Frontiers in Genetics

TYPE Methods
PUBLISHED 20 March 2024
DOI 10.3389/fgene.2024.1364951

OPEN ACCESS

EDITED BY
Ka-Chun Wong,
City University of Hong Kong, Hong Kong SAR, China

REVIEWED BY
Mengyang Xu,
Beijing Genomics Institute (BGI), China
Xin Gao,
King Abdullah University of Science and Technology, Saudi Arabia

*CORRESPONDENCE
Stephen S.-T. Yau,
✉ yau@uic.edu

RECEIVED 03 January 2024
ACCEPTED 06 March 2024
PUBLISHED 20 March 2024

CITATION
Yu H and Yau SS-T (2024), Automated recognition of chromosome fusion using an alignment-free natural vector method.
Front. Genet. 15:1364951.
doi: 10.3389/fgene.2024.1364951

Automated recognition of chromosome fusion using an alignment-free natural vector method

Hongyu Yu[1] and Stephen S.-T. Yau[1,2]*

[1]Department of Mathematical Sciences, Tsinghua University, Beijing, China, [2]Yanqi Lake Beijing Institute of Mathematical Science and Applications (BIMSA), Beijing, China

Chromosomal fusion is a significant form of structural variation, but research into algorithms for its identification has been limited. Most existing methods rely on synteny analysis, which necessitates manual annotations and always involves inefficient sequence alignments. In this paper, we present a novel alignment-free algorithm for chromosomal fusion recognition. Our method transforms the problem into a series of assignment problems using natural vectors and efficiently solves them with the Kuhn-Munkres algorithm. When applied to the human/gorilla and swamp buffalo/river buffalo datasets, our algorithm successfully and efficiently identifies chromosomal fusion events. Notably, our approach offers several advantages, including higher processing speeds by eliminating time-consuming alignments and removing the need for manual annotations. By an alignment-free perspective, our algorithm initially considers entire chromosomes instead of fragments to identify chromosomal structural variations, offering substantial potential to advance research in this field.

KEYWORDS

chromosomal fusion, alignment-free, natural vector, Kuhn-Munkres algorithm, automated recognition

1 Introduction

Chromosome fusion, a genetic event wherein two or more separate chromosomes merge to form a single chromosome, is a substantial restructuring of the genome. A primary factor leading to the chromosome fusion is Robertsonian translocation, a chromosomal abnormality where the long arms of two different chromosomes are linked Poot and Hochstenbach (2021). As is shown in Figure 1, initially, the short arms of these two chromosomes are also linked, but they are usually lost afterward. (The short arms, being too short to harbor significant genetic information, may not lead to lethality upon their loss.) Chromosome fusion holds significant implications for cellular processes and biological evolution Vara et al. (2021); Feulner and De-Kayne (2017). On one hand, chromosome fusion can interfere the process of meiosis and lead to the production of imbalanced gametes, potentially diminishing reproductive compatibility Cicconardi et al. (2021); Hauffe and Searle (1998). On the other hand, chromosome fusion physically connects genes that were originally located on different chromosomes, thereby reducing their potential for recombination, which can be instrumental in preserving co-adapted alleles Cicconardi et al. (2021); Guerrero and Kirkpatrick (2014). These two facets underscore the significance of chromosome fusion in the process of speciation.

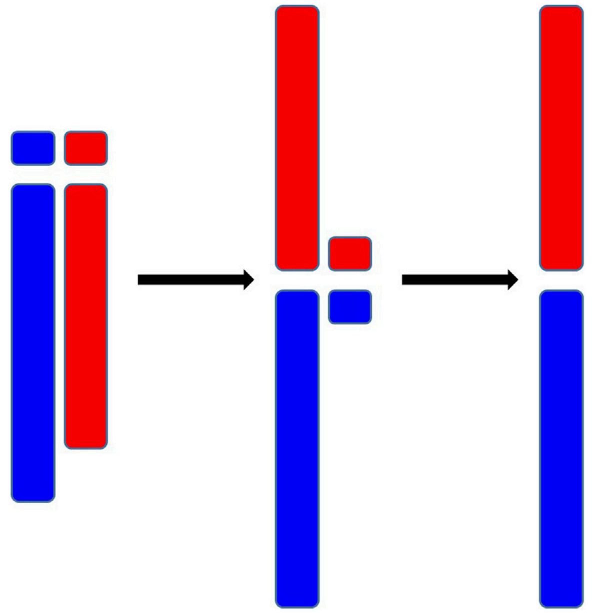

FIGURE 1
Demonstration of Robertsonian translocations.

