Linearization of Median Genomes Under the Double-Cut-and-Join-Indel Model

Evolutionary Bioinformatics
Volume 15: 1–12
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934318820534
SAGE

Pavel Avdeyev[1], Shuai Jiang[2] and Max A Alekseyev[1]

[1]Computational Biology Institute, The George Washington University, Washington, DC, USA. [2]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA.

ABSTRACT: Reconstruction of the median genome consisting of linear chromosomes from three given genomes is known to be intractable. There exist efficient methods for solving a relaxed version of this problem, where the median genome is allowed to have circular chromosomes. We propose a method for construction of an approximate solution to the original problem from a solution to the relaxed problem and prove a bound on its approximation error. Our method also provides insights into the combinatorial structure of genome transformations with respect to appearance of circular chromosomes.

KEYWORDS: Double cut and join, indels, genome median problem, circular chromosome

RECEIVED: November 16, 2018. **ACCEPTED:** November 27, 2018.

TYPE: Algorithm development for evolutionary biological computation - Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science Foundation under the grant no. IIS-1462107.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Max A Alekseyev, Computational Biology Institute, The George Washington University, 45085 University Dr., Suite 305, Ashburn, VA 20147, USA. Email: maxal@gwu.edu

Introduction

In the course of evolution, genomes become a subject to a number of large-scale evolutionary events such as genome rearrangements that shuffle genomic architectures, and gene insertions and deletions (indels) that insert or remove continuous intervals of genes. Since these evolutionary events are rare, the number of them between two genomes is used in phylogenomic studies to measure the evolutionary distance between them. Such measurement is often based on the *maximum parsimony assumption*, implying that the evolutionary distance can be estimated as the minimum number of events between genomes. A convenient model for the most common genome rearrangements is given by the *double-cut-and-join* (DCJ) operations,[1] also known as *2-breaks*,[2] which make two “cuts” in a genome and “glues” the resulting genomic fragments in a new order. Namely, DCJs mimic *reversals* (that inverse contiguous segments of chromosomes), *translocations* (that exchange tails of the two chromosomes), and *fissions/fusions* (that split/join chromosomes), while indels can be modeled by the DCJs on certain artificial circular chromosomes called *prosthetic*.[3,4]

The maximum parsimony assumption enables addressing the *ancestral genome reconstruction problem*, which asks to reconstruct ancestral genomes from given extant genomes, by minimizing the total distance between genomes along the branches of the phylogenetic tree. The basic case of this problem with just three given genomes is known as the *genome median problem* (GMP), which asks for a single ancestral genome (*median genome*) at the minimum total distance from the given genomes.

The GMP is NP-hard under a number of models of genome rearrangements, such as reversals-only[5] and DCJ.[6] While these problems can be posed for both *circular* genomes (consisting of circular chromosomes) and *linear* genomes (consisting of linear chromosomes), the DCJ model allows appearance of circular chromosomes in transformations between linear genomes. Correspondingly, a solution to the GMP under the DCJ model may contain circular chromosomes even if the given genomes are linear. Since appearance of circular chromosomes in the reconstructed ancestral genomes of extant linear genomes represents an artifact and inadequately describes the biological reality, it is important to distinguish between the GMP and the *linear genome median problem* (L-GMP), where the latter is restricted to linear genomes only.

To the best of our knowledge, there exist no solvers for the L-GMP, while there are some advanced GMP solvers,[7–9] which allow the median genome to contain circular chromosomes. This deficiency inspired us to pose the problem of using the solution for the GMP to obtain a linear genome approximating the solution to the L-GPM. In this study, we propose an algorithm that linearizes chromosomes of a given GMP solution in a certain optimal way as described in the “Background” section. Our approach also provides insights into the combinatorial structure of genome transformations by DCJs and indels with respect to appearance of circular chromosomes. We remark that a similar *linearization problem* appears in adjacency-based reconstructions of median genomes and is known to be intractable,[10] forcing the existing approaches[10–14] to solve its relaxation and allowing the constructed median genomes to contain circular chromosomes.

The article is organized as follows. In the “Background” section, we describe the graph-theoretical representation of genomes, DCJs, and indels. In the “Main Results” section, we formulate main theorems providing an approximate solution for L-GMP. In the “Methods” section, we develop necessary machienery and prove our main theorems. We conclude the article with the “Discussion” section.

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

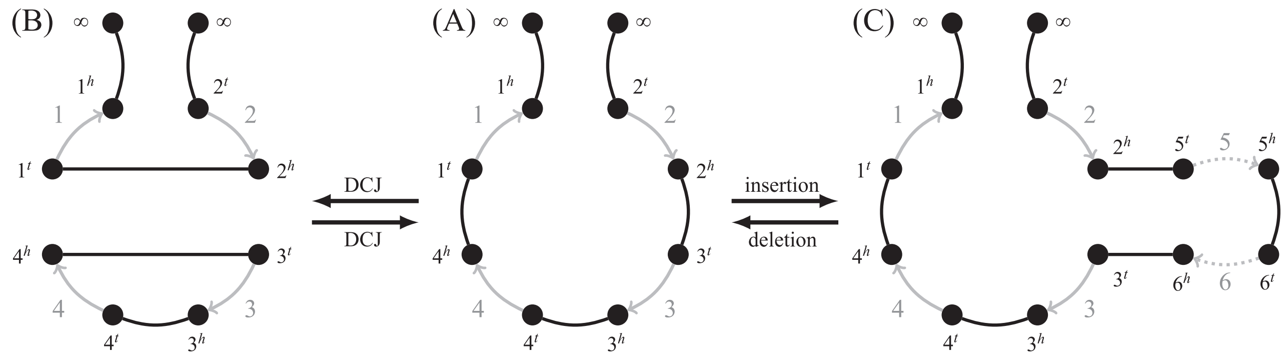

Figure 1. (A) A genome graph for a linear genome $(+2+3+4+1)$. (B) A genome graph for a genome $(+2+1)\{+3+4\}$ consisting of circular and linear chromosomes is obtained by a DCJ that splits the linear chromosome into two chromosomes. (C) A genome graph for a genome $(+2+5+6+3+4+1)$ is obtained by an insertion of gene sequence $+5$, $+6$. Dotted directed edges correspond to inserted genes.

Background

DCJ-Indel distance and genome graphs

In this study, we focus on genomes with no duplicated genes. Let P be a genome, which may contain both circular and linear chromosomes. We represent a circular chromosome consisting of n genes as a graph cycle with n directed *gene edges* encoding genes and their strands, which alternate with n undirected edges connecting the extremities of adjacent genes. Similarly, we represent a linear chromosome consisting of n genes as a path with n directed gene edges alternating with $n+1$ undirected edges, where $n-1$ undirected edges connect extremities of adjacent genes, and two more undirected edges connect each endpoint extremity to its own special vertex labeled ∞ corresponding to a telomere (Figure 1A). We label each gene edge with the corresponding gene x and further label its tail and head endpoints with x^t and x^h, respectively (Figure 1A). We define the operation $\bar{\cdot}$ as $\overline{x^t} = x^h$ and $\overline{x^h} = x^t$. A collection of cycles and paths representing the chromosomes of P forms the *genome graph* $\mathfrak{G}(P)$. The undirected edges in $\mathfrak{G}(P)$ are called *P-edges*. We denote by $S(P)$ the *gene content* of P (ie, the set of genes present in P) and by $V(P)$ the set of regular (non-∞) vertices of $\mathfrak{G}(P)$.

A DCJ transforming a genome P into a genome P' corresponds to one of the following operations transforming $\mathfrak{G}(P)$ into $\mathfrak{G}(P')$ (Figure 1A and B):

1. $\{x, y\}, \{u, v\} \rightarrow \{x, u\}, \{y, v\}$ (internal reversals, translocations)
2. $\{x, y\}, \{u, \infty\} \rightarrow \{x, u\}, \{y, \infty\}$ (reversals at chromosome ends, translocations involving a whole chromosome)
3. $\{x, \infty\}, \{y, \infty\} \rightarrow \{x, y\}$ (fusions)
4. $\{x, y\} \rightarrow \{x, \infty\}, \{y, \infty\}$ (fissions)

where x, y, u, and v are distinct vertices from $V(P)$.

A *DCJ scenario* between genomes P and Q with equal gene content (ie, $S(P) = S(Q)$) is a sequence of DCJs transforming P into Q. We define the *DCJ distance* $d_{\text{DCJ}}(P, Q)$ between genomes P and Q as the length of a shortest DCJ scenario between them.

To transform a genome P into a genome Q with unequal gene content, one needs to consider gene insertion and deletion operations (*indels*) in addition to DCJs. An insertion transforming a genome P into a genome P' corresponds to one of the following operations transforming $\mathfrak{G}(P)$ into $\mathfrak{G}(P')$ (Figure 1A and C):

(i) replace a P-edge $\{x, y\}$ with a path $(x, u_1, \bar{u}_1, u_2, \ldots, \bar{u}_l, y)$ (including the case of either $x = \infty$ or $y = \infty$),
(ii) add a path $(\infty, u_1, \bar{u}_1, u_2, \ldots, \bar{u}_l, \infty)$,
(iii) add a cycle $(u_1, \bar{u}_1, u_2, \ldots, \bar{u}_l, u_1)$,

where the edges alternate between P'-edges $\{\bar{u}_i, u_{i+1}\}$ and gene edges $(u_i, \bar{u}_i)$ with $u_i \notin S(P)$, resulting in $S(P') = S(P) \cup \{u_1, \ldots, u_l\}$.

A deletion can be viewed as an event reversing an insertion. A deletion transforming a genome P into a genome P' corresponds to one of the following operations transforming $\mathfrak{G}(P)$ into $\mathfrak{G}(P')$ (Figure 1A and C):

(i) replace a path $(x, u_1, \bar{u}_1, u_2, \ldots u_l, \bar{u}_l, y)$ with a P'-edge $\{x, y\}$ (including the case of either $x = \infty$ or $y = \infty$),
(ii) remove a path $(\infty, u_1, \bar{u}_1, u_2, \ldots, u_l, \bar{u}_l, \infty)$,
(iii) remove a cycle $(u_1, \bar{u}_1, u_2, \ldots, u_l, \bar{u}_l, u_1)$,

where the edges alternate between P-edges $\{\bar{u}_i, u_{i+1}\}$ and gene edges $(u_i, \bar{u}_i)$, resulting in $S(P') = S(P) \setminus \{u_1, \ldots, u_l\}$.

A *DCJ-Indel scenario* t between genomes P and Q is a sequence of DCJs and indels transforming P into Q, where deletions delete genes from $S(P) \setminus S(Q)$ and insertions insert genes from $S(Q) \setminus S(P)$ (ie, no gene can be inserted and then deleted, or deleted and then inserted), denoted as $t : P \rightarrow Q$. We also find it convenient to represent t as $P = P_0 \xrightarrow{\vartheta_1} P_1 \xrightarrow{\vartheta_2} \cdots \rightarrow P_{n-1} \xrightarrow{\vartheta_n} P_n = Q$, where each ϑ_i is a

DCJ or an indel. We define the *DCJ-Indel distance* $d_{\mathrm{DI}}(P, Q)$ as the length of a shortest DCJ-Indel scenario transforming genome P into genome Q. It is easy to see that any DCJ-Indel scenario transforming P into Q can be reversed (turning each insersion into a deletion, and vice versa) to obtain a DCJ-Indel scenario transforming Q into P, implying that $d_{\mathrm{DI}}(P, Q) = d_{\mathrm{DI}}(Q, P)$.

