be an adjacency in α_i. Due to $\mathbf{T}$ being a ISA multi-break scenario for A and B, the adjacency e is not joined by any of the multi-breaks preceding (α_i, β_i) in $\mathbf{T}$, which ensures that e is an adjacency of A. In what follows we show that e is not an adjacency of B, and conclude that $e \in E(A)$. Due to (α_i, β_i) being a multi-break, there exists an adjacency $f \neq e$ in β_i that includes u. Due to $\mathbf{T}$ being a ISA multi-break scenario for A and B, the adjacency f is an adjacency of B, as it is not broken by any of the multi-breaks that follow (α_i, β_i) in $\mathbf{T}$. By definition of a genome, f is the single adjacency of B that includes u, which means that e is not an adjacency of B. This way we obtain that $e \in E(A)$, and conclude that $\cup_{i=1}^{l} \alpha_i \subseteq E(A)$. Now, let $e = \{u, v\}$ be an adjacency in β_i. Due to (α_i, β_i) being a multi-break, there exists an adjacency $f \neq e$ in α_i that includes u. We have already shown that $f \in E(A)$, and, as f is the single adjacency of A that includes u, we obtain that $e \in E(B)$, and conclude that $\cup_{i=1}^{l} \beta_i \subseteq E(B)$.

We proceed by showing that $\{\alpha_1, \dots, \alpha_l\}$ and $\{\beta_1, \dots, \beta_l\}$ partitions $E(A)$ and $E(B)$. Let e and f be respectively adjacencies in $E(A)$ and $E(B)$. By definition we have that $E(A) \cap E(B) = \emptyset$, thus every adjacency in $E(A)$ must be broken by a multi-break in $\mathbf{T}$, and every adjacency in $E(B)$ must be joined by a multi-break in $\mathbf{T}$. This ensures that there exist $i, j \in \{1, \dots, l\}$ such that $e \in \alpha_i$ and $f \in \beta_i$. By definition of a genome, the adjacencies of A contain a single copy of e and the adjacencies of B contain a single copy of f. What is more, $\{e\} \cap \cup_{i=1}^{l} \beta_i \subseteq E(A) \cap E(B) = \emptyset$, thus e is not joined by any of the multi-breaks in $\mathbf{T}$, which ensures that it is only broken by (α_i, β_i) in $\mathbf{T}$. Similarly, f is not broken by any of the multi-breaks in $\mathbf{T}$, which ensures that it is only joined by (α_j, β_j) in $\mathbf{T}$. We conclude that $\{\alpha_1, \dots, \alpha_l\}$ and $\{\beta_1, \dots, \beta_l\}$ respectively partitions $E(A)$ and $E(B)$.

Let P_i be a subgraph of $G(A, B)$ induced by adjacencies α_i and β_i for $i \in \{1, \dots, l\}$. We will show that P_i is a subset of the connected components of $G(A, B)$. We do this by establishing that for every vertex u in P_i all the edges incident to u in $G(A, B)$ are also present in P_i. A vertex u of P_i is incident to one black and one gray edge in $G(A, B)$. By construction of P_i, u is in some adjacency in $\alpha_i \cup \beta_i$, however due to (α_i, β_i) being a multi-break, we obtain that u is both in some adjacency in α_i and in some adjacency in β_i. This ensures that P_i contains one black and one gray edge incident to u, and thus that all the edges incident to u in $G(A, B)$ are also present in P_i. This way we conclude that P_i is a subset of the connected components of $G(A, B)$.

Finally, due to $\{\alpha_1, \dots, \alpha_l\}$ and $\{\beta_1, \dots, \beta_l\}$ respectively partitioning $E(A)$ and $E(B)$, we obtain that $\mathbf{P} = (P_1, \dots, P_l)$ is an ordered partition of the connected components of the breakpoint graph $G(A, B)$. By construction of $\mathbf{P}$, we have that $\mathbf{T}(\mathbf{P}) = \mathbf{T} = ((\alpha_1, \beta_1), \dots, (\alpha_l, \beta_l))$, which ensures that $\mathbf{P} \mapsto \mathbf{T}(\mathbf{P})$ is a surjection between the ordered partitions of the connected components of the breakpoint graph $G(A, B)$ and the ISA multi-break scenarios for A and B. ◀

A.4 Theorem 2

▶ **Theorem.** *Let $\{H_1, \ldots, H_m\}$ be the connected components of the chromosomes graph $C(A, \mathcal{B})$. If A_i is a genome consisting of the chromosomes of A included in H_i, and $\mathcal{B}_i$ are the adjacencies in $\mathcal{B}$ included in H_i, then the chromosome number $c(A, \mathcal{B}, \beta)$ is equal to $\Sigma_{i=1}^{m} c(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$.*

Proof. Let A' be a ISA intermediate genome for A and $\mathcal{B}$ in which β affects $c(A, \mathcal{B}, \beta)$ chromosomes. A genome A'_i that consists of the chromosomes of A' included in a connected component H_i of the chromosome graph $C(A, \mathcal{B})$ is a ISA intermediate genome for A_i and $\mathcal{B}_i$. What is more, $c(A, \mathcal{B}, \beta) = \Sigma_{i=1}^{m} c_i$ where c_i is the number of chromosomes affected in A'_i by β. By construction of the chromosome graph, if an adjacency $e \in \beta$ affects a chromosome in A_i, then $e \in \mathcal{B}_i \cap \beta$, which ensures that the numbers of chromosomes affected by $\mathcal{B}_i \cap \beta$ in A'_i is equal to c_i. This allows us to conclude that $c(A, \mathcal{B}, \beta) \geq \Sigma_{i=1}^{m} c(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$.

Let A'_i be a ISA intermediate genome for A_i and $\mathcal{B}_i$ in which $\mathcal{B}_i \cap \beta$ affects $c(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$ chromosomes for $i \leq m$. A genome A' that consists of the union of the chromosomes of the genomes $\{A'_1, \ldots, A'_m\}$ is a ISA intermediate genome for A and $\mathcal{B}$. Whats is more, β affects $\Sigma_{i=1}^{m} c(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$ chromosomes in A', thus we have that $c(A, \mathcal{B}, \beta) \leq \Sigma_{i=1}^{m} c(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$, and finally conclude that $c(A, \mathcal{B}, \beta) = \Sigma_{i=1}^{m} c(A_i, \mathcal{B}_i, \mathcal{B}_i \cap \beta)$. ◀