Chromosome fusion presents several well-documented instances across different species. One of the most prominent examples is the case of human chromosome 2. Humans possess 23 pairs of chromosomes, while other great apes, such as chimpanzees and gorillas, have 24 pairs. Extensive research suggests that human chromosome 2 is the result of the fusion of two ancestral chromosomes from great apes Yunis and OM (1982); Ijdo et al. (1991). Additionally, chromosome fusion can also occur within the same species. For instance, the water buffalo (Bubalus bubalis) consists of two distinct subspecies, the swamp buffalo and the river buffalo, with chromosome numbers of 48 and 50, respectively Iannuzzi et al. (2021). This difference in chromosome number is also attributed to chromosome fusion events.

Despite the significance of chromosome fusion as a structural variation in chromosomes, research in this domain has been relatively limited in comparison to gene-level investigations. Currently, there are several structural variation detection algorithms designed for identifying gene structural variations within the same species. These algorithms prove effective in detecting genetic diseases caused by structural variations in humans Cameron et al. (2021); Layer et al. (2014). However, their applicability between different species is challenging. For fusion events between different species, the mainstream methods center around synteny analysis, with examples including Fish, Cinteny, and MCScan Calabrese et al. (2003); Sinha and Meller (2007); Tang et al. (2008b,a); Wang et al. (2012). The fundamental approach of these algorithms can be summarized in two main steps. Firstly, chromosomes are partitioned into multiple regions utilizing experimentally obtained annotation information, such as the coding sequence range. Subsequently, alignment algorithms are applied to compare and analyze these regions Altschul et al. (1990); Haas et al. (2005). These algorithms provide the advantage of delivering clear and visually interpretable results. Nevertheless, they come with limitations, given their reliance on manual annotation as well as alignment algorithms which can be computationally intensive.

Indeed, methods exist for detecting structural variations in the human genome without relying on alignment Liu et al. (2021). However, these methods still involve partitioning the sequence into multiple regions and using strategies like k-mer search as a substitute for alignment, which is logically similar to alignment. We aim to adopt a more purely alignment-free perspective by directly embedding each sequence into a vector and performing operations solely on vectors, rather than comparing sequences with each other. There are many methods that map sequences into vectors Qi et al. (2004); Jun et al. (2009), and the natural vector method is an effective approach among them Deng et al. (2011); Wen et al. (2014b). By incorporating statistical moments, the natural vector works well in sequence comparison and phylogenetic analysis. Additionally, the convex hulls formed by natural vectors from distinct families do not overlap, demonstrating the favorable separation properties of natural vectors Wen et al. (2014a); Sun et al. (2021); Tian et al. (2018).

In this paper, we addressed the issue of chromosome fusion recognition from a novel perspective. We utilized the natural vector approach to extract statistical information from sets of chromosomes. Subsequently, we framed the pairing of chromosome sets as an assignment problem and identified the most likely fusion scenarios by minimizing the assignment loss Kuhn (1955). Applying this algorithm to the human/gorilla and swamp buffalo/river buffalo datasets, we successfully and efficiently identified the correct chromosome fusion scenarios without the need for annotations or alignments. Moreover, the process is significantly faster than traditional synteny analysis.

2 Materials and methods

2.1 Materials

The data utilized in this paper comprises the reference chromosomes of four distinct species: human (*Homo sapiens*), gorilla (*Gorilla gorilla*), swamp buffalo (Bubalus carabanensis), and river buffalo (Bubalus Bubalis). We downloaded these sequences from the National Center for Biotechnology Information (NCBI) on 10 October 2023. The sequences can be accessed through the following URL:https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/.