A circular chromosome C in P is a *singleton* with respect to genome Q if it is composed of genes absent in Q, ie, $S(C) \cap S(Q) = \varnothing$. Let $\mathrm{sng}_Q(P)$ be the number of singletons in P with respect to Q. The total number of singletons in P and Q with respect to each other is $\mathrm{sng}(P, Q) := \mathrm{sng}_P(Q) + \mathrm{sng}_Q(P)$. The following lemma describes an important property of singletons.

Lemma 1 (Compeau[4])
For given genomes P and Q, let C be a singleton in P with respect to Q and P^0 be the genome obtained from P by removing C. Then $d_{\mathrm{DI}}(P^0, Q) = d_{\mathrm{DI}}(P, Q) - 1$.

From Lemma 1, the DCJ-Indel distance between two genomes can be computed with the following formula:[4]

$$d_{\mathrm{DI}}(P, Q) \quad = \mathrm{sng}(P, Q) + d_{\mathrm{DI}}({}^{Q}P, {}^{P}Q)\,, \qquad (1)$$

where ${}^{Q}P$ and ${}^{P}Q$ are obtained from P and Q by removing all singletons (ie, $\mathrm{sng}({}^{Q}P, {}^{P}Q) = 0$). We need the following lemma.

Lemma 2 (Compeau[4])
Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_n} P_n$ be a shortest DCJ-Indel scenario. Let C be a singleton in P_o with respect to P_n. Then for any $i \in \{0, \ldots, n\}$ and any chromosome D in P_i such that $S(C) \cap S(D) \neq \varnothing$, we have $S(D) \cap S(P_0) = \varnothing$.

Genome median problem

We pose the GMP under the DCJ-Indel model as follows.

Genome median problem (GMP)
Given genomes B_1, B_2, and B_3, *find a genome M with $S(M) \subseteq S(B_1) \cup S(B_2) \cup S(B_3)$ that minimizes the DCJ-Indel median score:*

$$ms_{\mathrm{DI}}(M, B_1, B_2, B_3) := \sum_{i=1}^{3} d_{\mathrm{DI}}(B_i, M).$$

Since the GMP is posed under the DCJ-Indel model, a median genome for given linear genomes may contain circular chromosomes. To address the issue of circular chromosome presence, we pose the following problem.

Linear genome median problem (L-GMP)
For given linear genomes B_1, B_2, and B_3, find a linear genome M with $S(M) \subseteq S(B_1) \cup S(B_2) \cup S(B_3)$ minimizing the DCJ-Indel median score $ms_{\mathrm{DI}}(M, B_1, B_2, B_3)$.

While we are not aware of efficient algorithms (let alone, software solvers) for solving the L-GMP, we pose the problem of constructing an approximate solution for the L-GMP from the given solution for GMP.

Results

Chromosome linearization

Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario and C be a circular chromosome in P_0. For each $i \in \{0, 1, \ldots, n\}$, let $\mathfrak{C}_i = \{C_i^1, \ldots, C_i^{m_i}\}$ be a collection of all circular chromosomes in P_i such that $S(C_i^l) \cap S(C) \neq \varnothing$ $(l \in \{1, \ldots, m_i\})$. We call $\mathfrak{C}_i$ a *meta-chromosome* of C in P_i and note that $\mathfrak{C}_i$ itself may be viewed as a genome, for which $S(\mathfrak{C}_i)$, $\mathfrak{S}(\mathfrak{C}_i)$, and $V(\mathfrak{C}_i)$ are defined. In particular, we have $S(\mathfrak{C}_i) = \bigcup\limits_{l=1}^{m_i} S(C_i^l)$. Below, we describe an important property of circular chromosomes appearing in DCJ-Indel scenarios (Figure 3).

Definition 1
A circular chromosome C is linearized within a DCJ-Indel scenario $t : P \to Q$ (or t linearizes C) if the following three conditions hold:

(E1) C is present in P;

(E2) $S(C) \cap S(Q) \neq \varnothing$;

(E3) $S(\mathfrak{C}) \cap S(P) \neq S(C) \cap S(Q)$, *where $\mathfrak{C}$ is the meta-circular chromosome of C in Q.*

Equivalently, a circular chromosome C of genome P is linearized within $t : P \to Q$ if there exists a gene in C that resides on a linear chromosome in Q, or together with a gene from another chromosome of P resides on a circular chromosome in Q.

We extend Definition 1 to a particular event in a DCJ-Indel scenario as follows.

Definition 2
Let $t : P = P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_n} P_n = Q$ be a DCJ-Indel scenario that linearizes a circular chromosome C. We say that an event ϑ_i linearizes C within t if C is linearized within $(\vartheta_1, \ldots, \vartheta_i)$ and C is not linearized within $(\vartheta_1, \ldots, \vartheta_k)$ for any $k < i$.

The following theorem shows that for given linear genomes, all circular chromosomes in their median genome are linearized within the corresponding DCJ-Indel scenarios.

Lemma 3
Let B_1, B_2, and B_3 be linear genomes, and M be a genome such that $S(M) \subseteq S(B_1) \cup S(B_2) \cup S(B_3)$. Let t_i be a shortest DCJ-Indel scenario between M and B_i for $i \in \{1, 2, 3\}$. Then each circular chromosome in M is linearized in at least one of the DCJ-Indel scenarios t_1, t_2, t_3.

Proof
Assume that there is a circular chromosome C in M that is not linearized in either of t_1, t_2, t_3. Then at least one of conditions

(E2) or (E3) does not hold for each $Q \in \{B_1, B_2, B_3\}$. Since $S(M) \subseteq S(B_1) \cup S(B_2) \cup S(B_3)$, for each circular chromosome C' in M, we have $S(C') \subseteq S(B_1) \cup S(B_2) \cup S(B_3)$. Hence, the condition (E2) must hold for at least one $Q \in \{B_1, B_2, B_3\}$. So, there is $l \in \{1, 2, 3\}$ such that the condition (E3) does not hold for the genome B_l (ie, $S(\mathcal{C}) \cap S(M) = S(C) \cap S(B_l)$), where $\mathcal{C}$ is the meta-chromosome of C in B_l. In other words, there exist circular chromosomes $\mathcal{C} = \{C'_1, \ldots, C'_k\}$ in the genome B_l, which contradicts its linearity. □

The following theorems represent a key to proving our main results on linearization of median genomes. Proofs of these theorems are rather technical and given in the "Methods" section.

Theorem 1
Let $t : P \to Q$ be a DCJ-Indel scenario that linearizes a circular chromosome C. Then there exists a DCJ-Indel scenario $\tilde{t} : P \xrightarrow{r} P' \xrightarrow{\tilde{t}'} Q$ such that r is a DCJ linearizing C within $\tilde{t}$ and $|\tilde{t}'| \leqslant |t|$.

For DCJ scenarios, we have a somewhat stronger result.

Theorem 2
Let $t : P \to Q$ be a DCJ scenario that linearizes a circular chromosome C. Then there exists a DCJ scenario $\tilde{t} : P \xrightarrow{r} P' \xrightarrow{\tilde{t}'} Q$ such that r is a DCJ linearizing C within $\tilde{t}$ and $|\tilde{t}'| = |t| - 1$.

Linearization of median genomes

For a genome P, let cchr(P) be the number of circular chromosomes in genome P. Our main results on linearization of median genomes are given by the following theorems.

Theorem 3
Let B_1, B_2, and B_3 be linear genomes, and M be a given median genome. Then for any $n \leqslant \mathrm{cchr}(M)$, there exists a genome $\hat{M}$ such that $\mathrm{cchr}(\hat{M}) = \mathrm{cchr}(M) - n$, $S(M) = S(\hat{M})$, and

$$ms_{\mathrm{DI}}(\hat{M}, B_1, B_2, B_3) - ms_{\mathrm{DI}}(M, B_1, B_2, B_3) \leqslant 2n.$$

Proof
We prove the theorem by induction on n. If $n = 0$, the theorem trivially holds for $\hat{M} = M$.

We assume that the theorem holds for $n < \mathrm{cchr}(M)$. Then there exists a genome M' such that $\mathrm{cchr}(M') = \mathrm{cchr}(M) - n$, $S(M) = S(M')$, and $ms_{\mathrm{DI}}(M', B_1, B_2, B_3) - ms_{\mathrm{DI}}(M, B_1, B_2, B_3) \leqslant 2n$. Let C' be a circular chromosome in M'. Since $S(M') = S(M) \subseteq S(B_1) \cup S(B_2) \cup S(B_3)$, we have $S(C') \subseteq S(B_1) \cup S(B_2) \cup S(B_3)$. Let t'_i be a shortest DCJ-Indel scenario between M' and B_i for $i \in \{1, 2, 3\}$ (Figure 2). By Lemma 3, there is at least one of the DCJ-Indel scenarios t'_1, t'_2, and t'_3 that linearizes C', say t'_1. By Theorem 1, we obtain a DCJ-Indel scenario $\tilde{t}_1$ of the form $M' \xrightarrow{\vartheta} \hat{M} \xrightarrow{\tilde{t}'_1} B_1$ such that ϑ linearizes C' within $\tilde{t}_1$ and $|\tilde{t}'_1| \leqslant |t'_1|$. Clearly, $d_{\mathrm{DI}}(M', B_1) = |\tilde{t}'_1| \leqslant |t'_1|$. By the triangle inequality, for $i = 2, 3$, we have $d_{\mathrm{DI}}(\hat{M}, B_i) \leqslant d_{\mathrm{DI}}(\hat{M}, M') + d_{\mathrm{DI}}(M', B_i) = 1 + |t_{i'}|$. Hence, we have $ms_{\mathrm{DI}}(\hat{M}, B_1, B_2, B_3) - ms_{\mathrm{DI}}(M', B_1, B_2, B_3) \leqslant 2$. Thus, we have

$$\begin{aligned} & ms_{\mathrm{DI}}(\hat{M}, B_1, B_2, B_3) - ms_{\mathrm{DI}}(M, B_1, B_2, B_3) \\ & = ms_{\mathrm{DI}}(\hat{M}, B_1, B_2, B_3) - ms_{\mathrm{DI}}(M', B_1, B_2, B_3) \\ & + ms_{\mathrm{DI}}(M', B_1, B_2, B_3) - ms_{\mathrm{DI}}(M, B_1, B_2, B_3) \\ & \leqslant 2 + 2n = 2 \cdot (n + 1). \end{aligned}$$ □

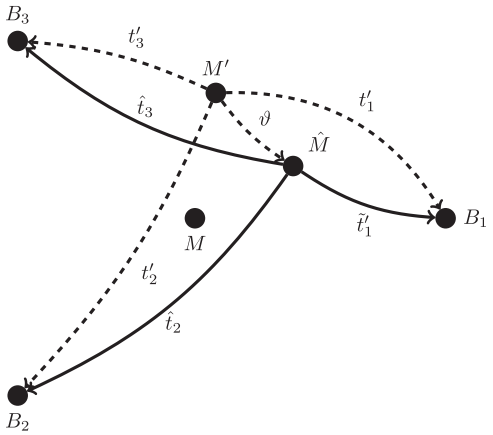

Figure 2. Linear genomes B_1, B_2, and B_3 and their median genome M represented as vertices. A genome M' containing $\mathrm{cchr}(M) - n$ ($n < \mathrm{cchr}(M)$) circular chromosomes is represented as vertex, and the corresponding shortest transformations t'_1, t'_2, and t'_3 are represented as directed dashed edges. We construct a shortest transformation from M' to B_1 composed of ϑ and $\tilde{t}'_1$ such that ϑ results in a genome $\hat{M}$ with $\mathrm{cchr}(\hat{M}) = \mathrm{cchr}(M') - 1$ and $|\tilde{t}'_1| \leqslant |t'_1|$. The corresponding shortest transformations from $\hat{M}$ to B_2 and B_3 are represented as bold directed edges and denoted by $\hat{t}_2$ and $\hat{t}_3$.