The IDs of these sequences will be listed in the Supplementary Material S1, and hereafter, we use numerical labels to represent these chromosomes. The autosomes for humans, swamp buffalo, and river buffalo are numbered from 1 to 22, 1 to 23, and 1 to 24, respectively, while gorilla autosomes are numbered from 1, 2A, 2B, 3 to 22, consistent with the sequence annotation labels.

2.2 Problem formulation

We will first elaborate on the specific representation of the chromosome fusion problem. For convenience, we only considered the scenario where only a pair of chromosomes fused. The situations involving multiple fusion or the fusion of multiple chromosomes are similar and only require an expansion of the enumerated cases. Then, the task of recognizing chromosome fusion can be precisely described as follows: given two sets of chromosomes, $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n, b_{n+1}\}$, it is known that a pair of chromosomes from B fuse to form a single chromosome in A. The objective is to identify which pair of chromosomes from B fuse together and establish a correspondence between the remaining chromosomes in both sets.

Instead of fragmenting chromosomes and employing local alignment techniques as done in synteny analysis, our approach takes a holistic approach to analyze sequences from a different perspective. In the subsequent sections, we will first introduce the extraction of sequence information using k-mer natural vectors. Following that, we will discuss the assignment problem and its corresponding solution, the Kuhn-Munkres algorithm. Finally, we will illustrate how to transform the chromosome fusion problem into an assignment problem using k-mer natural vectors and subsequently solve it.

2.3 Natural vectors and their properties in chromosome fusion

The natural vector method is an alignment-free technique that converts DNA sequences into moment vectors Deng et al. (2011). For a given DNA sequence $S = s_1 s_2 \ldots s_n$, we define:

$$w_k(s_i) = \begin{cases} 1, & s_i = k \\ 0, & otherwise \end{cases} \tag{1}$$

where $k, s_i \in \{A, T, C, G\}$. Then the order 2 natural vector $nv(S)$ can be defined as

$$nv(S) = \left(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T\right) \tag{2}$$

where

$$\begin{cases} n_k = \sum_{i=1}^{N} w_k(s_i) \\ \mu_k = \sum_{i=1}^{N} \frac{i}{n_k} w_k(s_i) \\ D_2^k = \sum_{i=1}^{N} \frac{(i-\mu_k)^2}{n_k N} w_k(s_i) \\ N = n_A + n_T + n_C + n_G \end{cases} \tag{3}$$

If $n_k = 0$, we let $\mu_k = D_2^k = 0$. In this paper, we concentrate on only the order 2 natural vectors; thus, we will omit the order designation in the following content.

Given two single-stranded sequences, S_1 and S_2, which are oriented from 5′ to 3', there are two possible representations of fusion, denoted as $S_1 + S_2$ (with S_1 in the front) and $S_2 + S_1$ (with S_2 in the front). However, for double-stranded sequences, it becomes much more complex. Each chromosome consists of two strands named the forward strand and the reverse strand respectively. Consequently, there are a total of 8 possible representations of fusion. If the number of segments increases to k, the number of possible representations of fusion increases rapidly to $k! \times 2^k$. Therefore, it is not efficient to generate each possible fusion and calculate their natural vectors separately. In fact, if we have already calculated $nv(S_1)$ and $nv(S_2)$, we can obtain $nv(S_3)$ where $S_3 = S_1 + S_2$ through the following calculation.
Let $nv(S_i) = (n_{iA}, n_{iC}, n_{iG}, n_{iT}, \mu_{iA}, \mu_{iC}, \mu_{iG}, \mu_{iT}, D_2^{iA}, D_2^{iC}, D_2^{iG}, D_2^{iT})$, $N_i = n_{iA} + n_{iC} + n_{iG} + n_{iT}$, and $C(S, k) = \{i | w_k(s_i) = 1\}$ where $S = s_1 s_2 \ldots s_n$, then we have

$$\begin{aligned} n_{3k} &= n_{1k} + n_{2k} \\ \mu_{1k} &= \frac{\sum_{i \in C(S_1,k)} i}{n_{1k}} \\ \mu_{2k} &= \frac{\sum_{i \in C(S_2,k)} i}{n_{2k}} \end{aligned} \tag{4}$$