For the GMP under the DCJ model, we can immediately improve the derived upper bound as follows.

Theorem 4
Let B_1, B_2, and B_3 be linear genomes with equal gene content, and M be a given median genome. Then for any $n \leqslant \mathrm{cchr}(M)$, there exists a genome $\hat{M}$ such that $\mathrm{cchr}(\hat{M}) = \mathrm{cchr}(M) - n$, and

$$ms_{\mathrm{DCJ}}(\hat{M}, B_1, B_2, B_3) - ms_{\mathrm{DCJ}}(M, B_1, B_2, B_3) \leqslant n.$$

Proof
The proof proceeds as the proof of Theorem 3 with the following difference. We use Theorem 2 instead of Theorem 1 to obtain a DCJ scenario $\tilde{t}_1$ of the form $M' \xrightarrow{\vartheta} \hat{M} \xrightarrow{\tilde{t}'_1} B_1$ such that ϑ linearizes C' within $\tilde{t}_1$ and $|\tilde{t}'_1| = |t'_1| - 1$. Hence, we have $ms_{\mathrm{DCJ}}(\hat{M}, B_1, B_2, B_3) - ms_{\mathrm{DCJ}}(M', B_1, B_2, B_3) \leqslant 1$. □

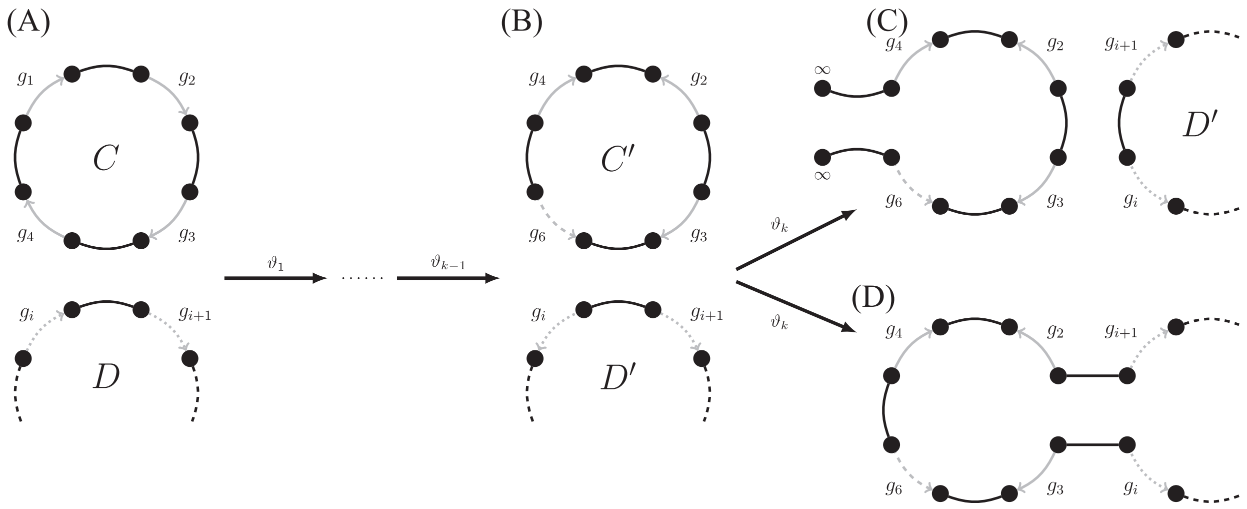

Figure 3. Illustration of a linearized circular chromosome C within a DCJ-Indel scenario $(\vartheta_1, \ldots, \vartheta_k)$ and Theorem 5. Dashed gray and black edges denote newly inserted genes and arbitrary gene sequences, respectively. Dotted edges represent genes that do not belong to meta-chromosomes of C. (A) Initial genome graph, where C is a linearized circular chromosome and D is a chromosome of any type. (B) The intermediate genome graph resulted from a DCJ-Indel scenario $(\vartheta_1, \ldots, \vartheta_{k-1})$, where C' is a meta-chromosome of C and D' is a chromosome obtained from D. (C) The resulting graph after a fission ϑ_k on a circular chromosome C'. (D) The graph resulted from a DCJ ϑ_k that combines a circular chromosome C' and a chromosome D'.

Methods

This section is devoted to the proof of Theorems 1 and 2.

We call any two DCJ-Indel scenarios between the same pair of genomes *equivalent*. Let $t : P \to Q$ be a DCJ-Indel scenario that linearizes a circular chromosome C. First, in Lemma 4, we will show that there exists an event r within t that linearizes a circular chromosome C. Second, in Theorem 5, we will show that r is a DCJ. Third, we will show how to obtain equivalent pair of events (i.e., a DCJ-Indel scenario of length 2) (α', β') from adjacent events (α, β) in t, where β and α' linearize C.

We will distinguish the pair of adjacent events based on their dependency. Namely, two adjacent events α and β in a DCJ-Indel scenario are called *independent* if the edges removed by β are not created by α. Otherwise, when β removes edge(s) created by α, we say that β *depends* on α. We will assume that β is a DCJ if not stated otherwise. We will consider the following cases:

(1) α and β are independent events (addressed in Lemma 6);
(2) β depends on a deletion α (addressed in Lemma 8);
(3) β depends on a DCJ α (addressed in Lemma 7);
(4) β depends on an insertion α (addressed in Lemmas 9 to 11).

Eventually, results of Lemmas 6 to 11 will enable us to prove Theorems 1 and 2.

Circular chromosomes and DCJ-Indel scenarios

The following lemma shows the connection between Definitions 1 and 2.

Lemma 4
Let $t : P \to Q$ be a DCJ-Indel scenario that linearizes a circular chromosome C. Then there exists an event r that linearizes C within t.

Proof
Suppose that t is of the form: $P = P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_n} P_n = Q$. For each $i \in \{0, 1, \ldots, n\}$, let $\mathfrak{C}_i$ be the meta-chromosome of C in P_i. In particular, $\mathfrak{C}_0 = \{C\}$. Then, the equality

$$S(\mathfrak{C}_i) \cap S(P) = S(C) \cap S(P_i) \tag{2}$$

holds for $i = 0$ but not for $i = n$ (since C is linearized within t). Hence, there exists $k \in \{1, \ldots, n\}$ such that equation (2) holds for $i = k - 1$ but not for $i = k$. Moreover, it is clear that $S(C) \cap S(P_{k-1}) \neq \emptyset$ and $S(C) \cap S(P_k) \neq \emptyset$. By Definition 1, C is not linearized within $(\vartheta_1, \ldots, \vartheta_{k-1})$ and is linearized within $(\vartheta_1, \ldots, \vartheta_k)$. Thus, $r = \vartheta_k$ linearizes C within t. □

An event linearizing a circular chromosome C can also be described in terms of removing edges in genome graphs as follows (Figure 3).

Theorem 5
Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario that linearizes a circular chromosome C. Let $\mathfrak{C}_i$ be the meta-chromosome of C in P_i for each $i \in \{0, 1, \ldots, n\}$. Then ϑ_k linearizes C within t if and only if ϑ_k is a DCJ with a minimal index k such that one of the following conditions holds:

(i) *ϑ_k removes edges $a \in \mathfrak{S}(\mathfrak{C}_{k-1})$ and $b \notin \mathfrak{S}(\mathfrak{C}_{k-1})$;*
(ii) *ϑ_k removes a single edge $a \in \mathfrak{S}(\mathfrak{C}_{k-1})$.*

Proof

Assume that ϑ_k is a DCJ and the above conditions (i) or (ii) holds, where k is the smallest such index. Since C is linearized within t, (E1) and (E2) hold for DCJ-Indel scenarios $(\vartheta_1, \ldots, \vartheta_{k-1})$ and $(\vartheta_1, \ldots, \vartheta_k)$. Now, we need to show that (E3) holds for $(\vartheta_1, \ldots, \vartheta_k)$, but not for $(\vartheta_1, \ldots, \vartheta_{k-1})$. We consider the following two cases.

If condition (i) holds, then a belongs to a circular chromosome $C' \in \mathfrak{C}_{k-1}$ and b belongs to a chromosome $D' \notin \mathfrak{C}_{k-1}$ (Figure 3D). If D' is circular, then ϑ_k creates a new circular chromosome $C'' \in \mathfrak{C}_k$ such that $S(C'') = S(C') \cup S(D')$ (ie, ϑ_k is a fusion of circular chromosomes). By Lemma 2, we have $S(D') \cap S(P_0) \neq \varnothing$. Since $S(D') \cap S(\mathfrak{C}_{k-1}) = \varnothing$, we have $S(\mathfrak{C}_k) \cap S(P_0) \neq S(C) \cap S(P_k)$, ie, (E3) holds. If D' is linear, then ϑ_k turns C' into a linear chromosome. Hence, we have $S(C') \cap S(\mathfrak{C}_k) = \varnothing$. Since $S(C) \cap S(C') \neq \varnothing$, we have $S(\mathfrak{C}_k) \cap S(P_0) \neq S(C) \cap S(P_k)$, ie, (E3) holds. Since k is the smallest index and $S(P_{k-1}) = S(P_k)$, (E3) does not hold for $(\vartheta_1, \ldots, \vartheta_{k-1})$.

If condition (ii) holds, the proof is similar (Figure 3C).

Now, assume that ϑ_k linearizes C within t. Then the equality

$$S(\mathfrak{C}_{k-1}) \cap S(P_0) = S(C) \cap S(P_{k-1}) \tag{3}$$

holds. There are three possible types of ϑ_k, namely, insertion, deletion, and DCJ. First, we assume that ϑ_k is an insertion. Then $S(P_{k-1}) \subset S(P_k)$. Recall that ϑ_k inserts genes from $S(P_n) \backslash S(P_0)$. In particular, since C is present in P_0 (ie, $S(C) \subseteq S(P_0)$), ϑ_k does not insert any genes from $S(C)$. Thus, we have $S(C) \cap S(P_{k-1}) = S(C) \cap S(P_k)$. Since insertions cannot change the chromosome types, we have $S(\mathfrak{C}_{k-1}) \cap S(P_0) = S(\mathfrak{C}_k) \cap S(P_0)$. By equation (3), we have a contradiction. Thus, ϑ_k is not an insertion.

Second, we assume that ϑ_k is a deletion. Then $S(P_k) \subset S(P_{k-1})$ and $S(C_k) \subseteq S(C_{k-1})$. Let

$$\begin{aligned} A &= (S(C) \cap S(P_{k-1})) \backslash (S(C) \cap S(P_k)) \\ &= S(C) \cap (S(P_{k-1}) \backslash S(P_k)) \end{aligned}.$$

Note that since $S(C) \cap S(P_n) \neq \varnothing$, we have $S(C) \cap S(P_i) \neq \varnothing$ for all $i \in \{0, 1, \ldots, n\}$. Then $A \neq S(C) \cap S(P_{k-1})$. Let

$$\begin{aligned} B &= (S(\mathfrak{C}_{k-1}) \cap S(P_0)) \backslash (S(\mathfrak{C}_k) \cap S(P_0)) \\ &= (S(\mathfrak{C}_{k-1}) \backslash S(\mathfrak{C}_k)) \cap S(P_0) \end{aligned}.$$

Our goal is to prove that $A = B$. Since A is the subset of genes removed by ϑ_k, $A \cap S(P_k) = \varnothing$. In particular, $A \cap S(\mathfrak{C}_k) = \varnothing$. Hence, $A \cap (S(\mathfrak{C}_k) \cap S(P_0)) = \varnothing$. By equation (3), we have that $A \subseteq B$. Now, let $g \in B$. Note that $g \in S(P_0)$, $g \in S(\mathfrak{C}_{k-1})$, and $g \notin S(\mathfrak{C}_k)$. Since deletion cannot change the chromosome types, it follows that g is removed by ϑ_k. Then $g \notin S(P_k)$. By equation (3), $g \in S(C) \cap S(P_{k-1})$, and thus we have $g \in A$. Since the choice of g was arbitrary, we have proved that $A = (S(\mathfrak{C}_{k-1}) \cap S(P_0)) \backslash (S(\mathfrak{C}_k) \cap S(P_0))$. Note that $S(\mathfrak{C}_k) \cap S(P_0) \subset S(\mathfrak{C}_{k-1}) \cap S(P_0)$ and $S(C) \cap S(P_k) \subset S(C) \cap S(P_{k-1})$. Therefore, $S(\mathfrak{C}_k) \cap S(P_0) = S(C) \cap S(P_k)$, a contradiction to ϑ_k linearizing C. Thus, ϑ_k is not a deletion.

We proved that ϑ_k is a DCJ. Then $S(P_{k-1}) = S(P_k)$. Hence,

$$\begin{aligned} S(\mathfrak{C}_{k-1}) \cap S(P_0) &= S(C) \cap S(P_{k-1}) \\ &= S(C) \cap S(P_k) \neq S(\mathfrak{C}_k) \cap S(P_0) \end{aligned}.$$

Thus, $S(\mathfrak{C}_{k-1}) \neq S(\mathfrak{C}_k)$ holds. Since ϑ_k does not change the gene content, ϑ_k either breaks one circular chromosome $C' \in \mathfrak{C}_{k-1}$, or combines circular chromosomes $C' \in \mathfrak{C}_{k-1}$ and $C'' \notin \mathfrak{C}_{k-1}$ into a single circular chromosome, or combines a circular chromosome $C' \in \mathfrak{C}_{k-1}$ and linear chromosome into a single linear chromosome. In the first case, ϑ_k removes a single edge that belongs to $\mathfrak{G}(\mathfrak{C}_{k-1})$ (Figure 3C). In the last two cases, among the two edges removed by ϑ_k, one must belong to $\mathfrak{G}(\mathfrak{C}_{k-1})$ and the other does not belong to $\mathfrak{G}(\mathfrak{C}_{k-1})$ (Figure 3D). □

The following lemma describes an important property of meta-chromosomes.

Lemma 5

Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario, where ϑ_n linearizes a circular chromosome C within t. Let $k \in \{0, \ldots, n-1\}$, and $\mathfrak{C}_k$ and $\mathfrak{C}_{k+1}$ be the meta-chromosomes of C in P_k and P_{k+1}, respectively. Then for any vertex $x \in V(P_k) \cap V(P_{k+1})$, if $x \in V(\mathfrak{C}_{k+1})$, then $x \in V(\mathfrak{C}_k)$.

Proof

Let g_x be the gene corresponding to x. Note that $x \in V(\mathfrak{C}_i)$ if and only if $g_x \in S(\mathfrak{C}_i)$ for $i \in \{k, k+1\}$. Since $g_x \in S(P_k)$ and $g_x \in S(P_{k+1})$, g_x cannot be inserted or removed by ϑ_{k+1}. Suppose that $x \in V(\mathfrak{C}_{k+1})$, ie, $g_x \in S(\mathfrak{C}_{k+1})$. We consider two cases depending on whether ϑ_{k+1} is an indel or a DCJ.

First, assume that ϑ_{k+1} is an indel. Since $g_x \in S(\mathfrak{C}_{k+1})$, there is a circular chromosome $C' \in \mathfrak{C}_{k+1}$ such that $g_x \in S(C')$. Let C'' be a chromosome in P_k such that $g_x \in S(C'')$, ie, $S(C'') \cap S(C') \neq \varnothing$. If $C'' = C'$ (ie, ϑ_{k+1} does not affect C'), we have $C'' \in \mathfrak{C}_k$, implying that $g_x \in S(\mathfrak{C}_k)$. If $C'' \neq C'$, we have either $S(C') \subset S(C'')$ or $S(C'') \subset S(C')$. Since $C' \in \mathfrak{C}_{k+1}$, in both cases, $C' \in \mathfrak{C}_k$. Therefore, $g_x \in S(\mathfrak{C}_k)$.

Second, assume that ϑ_{k+1} is a DCJ. Then, since ϑ_{k+1} does not linearize C, ϑ_{k+1} operates on four vertices that belong to $V(\mathfrak{C}_{k+1})$. Since ϑ_{k+1} is a DCJ, $S(\mathfrak{C}_k) = S(\mathfrak{C}_{k+1})$. Hence, these four vertices belong to $V(\mathfrak{C}_k)$. Thus, if $g_x \in S(\mathfrak{C}_{k+1})$ then $g_x \in S(\mathfrak{C}_k)$. □

From Lemma 5, the following corollary follows immediately.

Corollary 1

Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario, where ϑ_n linearizes a circular chromosome C. Let $k \in \{0, \ldots, n-1\}$ and x, y, z be vertices from $V(P_k) \cap V(P_{k+1})$ such that

$\{x, y\} \in \mathfrak{S}(P_k)$ and $\{x, z\} \in \mathfrak{S}(P_{k+1})$. Let $\mathfrak{C}_k$ and $\mathfrak{C}_{k+1}$ be the meta-chromosomes of C in P_k and P_{k+1}, respectively. If $\{x, z\} \in \mathfrak{S}(\mathfrak{C}_{k+1})$, then $\{x, y\} \in \mathfrak{S}(\mathfrak{C}_k)$.

Independent adjacent events

In this section, we address the case (1), ie, α and β are independent events. It is easy to see that the order of any two adjacent independent events in a DCJ-Indel scenario can be changed without affecting the starting and ending genomes.[15]

Lemma 6

Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta_{n-1}} P_{n-1} \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario that linearizes a circular chromosome C, where ϑ_{n-1} and ϑ_n are independent events. If ϑ_n linearizes C within t, then ϑ_n also linearizes C within the DCJ-Indel scenario $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta_n} P' \xrightarrow{\vartheta_{n-1}} P_n$.

Proof

Let $\mathfrak{C}_{n-2}$ and $\mathfrak{C}_{n-1}$ be the meta-chromosomes of C in P_{n-2} and P_{n-1}, respectively. Since ϑ_n linearizes C within t, by Theorem 5, ϑ_n is a DCJ. If ϑ_n removes two edges in $\mathfrak{S}(P_{n-1})$, say $\{x, y\} \in \mathfrak{S}(\mathfrak{C}_{n-1})$ and $\{z, w\} \notin \mathfrak{S}(\mathfrak{C}_{n-1})$, then since ϑ_{n-1} and ϑ_n are independent, the edges $\{x, y\}$ and $\{z, w\}$ are present in $\mathfrak{S}(P_{n-2})$. By Corollary 1, we have $\{x, y\} \in \mathfrak{S}(\mathfrak{C}_{n-2})$ and $\{z, w\} \notin \mathfrak{S}(\mathfrak{C}_{n-2})$. By Theorem 5, ϑ_n linearizes C within the DCJ-Indel scenario $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta_n} P' \xrightarrow{\vartheta_{n-1}} P_n$. If ϑ_n removes a single edge, the proof is similar. □

DCJ depends on a deletion

In this section, we consider case (2), ie, a DCJ β depends on a deletion α. For such pair of events the following lemma holds.

Lemma 7

Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta_{n-1}} P_{n-1} \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario that linearizes a circular chromosome C, where DCJ ϑ_n depends on deletion ϑ_{n-1}. If ϑ_n linearizes C, then there exists a DCJ-Indel scenario $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$, where ϑ'_{n-1} linearizes C and ϑ'_n is a deletion.

Proof

Let $\mathfrak{C}_{n-2}$ and $\mathfrak{C}_{n-1}$ be the meta-chromosomes of C in P_{n-2} and P_{n-1}, respectively. Let $(x, u_1, \bar{u}_1 \ldots, u_l, \bar{u}_l, y)$ be the path in $\mathfrak{S}(P_{n-2})$ that is replaced with $\{x, y\}$ in $\mathfrak{S}(P_{n-1})$ by ϑ_{n-1}.

Suppose that ϑ_n removes two edges. Since ϑ_n depends on ϑ_{n-1}, we can assume that ϑ_n removes edges $\{x, y\}$, $\{z, w\}$ in $\mathfrak{S}(P_{n-1})$ and creates $\{x, z\}$, $\{y, w\}$ in $\mathfrak{S}(P_n)$ (Figure 4A, B, and D). By Theorem 5, without loss of generality, we assume that $\{x, y\} \in \mathfrak{S}(\mathfrak{C}_{n-1})$ and $\{z, w\} \notin \mathfrak{S}(\mathfrak{C}_{n-1})$. We define ϑ'_{n-1} as the DCJ that removes edges $\{x, u_1\}$, $\{z, w\}$ in $\mathfrak{S}(P_{n-2})$, and creates $\{x, z\}$, $\{u_1, w\}$ in $\mathfrak{S}(P')$, where P' is the genome resulted from ϑ'_{n-1}. Moreover, we define ϑ'_n as the deletion that replaces a path $(w, u_1, \bar{u}_1, \ldots, u_l, \bar{u}_l, y)$ in $\mathfrak{S}(P')$ with an edge $\{y, w\}$ in $\mathfrak{S}(P_n)$ (Figure 4A, C, and D). Since ϑ_n depends on ϑ_{n-1}, $\{z, w\}$ is present in $\mathfrak{S}(P_{n-2})$. By Corollary 1, $\{x, u_1\} \in \mathfrak{S}(\mathfrak{C}_{n-2})$ and $\{z, w\} \notin \mathfrak{S}(\mathfrak{C}_{n-2})$. Thus, by Theorem 5, ϑ'_{n-1} linearizes C within $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$.