$$\begin{aligned} \mu_{3k} &= \frac{\sum_{i \in C(S_1,k)} i + \sum_{i \in C(S_2,k)} (i + N_1)}{n_{1k} + n_{2k}} \\ &= \frac{n_{1k}}{n_{1k} + n_{2k}} \mu_{1k} + \frac{n_{2k}}{n_{1k} + n_{2k}} \left(\mu_{2k} + N_1\right) \end{aligned} \tag{5}$$

$$\begin{aligned} n_{1k} N_1 D_2^{1k} &= \sum_{i \in C(S_1,k)} \left(i - \mu_{1k}\right)^2 \\ n_{2k} N_2 D_2^{2k} &= \sum_{i \in C(S_2,k)} \left(i - \mu_{2k}\right)^2 \\ &\sum_{i \in C(S_1,k)} \left(i - \mu_{3k}\right)^2 + \sum_{i \in C(S_2,k)} \left(i + N_1 - \mu_{3k}\right)^2 \\ = &\sum_{i \in C(S_1,k)} \left(i - \mu_{1k}\right)^2 + \sum_{i \in C(S_1,k)} \left(\mu_{1k} - \mu_{3k}\right)^2 \\ &+2 \sum_{i \in C(S_1,k)} \left(\mu_{1k} - \mu_{3k}\right)\left(i - \mu_{1k}\right) + \sum_{i \in C(S_2,k)} \left(i - \mu_{2k}\right)^2 \\ &+ \sum_{i \in C(S_2,k)} \left(\mu_{2k} + N_1 - \mu_{3k}\right)^2 + 2 \sum_{i \in C(S_2,k)} \left(\mu_{2k} + N - \mu_{3k}\right)\left(i - \mu_{2k}\right) \\ = &\sum_{i \in C(S_1,k)} \left(i - \mu_{1k}\right)^2 + \sum_{i \in C(S_2,k)} \left(i - \mu_{2k}\right)^2 \\ &+ n_{1k}\left(\mu_{1k} - \mu_{3k}\right)^2 + n_{2k}\left(\mu_{2k} + N_1 - \mu_{3k}\right)^2 \end{aligned} \tag{6}$$

$$\begin{aligned} D_2^{3k} &= \frac{\sum_{i \in C(S_1,k)} \left(i - \mu_{3k}\right)^2 + \sum_{i \in C(S_2,k)} \left(i + N_1 - \mu_{3k}\right)^2}{\left(N_1 + N_2\right)\left(n_{1k} + n_{2k}\right)} \\ &= \frac{n_{1k} N_1 D_2^{1k} + n_{2k} N_2 D_2^{2k} + n_{1k}\left(\mu_{1k} - \mu_{3k}\right)^2 + n_{2k}\left(\mu_{2k} + N_1 - \mu_{3k}\right)^2}{\left(N_1 + N_2\right)\left(n_{1k} + n_{2k}\right)} \end{aligned} \tag{7}$$

Also, Given a vector of the forward strand

$$nv(S) = \left(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T\right)$$

and $N = n_A + n_T + n_C + n_G$, the vector of the corresponding reverse strand $R(S)$ can be represented as

$$\begin{aligned} nv(R(S)) = (&n_T, n_G, n_C, n_A, N + 1 - \mu_T, N + 1 - \mu_G, N + 1 - \mu_C, \\ &N + 1 - \mu_A, D_2^T, D_2^G, D_2^C, D_2^A). \end{aligned} \tag{8}$$

It is worth noticing that we should not only take the complementary chain but also reverse it since the orientation of two strands is opposite.