Suppose that ϑ_n removes a single edge a. Since ϑ_n depends on ϑ_{n-1}, we have $a = \{x, y\}$. We define ϑ'_{n-1} as the DCJ that removes a single edge $\{x, u_1\}$ and creates $\{x, \infty\}$, $\{u_1, \infty\}$, and ϑ'_n as the deletion that replaces a path $(\infty, u_1, \ldots, \bar{u}_l, y)$ with an edge $\{y, \infty\}$. The proof that ϑ'_{n-1} linearizes C within $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$ is similar. □

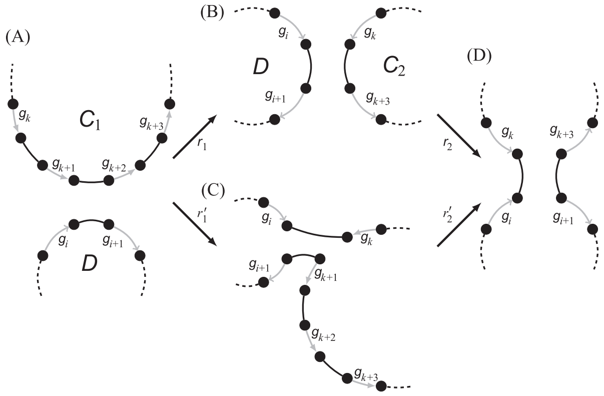

Figure 4. Illustration of Lemma 8. (A) Initial genome graph, where the dashed edges denote arbitrary gene sequences. C_1 is a circular chromosome linearized by r_2 within DCJ-Indel scenario (r_1, r_2), where r_1 is a deletion of gene sequence (g_{k+1}, g_{k+2}) and r_2 is a DCJ. (B) The intermediate genome resulted from deletion r_1. (C) The intermediate genome resulted from DCJ r'_1. (D) The graph resulted from the equivalent pair of DCJ-Indel scenarios (r_1, r_2) and (r'_1, r'_2), where C_1 is linearized by DCJs r'_1 and r_2, and r_1, r'_2 are deletions.

DCJ depends on a DCJ

In this section, we address case (3), ie, DCJ β depends on a DCJ α. Let A be the set of edges created by α, and B be the set of edges removed by β. Since β depends on α, $A \cap B \neq \emptyset$. We say that β *strongly depends* on α if $A = B$ and *weakly depends* on α otherwise (such pairs of adjacent DCJs are also known as *enchained*[15]). In a genome graph, a pair of adjacent dependent DCJs replaces three edges with three other edges on the same six vertices (this operation is known as a *3-break*[2]). It is easy to see that for a pair of weakly dependent DCJs, there exist equivalent pairs of weakly dependent DCJs.[15] Then the following lemma holds.

Lemma 8

Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta_{n-1}} P_{n-1} \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario that linearizes a circular chromosome C, where ϑ_{n-1} and ϑ_n are dependent DCJs. If ϑ_n linearizes C, then there exists a pair of DCJs $(\vartheta'_{n-1}, \vartheta'_n)$ equivalent to $(\vartheta_{n-1}, \vartheta_n)$ such that ϑ'_{n-1} linearizes C within the DCJ-Indel scenario $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$.

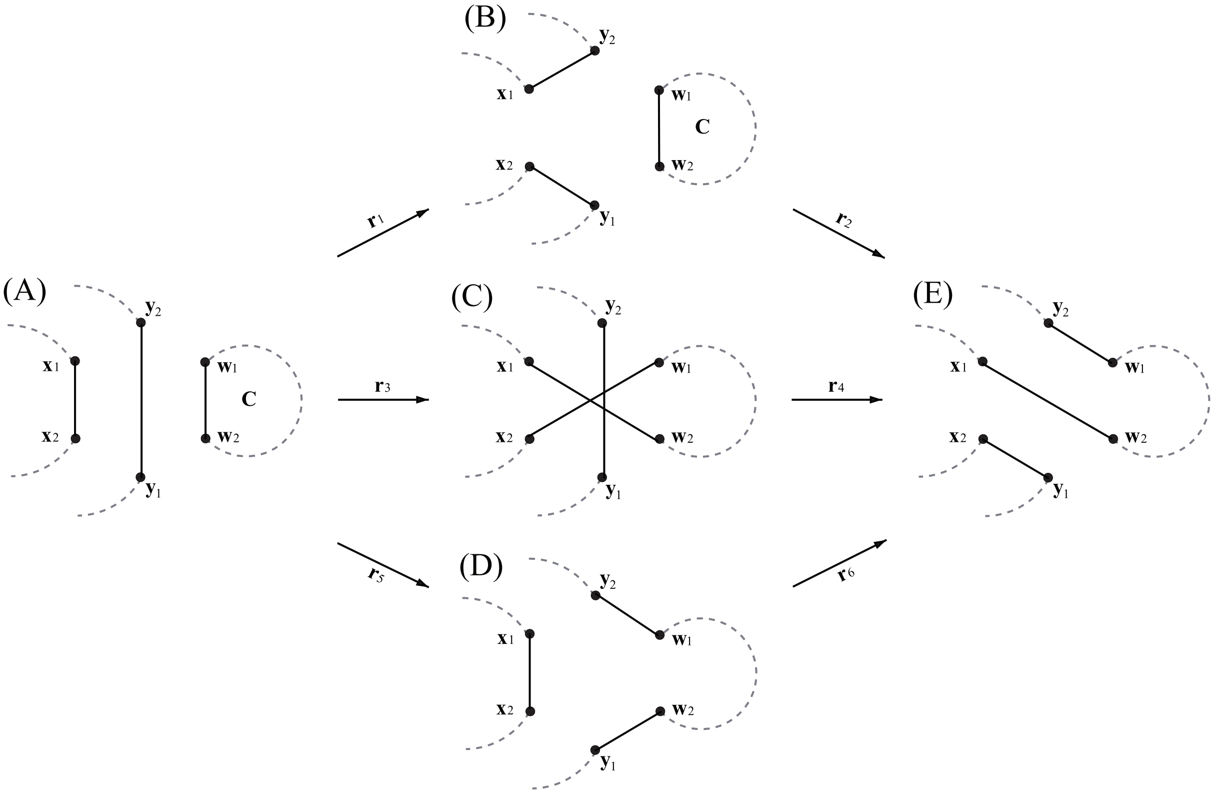

Figure 5. Illustration of Lemma 7. (A) Initial genome graph, where the dashed edges denote arbitrary gene sequences. The dashed edge and black undirected edge between w_1 and w_2 form a circular chromosome C that is linearized within (r_1, r_2) by r_2. (B-D) The intermediate genomes after first DCJs in the three equivalent pairs of weakly dependent DCJs. (E) The resulting genome graph after the equivalent pairs of DCJs, where C is linearized by DCJs r_2 and either r_3 or r_5 (depending on the belonging other chromosomes to meta-chromosome corresponding to C).

Proof

Let $\mathcal{C}_{n-2}$ and $\mathcal{C}_{n-1}$ be the meta-chromosomes of C in P_{n-2} and P_{n-1}, respectively. Let A be the set of edges created by α, and B be the set of edges removed by β. We consider two cases depending on whether ϑ_n strongly depends or weakly depends on ϑ_{n-1}.

First, assume that ϑ_n strongly depends on ϑ_{n-1} (ie, $A = B$). If $|A| = 2$, then let $\{x, y\}$ and $\{z, w\}$ be the edges removed by ϑ_n in P_{n-1}. By Theorem 5, without loss of generality, we assume that $\{x, y\} \in \mathfrak{S}(\mathcal{C}_{n-1})$ and $\{z, w\} \notin \mathfrak{S}(\mathcal{C}_{n-1})$. Since $\{x, y\}$ and $\{z, w\}$ are created by ϑ_{n-1}, the edges $\{x, z\}$ and $\{y, w\}$, or $\{x, w\}$ and $\{y, z\}$ are present in $\mathfrak{S}(P_{n-2})$. In both cases, we have a contradiction to Corollary 1. If $|A| = 1$, the proof is similar.

For the rest of the proof, we assume that ϑ_n weakly depends on ϑ_{n-1} (ie, $A \cap B \neq \emptyset$ and $A \neq B$). We consider two cases depending on the number of edges removed by ϑ_n.

If ϑ_n removes two edges in P_{n-1}, let $\{x, y\}$ and $\{z, w\}$ be these edges, and $\{x, z\}$ and $\{y, w\}$ be the edges created by ϑ_n in P_n. By Theorem 5, without loss of generality, we assume that $\{x, y\} \in \mathfrak{S}(\mathcal{C}_{n-1})$ and $\{z, w\} \notin \mathfrak{S}(\mathcal{C}_{n-1})$. Since ϑ_n weakly depends on ϑ_{n-1}, either $\{x, y\}$ or $\{z, w\}$ is created by ϑ_{n-1} (Figure 5A, B, and E). We consider these two subcases below.

Suppose that $\{x, y\}$ is created by ϑ_{n-1}. If ϑ_{n-1} creates a single edge, then ϑ_{n-1} removes edges $\{x, \infty\}$ and $\{y, \infty\}$ in $\mathfrak{S}(P_{n-2})$, a contradiction to Corollary 1. Thus, we assume that ϑ_{n-1} removes two edges, say $\{x, x_1\}$ and $\{y, y_1\}$. By Corollary 1, both $\{x, x_1\}$ and $\{y, y_1\}$ belong to $\mathfrak{S}(\mathcal{C}_{n-2})$. Since $\{z, w\} \in \mathfrak{S}(P_{n-2})$ and $\{z, w\} \notin \mathfrak{S}(\mathcal{C}_{n-1})$, by Corollary 1, $\{z, w\} \notin \mathfrak{S}(\mathcal{C}_{n-2})$. We define ϑ'_{n-1} as a DCJ that removes $\{z, w\}$ and $\{x, x_1\}$ in $\mathfrak{S}(P_{n-2})$ and creates $\{x, z\}$ and $\{w, x_1\}$ in $\mathfrak{S}(P')$. We further define ϑ'_n as a DCJ that removes $\{w, x_1\}$ and $\{y, y_1\}$ in $\mathfrak{S}(P')$ (Figure 5A, C, D, and E) and creates $\{y, w\}$ and $\{x_1, y_1\}$ in $\mathfrak{S}(P_n)$. Then by Theorem 5, ϑ'_{n-1} linearizes C within t'.