2.4 K-mer natural vectors and their properties in chromosome fusion

The k-mer natural vector method is an extension of the natural vector method which considers k-mers instead of nucleotides as basic elements in sequences Wen et al. (2014b); Sun et al. (2021). K-mer is a string composed of k nucleotides and there are 4^k possible k-mers (denoted by $l_1, \ldots, l_{4^k}$). For the sequence $S = s_1 s_2 \ldots s_n$, we can regard it as a sequence consisting of $n - k + 1$ k-mers $(s_1 \ldots s_k) \ldots (s_{n-k+1} \ldots s_n)$. Similar to traditional natural vectors, we can define the k-mer natural vector

$$nv_k(S) = \left(n_{l_1}, \ldots, n_{l_k}, \mu_{l_1}, \ldots, \mu_{l_k}, D_2^{l_1}, \ldots, D_2^{l_k}\right).$$

If $n_{l_i} = 0$, we let $\mu_{l_i} = D_2^{l_i} = 0$. In the case of traditional natural vectors, we can compute $nv_k(S_3)$ (where $S_3 = S_1 + S_2$) using $nv_k(S_1)$ and $nv_k(S_2)$. When dealing with k-mer natural vectors, we can follow a similar process, but it's important to note that the results obtained from formulas (5) and (7) are no longer exact but approximate.

Let's take a 2-mer example for clarity. If we have the sequence $S_1 = s_1 s_2 \ldots s_n$ and the sequence $S_2 = t_1 t_2 \ldots t_n$, we can consider them as $(s_1 s_2) \ldots (s_{n-1} s_n)$ and $(t_1 t_2) \ldots (t_{n-1} t_n)$, respectively. The results obtained from formulas (5) and (7) represent the natural vector of $(s_1 s_2) \ldots (s_{n-1} s_n)(t_1 t_2) \ldots (t_{n-1} t_n)$ instead of $(s_1 s_2) \ldots (s_{n-1} s_n)(s_n t_1)(t_1 t_2) \ldots (t_{n-1} t_n)$ that corresponds to $S_1 + S_2$. However, given the considerable length of chromosomes, the difference introduced by a single k-mer becomes negligible. Therefore, we can still apply the previous formulas to perform the calculations effectively.

Previous studies have indicated that the optimal value for k should fall within the range of [*ceil* (*log*$_4$*min* (*LS*)), *ceil* (*log*$_4$*max* (*LS*)) + 1] where *LS* represents the set of lengths of genetic sequences in the study Wen et al. (2014b). For the datasets under consideration, the optimal k to extract the information of the sequences ranges from 13 to 15. However, this k is excessively large and does not fully leverage the time complexity advantage of our algorithm. Therefore, in this paper, we set $k = 10$, the smallest k to ensure that the algorithm avoids errors on our datasets.

2.5 The assignment problem and the Kuhn-Munkres algorithm

An assignment problem represents a specific instance of the more general transportation problem. In this particular case, the goal is to assign a set of resources to an equal number of activities while minimizing the total cost or maximizing the total profit of the allocation. To elaborate further, the problem can be formally stated as follows: given an $n \times n$ matrix $M = (m_{ij})$, we aim to determine an optimal permutation $p_1, p_2, \ldots, p_n$ from the set $1, 2, \ldots, n$ in order to minimize or maximize the objective function $\sum_{i=1}^{n} m_{ip_i}$.

Simply enumerating all possible permutations is feasible for small values of n. However, for larger values of n, this approach becomes computationally expensive and impractical as there are $n!$ possible permutations. In such cases, the Kuhn-Munkres algorithm, also known as the Hungarian method, offers an efficient solution to this problem Kuhn (1955, 1956); Munkres (1957).

The Kuhn-Munkres algorithm is a combinatorial optimization algorithm that can solve the assignment problem in polynomial time. The original version of the algorithm has a time complexity of $O(n^4)$, but later improvements have reduced it to $O(n^3)$ Edmonds and Karp (2003); Tomizawa (1971).

In Algorithm 1, we demonstrate how to minimize the objective function given matrix M by the Kuhn-Munkres algorithm:

```
Subtract the minimum entry in each row from all other
entries in the same row.
Subtract the minimum entry in each column from all other
entries in the same column.
while There are no $m$ lines (rows or columns) that cover
all zeros, where $m < n$do
    Find the minimum entry not covered by any line and its
    value is $e$.
    Subtract $e$ from each uncovered row and add $e$ to each
    covered column.
end while
We can select $n$ zeros with distinct rows and columns,
which corresponds to the optimal choice.
```

Algorithm 1. The Kuhn-Munkres algorithm.

2.6 The algorithm to recognize chromosome fusion

In order to transform the chromosome fusion problem into an assignment problem, we need to define a good measure for the similarity between chromosomes. One straightforward approach is to employ the Euclidean distance between the k-mer natural vectors of the chromosomes. However, sequencing chromosomes can introduce substantial errors and lead to length variation. Directly using the Euclidean distance might be problematic as the Euclidean distance between natural vectors is length-sensitive which could amplify the errors. To mitigate the effects of sequence length variations, we propose two distinct measures to dissociate the impact of the length from the k-mer patterns:

$$D_1(a,b) = 1 - \max\left(\cos\angle(nv_K(a), nv_K(b)),\ \cos\angle(nv_K(a), nv_K(R(b)))\right) \tag{9}$$

$$D_2(a,b) = |\text{length}(a) - \text{length}(b)| \tag{10}$$

In the above equations, $\angle(.,.)$ represents the angle between the two vectors. In D_1, we need to take both two strands into account so $R(b)$ should also be considered.

It is worth noticing that formula (9) will be slightly modified for fused chromosomes because there are eight possible representations for fused chromosomes, as opposed to the two representations for normal chromosomes. Assuming the fused chromosome is $\tilde{b}_1$ from the set $v_1, \ldots, v_8$, then formula (9) is adjusted as follows:

$$D_1\left(a_i, \tilde{b}_1\right) = 1 - \max_{j=1,\ldots,8}\left(\cos\angle\left(nv_K(a_i), nv_K\left(v_j\right)\right)\right). \tag{11}$$

Given two sets of chromosomes of equal count, $A = \{a_1, a_2, \ldots, a_n\}$ and $\tilde{B} = \{\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n\}$. The correspondence between these sets can be denoted by $(p_1, \ldots, p_n)$, a permutation of $\{1, \ldots, n\}$. That is, $\tilde{b}_{p_i}$ is similar to a_i. The task of establishing chromosome correspondence can be formulated as an optimization problem:

$$L\left(A, \tilde{B}\right) = \min_{p_1,\ldots,p_n} \sum_{i=1}^{n} \left(\sum_{l=1}^{2} N\left(D_l\left(a_i, \tilde{b}_{p_i}\right) | D_l(a_i, .)\right) \right) \tag{12}$$

where

$$N(x|A) = \frac{x - min(A)}{max(A) - min(A)} \tag{13}$$

serves as a normalization function to balance the importance of D_1 and D_2 with different orders of magnitude.

Eq 12 can be comprehended from two perspectives. First, by solving this optimization problem with the help of the Kuhn-Munkres algorithm, we can determine the optimal correspondence between the two chromosome sets. Second, the defined function L can serve as a metric, indicating the closeness of the two chromosome sets. Considering various possible fusion that convert $B = \{b_1, b_2, \ldots, b_n, b_{n+1}\}$ into $\tilde{B} = \{\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n\}$, the fusion resulting in the smallest loss function L would be our desired transformation.

To provide a clearer representation of the algorithm's process, we briefly summarize it in the following Algorithm 2. The code can be found in https://github.com/BobYHY/Fusion/.

```
Calculate $k$-mer natural vectors for all chromosomes.
Calculate $D_1$ and $D_2$ for all pairs of chromosomes, including potential new ones resulting from fusion.
for $j_1 = 1$ to $n + 1$ do
  for$j_2 = j_1 + 1$ to $n + 1$do
    Fuse $b_{j_1}$ and $b_{j_2}$ in eight possible ways ($v_1, v_2, \ldots, v_8$) and calculate their $k$-mer natural vectors using Eqs 5, 7. The resulting set after fusion is denoted as $\tilde{B}_{j_1 j_2}$.
    Calculate $L_{j_1 j_2} = L(A, \tilde{B}_{j_1 j_2})$ using the Kuhn-Munkres algorithm given the precomputed $D_1$ and $D_2$.
  end for
end for
The smallest $L_{j_1 j_2}$ corresponds to the most likely fusion.
```

Algorithm 2. Chromosome Fusion Recognition Algorithm.