Suppose that $\{z, w\}$ is created by ϑ_{n-1}. Let us first assume that ϑ_{n-1} removes two edges $\{z, z_1\}$ and $\{w, w_1\}$. Since $\{z, w\} \notin \mathfrak{S}(\mathcal{C}_{n-1})$, by Corollary 1, $\{z, z_1\}$ and $\{w, w_1\}$ do not belong to $\mathfrak{S}(\mathcal{C}_{n-2})$. Moreover, since $\{x, y\} \in \mathfrak{S}(P_{n-2})$ and $\{x, y\} \in \mathfrak{S}(\mathcal{C}_{n-1})$, we have $\{x, y\} \in \mathfrak{S}(\mathcal{C}_{n-2})$. We define ϑ'_{n-1} as the DCJ that removes $\{x, y\}$ and $\{z, z_1\}$ in $\mathfrak{S}(P_{n-2})$ and creates $\{x, z\}$ and $\{y, z_1\}$ in $\mathfrak{S}(P')$. We define ϑ'_n as the DCJ that removes $\{w, w_1\}$ and $\{y, z_1\}$ in $\mathfrak{S}(P')$ and creates $\{y, w\}$ and $\{w_1, y_1\}$ in $\mathfrak{S}(P_n)$ (Figure 5). By Theorem 5, ϑ'_{n-1} linearizes C within t'. If ϑ_{n-1} removes a single edge, then the proof is similar.

If ϑ_n removes a single edge $\{x,y\}$ in P_{n-1}, then by Theorem 5, $\{x,y\} \in \mathfrak{S}(\mathcal{C}_{n-1})$. Since ϑ_{n-1} creates $\{x,y\}$, it removes two edges. We assume that these edges are $\{x,x_1\}$ and $\{y,y_1\}$. By Corollary 1, $\{x,x_1\}$ and $\{y,y_1\}$ belong to $\mathfrak{S}(\mathcal{C}_{n-2})$. We define ϑ'_{n-1} as a DCJ that removes a single edge in $\mathfrak{S}(P_{n-2})$, say $\{x,x_1\}$, and creates two edges $\{x,\infty\}$ and $\{x_1,\infty\}$ in $\mathfrak{S}(P')$. We define ϑ'_n as a DCJ that removes $\{y,y_1\}$ and $\{x_1,\infty\}$ in $\mathfrak{S}(P')$ and creates $\{y,\infty\}$ and $\{x_1,y_1\}$ in $\mathfrak{S}(P_n)$. By Theorem 5, ϑ'_{n-1} linearizes C within t'. It is easy to see that by construction, in all cases, $(\vartheta'_{n-1},\vartheta'_n)$ is equivalent to $(\vartheta_{n-1},\vartheta_n)$, which completes the proof. □

DCJ depends on an insertion

In this section, we consider case (4), ie, a DCJ β depends on an insertion α. We say that β *strongly depends* on α if β removes two edges created by α. If β removes one edge created by α, we say that β *weakly depends* on α. In contrast to cases (2) and (3), when β weakly depends on α, there may not always exist an equivalent pair (α',β'), where α' is a DCJ and β' is an insertion.

To better capture and analyze the combinatorial structure of events in a DCJ-Indel scenario t, we construct the *dependency graph*[16] DG(t) (also known as *overlap graph*[17,18]), whose vertices are labeled with events from t and there is an arc (δ,γ) whenever an event γ depends on an event δ. We remark that a DCJ can weakly depend on at most two insertions in a DCJ-Indel scenario. The following definition describes DCJs β in t for which the pair of adjacent events (α,β) does not have an equivalent pair (α',β'), where α',β are DCJs and α,β' are an insertion.

Definition 3
A DCJ β in a DCJ-Indel scenario t is called upper-movable if the following property holds:

- *If there exists exactly one insertion α in t such that there is a path from α to β in* DG(t), *say $(\alpha,\gamma,\ldots,\beta)$, then γ removes either the first or the last edge of the path inserted by α.*

First, we consider the case when a DCJ depends on two insertions (Figure 6). Second, we address the case when a DCJ is upper-movable and depends on only one insertion (Figure 7). Finally, we consider the case when a DCJ is not upper-movable.

Lemma 9
Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-3}} P_{n-3} \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta_{n-1}} P_{n-1} \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario that linearizes a circular chromosome C, where DCJ ϑ_n weakly depends on insertions ϑ_{n-1} and ϑ_{n-2}. If ϑ_n linearizes C, then there exists a DCJ-Indel scenario $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-3}} P_{n-3} \xrightarrow{\vartheta'_{n-2}} P' \xrightarrow{\vartheta'_{n-1}} P'' \xrightarrow{\vartheta'_n} P_n$, where ϑ'_{n-2} linearizes C and ϑ'_{n-1}, ϑ'_n are insertions.

Proof
Let $\mathfrak{C}_{n-3}$, $\mathfrak{C}_{n-2}$, and $\mathfrak{C}_{n-1}$ be the meta-chromosomes of C in P_{n-3}, P_{n-2}, and P_{n-1}, respectively. Let $\mathfrak{P}_1 = (x, u_1, \bar{u}_1, \ldots, u_l, \bar{u}_l, y)$ and $\mathfrak{P}_2 = (z, v_1, \bar{v}_1 \ldots, v_k, \bar{v}_k, w)$ be paths inserted by ϑ_{n-2} and ϑ_{n-1}, respectively. Since DCJ ϑ_n weakly depends on insertions ϑ_{n-1} and ϑ_{n-2}, without loss of generality, we assume that ϑ_n removes $\{\bar{u}_{p-1}, u_p\}$ and $\{\bar{v}_{q-1}, v_q\}$, and creates $\{\bar{u}_{p-1}, \bar{v}_{q-1}\}$ and $\{u_p, v_q\}$ for $p \in \{2,\ldots,l\}$ and $q \in \{2,\ldots,k\}$ (Figure 6A, B, and D). By Theorem 5, we have $\{\bar{u}_{p-1}, u_p\} \in \mathfrak{S}(\mathfrak{C}_{n-1})$ and $\{\bar{v}_{q-1}, v_q\} \notin \mathfrak{S}(\mathfrak{C}_{n-1})$. Then, all edges in $\mathfrak{P}_1$ belong to $\mathfrak{S}(\mathfrak{C}_{n-1})$ and all edges in $\mathfrak{P}_2$ do not belong to $\mathfrak{S}(\mathfrak{C}_{n-1})$. If ϑ_{n-1} depends on ϑ_{n-2} (ie, $z = \bar{u}_{s-1}$ and $w = u_s$ for some $s \in \{2,\ldots,l\}$), then all edges in $\mathfrak{P}_1$ and $\mathfrak{P}_2$ belong to $\mathfrak{S}(\mathfrak{C}_{n-1})$, a contradiction. Thus, ϑ_{n-1} and ϑ_{n-2} are independent events. We define ϑ'_{n-2} as a DCJ that removes $\{x,y\}$ and $\{z,w\}$ in P_{n-3} and creates $\{x,z\}$ and $\{y,w\}$ in P'. We define ϑ'_{n-1} and ϑ'_n as insertions that replace $\{x,z\}$ in P' with a path $(x, u_1, \bar{u}_1, \ldots, u_{p-1}, \bar{u}_{p-1}, \bar{v}_{q-1}, v_{q-1} \ldots, \bar{v}_1, v_1, z)$ in P'' and $\{y,w\}$ in P'' with a path $(y, \bar{u}_l, u_l \ldots, \bar{u}_p, u_p, v_q, \bar{v}_q, \ldots, v_k, \bar{v}_k, w)$ in P_n, respectively (Figure 6A, C, and D). By Corollary 1, $\{x,u_1\} \in \mathfrak{S}(\mathfrak{C}_{n-2})$ and $\{z,w\} \notin \mathfrak{S}(\mathfrak{C}_{n-2})$ and, moreover, $\{x,y\} \in \mathfrak{S}(\mathfrak{C}_{n-3})$ and $\{z,w\} \notin \mathfrak{S}(\mathfrak{C}_{n-3})$. By Theorem 5, ϑ'_{n-2} linearizes C within a DCJ-Indel scenario $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-3}} P_{n-3} \xrightarrow{\vartheta'_{n-2}} P' \xrightarrow{\vartheta'_{n-1}} P'' \xrightarrow{\vartheta'_n} P_n$. □

Lemma 10
Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta_{n-1}} P_{n-1} \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario that linearizes a circular chromosome C, where DCJ ϑ_n depends on insertion ϑ_{n-1}, and there is no $\alpha \in \{\vartheta_1,\ldots,\vartheta_{n-2}\}$ such that α is insertion connected by a path to ϑ_n in DG(t). *If ϑ_n is upper-movable and linearizes C, then there exists a DCJ-Indel scenario $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$, where ϑ'_{n-1} linearizes C and ϑ'_n is an insertion.*

Proof
Let $\mathfrak{C}_{n-2}$ and $\mathfrak{C}_{n-1}$ be the meta-chromosomes of C in P_{n-2} and P_{n-1}, respectively. Let $\mathfrak{P} = (u, u_1, \bar{u}_1, \ldots, u_l, \bar{u}_l, v)$ be a path inserted by ϑ_{n-1}.

Assume that ϑ_n strongly depends on ϑ_{n-1}. Let $\{x,y\}$ and $\{z,w\}$ be the edges removed by ϑ_n in P_{n-1}. Then $\{x,y\}$ and $\{z,w\}$ are inserted by ϑ_{n-1}, and thus belong to the same chromosome. Then by Theorem 5, ϑ_n cannot linearize C, a contradiction.

For the rest of the proof, we assume that ϑ_n weakly depends on ϑ_{n-1}. Since there is no insertion $\alpha \in \{\vartheta_1,\ldots,\vartheta_{n-2}\}$ connected by a path to ϑ_n in DG(t), ϑ_{n-1} is the only insertion in t that has a path to ϑ_n in DG(t). Since ϑ_n is upper-movable, ϑ_n removes $\{u,u_1\}$ or $\{\bar{u}_l, v\}$. We consider two cases depending on the number of edges removed by ϑ_n.

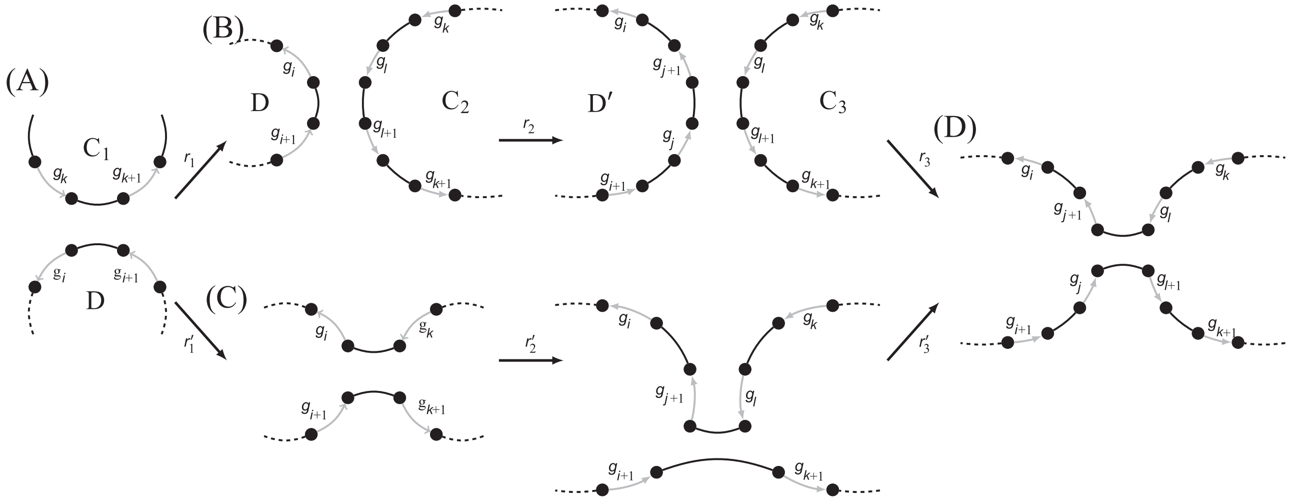

Figure 6. Illustration of Lemma 9. (A) Initial genome graph, where the dashed edges denote arbitrary gene sequences. C_1 is a circular chromosome linearized by r_3 within DCJ-Indel scenario (r_1, r_2, r_3), where r_3 is a DCJ and r_1, r_2 are insertions of gene sequences (g_l, g_{l+1}) and (g_j, g_{j+1}). (B) The intermediate genomes before and after an insertion r_2. (C) The intermediate genomes before and after an insertion r'_2. (D) The resulting graph after the equivalent pair of DCJ-Indel scenarios (r_1, r_2, r_3) and (r'_1, r'_2, r'_3), where C_1 is linearized by DCJs r'_1 and r_3, and r_1, r_2, r'_2, r'_3 are insertions.

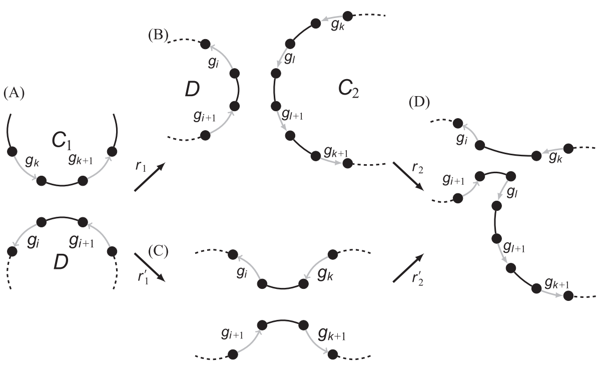

Figure 7. Illustration of Lemma 10. (A) Initial genome graph, where the dashed edges denote arbitrary gene sequences. C_1 is a circular chromosome linearized by r_2 within DCJ-Indel scenario (r_1, r_2), where r_1 is an insertion of gene sequences (g_l, g_{l+1}) and r_2 is a DCJ. (B) The intermediate genome after an insertion r_1. (C) The intermediate genome after a DCJ r'_1. (D) The resulting graph after the equivalent pair of DCJ-Indel scenarios (r_1, r_2) and (r'_1, r'_2), where C_1 is linearized by DCJs r'_1 and r_2, and r_1, r'_2 are insertions.

If ϑ_n removes two edges, then without loss of generality, we assume that ϑ_n removes $\{u, u_1\}$ and $\{x, y\}$ in $\mathfrak{S}(P_{n-1})$ and creates $\{u, x\}$ and $\{u_1, y\}$ in $\mathfrak{S}(P_n)$ (Figure 7A, B, and D). Let us define ϑ'_{n-1} as a DCJ that removes $\{u, v\}$ and $\{x, y\}$ in $\mathfrak{S}(P_{n-2})$ and creates $\{u, x\}$ and $\{y, v\}$ in $\mathfrak{S}(P')$. We define ϑ'_n as an insertion that replaces the edge $\{y, v\}$ in $\mathfrak{S}(P')$ with a path $(y, u_1, \bar{u}_1, \ldots, u_l, \bar{u}_l, v)$ in $\mathfrak{S}(P_n)$ (Figure 7A, C, and D). By Theorem 5, without loss of generality, $\{u, u_1\} \in \mathfrak{S}(\mathcal{C}_{n-1})$ and $\{x, y\} \notin \mathfrak{S}(\mathcal{C}_{n-1})$. By Corollary 1, $\{u, v\} \in \mathfrak{S}(\mathcal{C}_{n-2})$ and $\{x, y\} \notin \mathfrak{S}(\mathcal{C}_{n-2})$. Thus, by Theorem 5, ϑ'_{n-1} linearizes C within $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$.

If ϑ_n removes a single edge, the proof is similar. □

Lemma 11

Let $t : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta_{n-1}} P_{n-1} \xrightarrow{\vartheta_n} P_n$ be a DCJ-Indel scenario that linearizes a circular chromosome C, where DCJ ϑ_n weakly depends on insertion ϑ_{n-1}. If ϑ_n linearizes C within t and is not upper-movable, then there exists a DCJ-Indel scenario $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P'' \xrightarrow{\vartheta'_{n+1}} P_n$, where ϑ'_{n-1} linearizes C and $\vartheta'_n, \vartheta'_{n+1}$ are insertions.

Proof

Let $\mathfrak{C}_{n-2}$ and $\mathfrak{C}_{n-1}$ be the meta-chromosomes of C in P_{n-2} and P_{n-1}, respectively. Let $\mathfrak{P} = (u, u_1, \bar{u}_1 \ldots, u_l, \bar{u}_l, v)$ be a path inserted by ϑ_{n-1}. Since ϑ_n weakly depends on ϑ_{n-1} and is not upper-movable, ϑ_{n-1} breaks $\mathfrak{P}$ into two non-trivial subpaths. We consider two cases depending on the number of edges removed by DCJ ϑ_n.

If ϑ_n removes two edges, then without loss of generality, we assume that ϑ_n removes edges $\{\bar{u}_k, u_{k+1}\}$ $(k \in \{1, \ldots, l-1\})$ and $\{x, y\}$ in $\mathfrak{S}(P_{n-1})$ and creates edges $\{\bar{u}_k, x\}$ and $\{u_{k+1}, y\}$ in $\mathfrak{S}(P_n)$. By Theorem 5, we can assume that $\{\bar{u}_k, u_{k+1}\} \in \mathfrak{S}(\mathfrak{C}_{n-1})$ and $\{x, y\} \notin \mathfrak{S}(\mathfrak{C}_{n-1})$. Note that $\{x, y\}$ and $\{u, v\}$ are present in $\mathfrak{S}(P_{n-2})$. By Corollary 1, $\{u, v\} \in \mathfrak{S}(\mathfrak{C}_{n-2})$ and $\{x, y\} \notin \mathfrak{S}(\mathfrak{C}_{n-2})$. We define ϑ'_{n-1} as a DCJ that removes $\{x, y\}$ and $\{u, v\}$ in $\mathfrak{S}(P_{n-2})$ and creates $\{x, u\}$ and $\{y, v\}$ in $\mathfrak{S}(P')$. We define ϑ'_n and ϑ'_{n+1} as insertions that replace edges $\{x, u\}$ in $\mathfrak{S}(P')$ and $\{y, v\}$ in $\mathfrak{S}(P'')$ with paths $(u, u_1, \bar{u}_1, \ldots, u_k, \bar{u}_k, x)$ in $\mathfrak{S}(P'')$ and $(y, u_{k+1}, \bar{u}_{k+1} \ldots, u_l, \bar{u}_l, v)$ in $\mathfrak{S}(P_n)$, respectively. By Theorem 5, ϑ'_{n-1} linearizes C within $P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P'' \xrightarrow{\vartheta'_{n+1}} P_n$.

If ϑ_n removes a single edge, the proof is similar. □

Proof of Theorems 1 and 2

We remark that for each pair of adjacent events (α, β), there is an equivalent pair of adjacent events (α', β'), where α', β are insertions and α, β' have the same type. Below we prove Theorem 6, which will imply Theorems 1 and 2.

Theorem 6
Let $t : P \to Q$ be a DCJ-Indel scenario that linearizes a circular chromosome C. Then there exists a DCJ-Indel scenario $P \xrightarrow{r} P' \xrightarrow{t'} Q$ such that r is a DCJ linearizing C, and if C is linearized by an upper-movable DCJ within t, then $|t'| = |t| - 1$, otherwise $|t'| = |t|$.

Proof
We prove the theorem statement by induction on $|t|$. If $|t| = 1$, then by Lemma 4 and Theorem 5, the statement trivially holds.

For an integer $n \geqslant 2$, we assume that the theorem holds for all $|t| < n$. Suppose that t has length n, ie, t has the form $t: \ P = P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_n} P_n = Q$. We consider two cases depending on whether ϑ_n linearizes C within t.

Case 1. ϑ_n does not linearize C within t. By Lemma 4, there exists an event ϑ_k for $k < n$ that linearizes C within t. By induction, we obtain a DCJ-Indel scenario $t_1 : P_0 \xrightarrow{r} P_1' \xrightarrow{\vartheta_2'} \cdots \xrightarrow{\vartheta_l'} P_k \xrightarrow{\vartheta_{k+1}} \cdots \xrightarrow{\vartheta_n} P_n$, where r linearizes C and $|t| \leqslant |t_1| \leqslant |t| + 1$. We let $t' = (\vartheta_2', \ldots, \vartheta_l', \vartheta_{k+1}, \ldots, \vartheta_n)$. It is clear that $|t'| = |t| - 1$ if ϑ_k is upper-movable, and $|t'| = |t|$ otherwise.
Case 2. ϑ_n linearizes C within t. By Theorem 5, ϑ_n is a DCJ. We consider two cases depending on whether ϑ_n depends on ϑ_{n-1}. If ϑ_n does not depend on ϑ_{n-1}, then, by Lemma 6, we obtain a DCJ-Indel scenario $t_1 : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$, where $\vartheta'_{n-1} = \vartheta_n$ and $\vartheta'_n = \vartheta_{n-1}$, and ϑ'_{n-1} linearizes C. If ϑ_n depends on ϑ_{n-1} and ϑ_{n-1} is a DCJ or a deletion, then by Lemma 7 or 8, we obtain a DCJ-Indel scenario $t_1 : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$, where ϑ'_{n-1} linearizes C. In both cases, applying the induction to t_1, we obtain a DCJ-Indel scenario $t_2 : P_0 \xrightarrow{r} P_1' \xrightarrow{\vartheta_2''} \cdots \xrightarrow{\vartheta_l''} P' \xrightarrow{\vartheta'_n} P_n$, where r linearizes C and $|t| \leqslant |t_2| \leqslant |t| + 1$. Now, we let $t' = (\vartheta_2'', \ldots, \vartheta_l'', \vartheta_n')$. It is clear that $|t'| = |t| - 1$ if ϑ'_{n-1} is upper-movable, and $|t'| = |t|$ otherwise.

It remains to consider the case when DCJ ϑ_n depends on ϑ_{n-1}, ϑ_{n-1} is an insertion, which we split into two subcases depending on whether ϑ_n is upper-movable.

Case 2.1. ϑ_n is upper-movable. Here we consider two cases depending on whether there exists any insertion other than ϑ_{n-1} that is connected by a path to ϑ_n in $\mathrm{DG}(t)$.