Suppose each chromosome has a length of $O(l)$, then the time complexity of the k-mer natural vector algorithm is $O\ (nl)$. Next, measuring the distances between all pairs of chromosomes, including existing chromosomes and potentially new ones resulting from fusion, has a time complexity of $O\ (n^3 4^k)$. The Kuhn-Munkres algorithm has a complexity of $O\ (n^3)$, and it needs to be run for all possible fusion scenarios. Therefore, the overall complexity is $O\ (nl + n^3 4^k + n^5)$.

2.7 Multidimensional scaling

Multidimensional scaling (MDS) is a technique for visualizing the similarity between individual cases within a dataset Mead (1992). Its underlying concept is quite straightforward: how to find a set of points in a plane in such a way that the distances between them closely approximate a given distance matrix. More precisely, if we have a distance matrix $D = (d_{ij})$ for n chromosomes and we wish to map them to positions $x_1, \ldots, x_n$ on a plane, then MDS formulates an optimization problem to minimize the following expression:

$$f(x_1, \ldots, x_n) = \sum_{i \neq j} \left(d_{ij} - \|x_i - x_j\| \right)^2. \tag{14}$$

It is worth noting that in the earlier process of identifying chromosomal fusions, we did not calculate the distances within the same chromosome group. In fact, we can employ the normalized distances as defined in Eq. 12 for cases where A and B are the same sets, and then symmetrized the results to obtain the distances within the chromosome groups.

2.8 Synteny analysis

Synteny plots were generated using the MCScan module from the jcvi library Tang et al. (2008a). The data used for this analysis included chromosome sequences and GTF annotation data. The 'minspan' parameter was set to 50 to control the minimum span of syntenic blocks in the analysis. The synteny plot provides a visual representation of conserved gene order and genomic rearrangements between different species.

3 Results and discussion

3.1 Chromosome fusion recognition for human/gorilla and swamp buffalo/river buffalo

We employed our algorithm to identify chromosome fusion in human/gorilla and swamp buffalo/river buffalo datasets. It's worth noting that we did not include sex chromosomes in our analysis for two main reasons. First, the presence of palindromes in sex chromosomes, especially in Y chromosome, complicates its sequencing and results in a higher error rate compared to other chromosomes Trombetta and Cruciani (2017). Second, identifying sex chromosomes in the XY pair is straightforward due to their unequal lengths, obviating the need for explicit matching.

The results of our algorithm reveal that the gorilla chromosomes 2A and chromosome 2B have fused into a single sequence, aligning with human chromosome 2. Additionally, the river buffalo chromosome 4 and chromosome nine have fused into a single sequence, aligning with the swamp buffalo chromosome 1. This outcome is consistent with the data annotations. After identifying the correct fusion scenarios, all chromosomes can also find their corresponding chromosomes in the other set.

It is worth noting that in our algorithm, chromosome pairing between the two sets is achieved globally by minimizing the total loss. This means that at the algorithmic level, we do not require the paired sequences to be each other's nearest neighbors Cover and Hart (1967). (In this context, 'near' refers to a smaller pairing loss.) This design enhances the algorithm's robustness, preventing scenarios where multiple sequences might share the same nearest neighbor due to other mutations, thus avoiding situations that could disrupt the one-to-one correspondence. However, in terms of the results, almost all pairings meet the nearest neighbor condition. All swamp buffalo chromosomes match their nearest neighbors, while all human chromosomes except one have their counterparts as nearest neighbors. The only exception is human chromosome 17, which has its corresponding counterpart as the second nearest neighbor. This exception aligns with the reality. Figures 2, 3 display the synteny analysis of chromosomes for human/gorilla and swamp buffalo/river buffalo using the MCScan method. It is

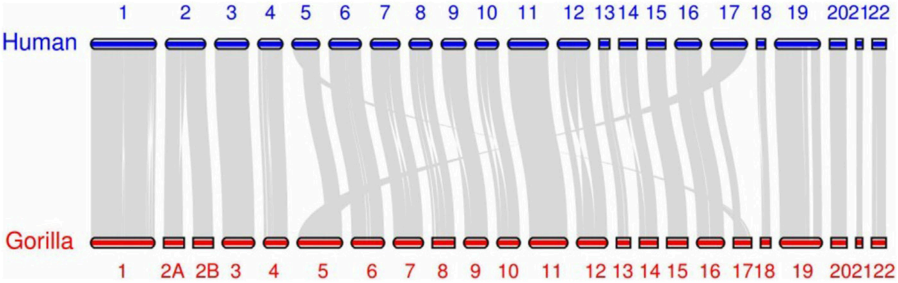

FIGURE 2
Synteny plots based on MCScan human/gorilla.