Case 2.1.1. ϑ_{n-1} is a single insertion such that there is a path to ϑ_n in $\mathrm{DG}(t)$. By Lemma 10, we obtain a DCJ-Indel scenario $t_1 : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$, where ϑ'_{n-1} linearizes C. Since ϑ'_{n-1} is upper-movable in t_1, by induction, we obtain a DCJ-Indel scenario $t_2 : P_0 \xrightarrow{r} P_1' \xrightarrow{\vartheta_2''} \cdots \xrightarrow{\vartheta''_{n-2}} P'_{n-2} \xrightarrow{\vartheta''_{n-1}} P' \xrightarrow{\vartheta'_n} P_n$, where r linearizes C and $|t_2| = |t|$. We let $t' = (\vartheta_2'', \ldots, \vartheta''_{n-2}, \vartheta'_{n-1}, \vartheta'_n)$ to complete the proof.
Case 2.1.2. There exists an insertion ϑ_i with $i < n - 1$ connected by a path to ϑ_n in $\mathrm{DG}(t)$. We consider two subcases depending on whether $i = n - 2$. If $i = n - 2$, then by Lemma 9, we obtain a DCJ-Indel scenario $t_1 : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-3}} P_{n-3} \xrightarrow{\vartheta'_{n-2}} P' \xrightarrow{\vartheta'_{n-1}} P'' \xrightarrow{\vartheta'_n} P_n$, where ϑ'_{n-2} linearizes C and $|t_1| = |t|$. Since ϑ'_{n-2} is upper-movable in t_1, by induction, we obtain a DCJ-Indel scenario $t_2 : P_0 \xrightarrow{r} P_1' \xrightarrow{\vartheta_2''} \cdots \xrightarrow{\vartheta''_{n-2}} P' \xrightarrow{\vartheta'_{n-1}} P'' \xrightarrow{\vartheta'_n} P_n$, where r linearizes C and $|t_2| = |t|$. We let $t' = (\vartheta_2'', \ldots, \vartheta''_{n-1}, \vartheta'_n)$ to complete the proof. If $i \neq n - 2$, then we replace the pair of adjacent events $(\vartheta_{n-2}, \vartheta_{n-1})$ in t with an equivalent pair of adjacent events $(\vartheta'_{n-2}, \vartheta'_{n-1})$, where ϑ'_{n-2} is an insertion and ϑ'_{n-1} has the same type as ϑ_{n-2}, resulting in $t_1 : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-3}} P_{n-2} \xrightarrow{\vartheta'_{n-2}} P' \xrightarrow{\vartheta'_{n-1}} P_{n-1} \xrightarrow{\vartheta_n} P_n$. By Lemmas 6 to 8 for the pair of adjacent events $(\vartheta'_{n-1}, \vartheta_n)$ (depending on the type of ϑ'_{n-1} and dependency with ϑ_n), we obtain a DCJ-Indel scenario $t_2 : P_0 \xrightarrow{\vartheta_1} \cdots \xrightarrow{\vartheta_{n-3}} P_{n-2} \xrightarrow{\vartheta'_{n-2}} P' \xrightarrow{\vartheta''_{n-1}} P'' \xrightarrow{\vartheta''_n} P_n$, where ϑ''_n linearizes C and $|t_2| = |t|$. Since ϑ''_{n-1} is upper-movable in t_2, by induction, we obtain a DCJ-Indel scenario $t_3 : P_0 \xrightarrow{r} P_1' \xrightarrow{\vartheta_2'''} \cdots \xrightarrow{\vartheta'''_{n-1}} P'' \xrightarrow{\vartheta''_n} P_n$, where r linearizes C and $|t_3| = |t|$. We let $t' = (\vartheta_2''', \ldots, \vartheta'''_{n-1}, \vartheta''_n)$ to complete the proof.
Case 2.2. ϑ_n is not upper-movable. By Lemma 11, we obtain a DCJ-Indel scenario $t_1 : P_0 \xrightarrow{\vartheta_1} P_1 \cdots \xrightarrow{\vartheta_{n-2}} P_{n-2} \xrightarrow{\vartheta'_{n-1}} P' \xrightarrow{\vartheta'_n} P'' \xrightarrow{\vartheta'_{n+1}} P_n$, where ϑ'_{n-1} linearizes C and $|t_1| = |t| + 1$. Since there is no insertion α in the DCJ-Indel scenario t_1 connected by a path to ϑ'_{n-1} in $\mathrm{DG}(t_1)$, ϑ'_{n-1} is upper-movable in t_1. By induction, we obtain a DCJ-Indel scenario $t_2 : P_0 \xrightarrow{r} P_1' \xrightarrow{\vartheta_2''} \cdots \xrightarrow{\vartheta''_{n-1}} P' \xrightarrow{\vartheta'_n} P'' \xrightarrow{\vartheta'_{n+1}} P_n$, where r linearizes C and $|t_2| = |t| + 1$. We let $t' = (\vartheta_2'', \ldots, \vartheta''_{n-1}, \vartheta'_n, \vartheta'_{n+1})$ to complete the proof.

Theorems 1 and 2 immediately follow from Theorem 6.

Discussion

For three given linear genomes and their DCJ median genome M (which may contain circular chromosomes), we described an algorithm that constructs a linear genome M' such that the approximation error of M' (ie, the difference in the DCJ

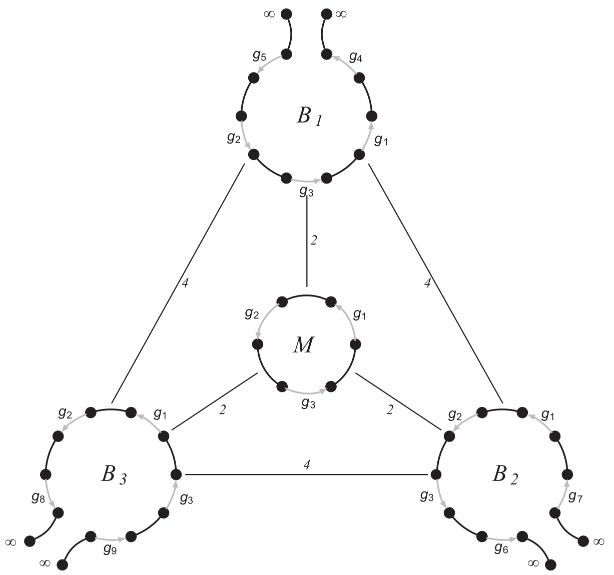

Figure 8. A circular median genome M on genes $\{g_1, g_2, g_3\}$ of three unichromosomal linear genomes B_1, B_2, and B_3 on genes $\{g_1, g_2, g_3, g_4, g_5\}$, $\{g_1, g_2, g_3, g_6, g_7\}$, and $\{g_1, g_2, g_3, g_8, g_9\}$, respectively, with specified pairwise DCJ-Indel distances, where $g_4, g_5, g_6, g_7, g_8, g_9$ are inserted genes.

median scores of M' and M) is bounded by twice the number of circular chromosomes in M.

We claim (and will prove elsewhere) that the bound in Theorem 3 is tight. We illustrate this claim with Figure 8, where each of the linear genomes B_1, B_2, and B_3 can be obtained from the genome M by an insertion followed by a fission. Note that all the pairwise DCJ distances between $B_1, B_2,$ andB_3 equal 4. We claim that the DCJ-Indel median score of M is 6, while any linearization of M has the DCJ-Indel median score at least 8, implying that the bound in Theorem 3 is tight.

At the same time, it was previously observed by Xu[8] on simulated data that the number of circular chromosomes produced by their GMP solver is typically very small, implying negligible approximation error for our algorithm.

The proposed algorithm is implemented in the AGRP solver MGRA2.[19]

Acknowledgements

The authors thank the anonymous reviewers for their thoughtful comments regarding the earlier version of this paper. Some preliminary results of the present work appeared in the Proceedings of the 14th Workshop on Algorithms in Bioinformatics (WABI 2014).

Author Contributions

All authors participated in the manuscript preparation.

ORCID iD

Max A Alekseyev https://orcid.org/0000-0002-5140-8095

REFERENCES

1. Yancopoulos S, Attie O, Friedberg R. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*. 2005;21:3340–3346. doi:10.1093/bioinformatics/bti535.
2. Alekseyev MA, Pevzner PA. Multi-break rearrangements and chromosomal evolution. *Theor Comput Sci*. 2008;395:193–202. doi:10.1016/j.tcs.2008.01.013.
3. Braga MDV, Willing E, Stoye J. Genomic distance with DCJ and indels. In: Moulton V, Singh M, eds. *Algorithms in Bioinformatics*. Vol 6293. Berlin, Germany: Springer; 2010:90–101. doi:10.1007/978-3-642-15294-8_8.
4. Compeau P. DCJ-indel sorting revisited. *Algorithm Mol Biol*. 2013;8:6. doi:10.1186/1748-7188-8-6.
5. Caprara A. The reversal median problem. *INFORMS J Comput*. 2003;15:93–113.
6. Tannier E, Zheng C, Sankoff D. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*. 2009;10:120.
7. Xu AW. A fast and exact algorithm for the median of three problem: a graph decomposition approach. *J Comput Biol*. 2009;16:1369–1381.
8. Xu AW. DCJ median problems on linear multichromosomal genomes: graph representation and fast exact solutions. In: Ciccarelli FD, Miklos I, eds. *Comparative Genomics*. Vol 5817. Berlin, Germany: Springer; 2009:70–83. doi:10.1007/978-3-642-04744-27.
9. Zhang M, Arndt W, Tang J. An exact solver for the DCJ median problem. *Paper presented at: Pacific Symposium on Biocomputing;* November 5-9, 2009; Big Island, HI:138–149. Singapore: World Scientific.
10. Maňuch J, Patterson M, Wittler R, et al. Linearization of ancestral multichromosomal genomes. *BMC Bioinformatics*. 2012;13:S11.
11. Ma J, Zhang L, Suh BB, et al. Reconstructing contiguous regions of an ancestral genome. *Genome Res*. 2006;16:1557–1565.
12. Ma J, Ratan A, Raney BJ, et al. DUPCAR: reconstructing contiguous ancestral regions with duplications. *J Comput Biol*. 2008;15:1007–1027.
13. Muffato M, Louis A, Poisnel CE, et al. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*. 2010;26:1119–1121.
14. Avdeyev P, Alexeev N, Rong Y, et al. A unified ILP framework for genome median, halving, and aliquoting problems under DCJ. *Paper presented at: Proceedings of the 15th Annual Research in Computational Molecular Biology Satellite Workshop on Comparative Genomics (RECOMB-CG)*; October 4-6, 2017; Barcelona, Spain. Vol 10562:156–178. Berlin, Germany: Springer. doi:10.1007/978-3-319-67979-29.
15. Braga MD, Stoye J. The solution space of sorting by DCJ. *J Comput Biol*. 2010;17:1145–1165.
16. Avdeyev P, Jiang S, Alekseyev MA. Implicit transpositions in DCJ scenarios. *Front Genet*. 2017;8:212. doi:10.3389/fgene.2017.00212.
17. Ozery-Flato M, Shamir R. Sorting by translocations via reversals theory. *Paper presented at: Proceedings of the 4th RECOMB International Workshop on Comparative Genomics (RECOMB-CG)*; September 24-26, 2006; Montreal, QC, Canada. Vol 4205:87–98. Berlin, Germany: Springer. doi:10.1007/118641278.
18. Ouangraoua A, Bergeron A. Combinatorial structure of genome rearrangements scenarios. *J Comput Biol*. 2010;17:1129–1144.
19. Avdeyev P, Jiang S, Aganezov S, et al. Reconstruction of ancestral genomes in presence of gene gain and loss. *J Comput Biol*. 2016;23:150–164. doi:10.1089/cmb.2015.0160.