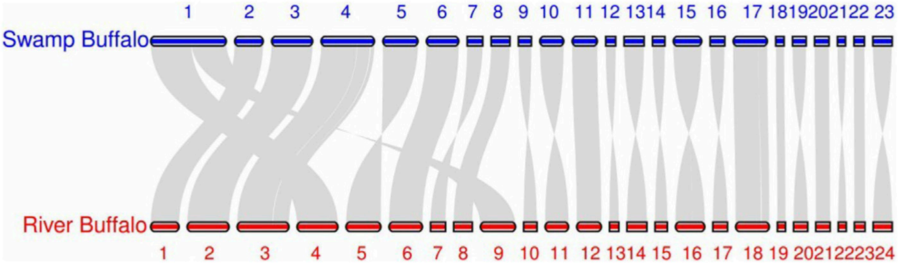

FIGURE 3
Synteny plots based on MCScan swamp buffalo/river buffalo.

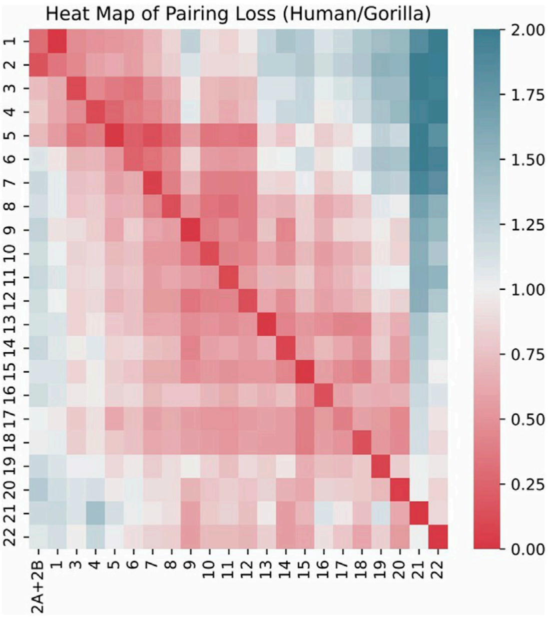

FIGURE 4
Heat map of pairing loss for chromosomes human/gorilla.

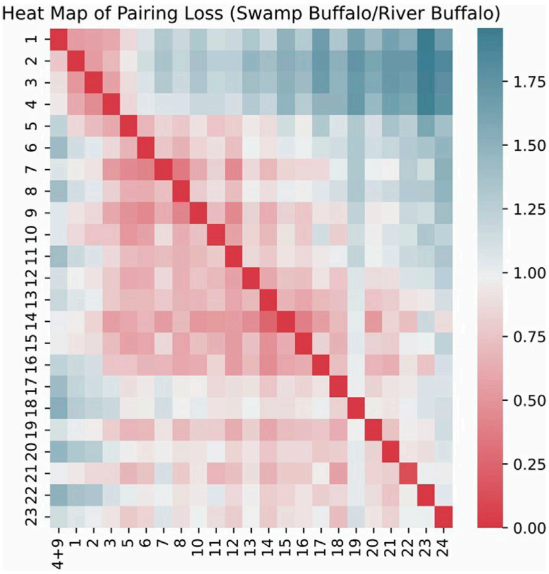

FIGURE 5
Heat map of pairing loss for chromosomes swamp buffalo/river buffalo.

evident that, in the case of human chromosome 17, due to some non-fusion structural variations, a portion of it actually originates from gorilla chromosome 5.

We visualized the pairing loss between chromosomes in two sets after fusion using two different approaches. In Figures 4, 5, we employed the conventional heatmap representation. The smaller the

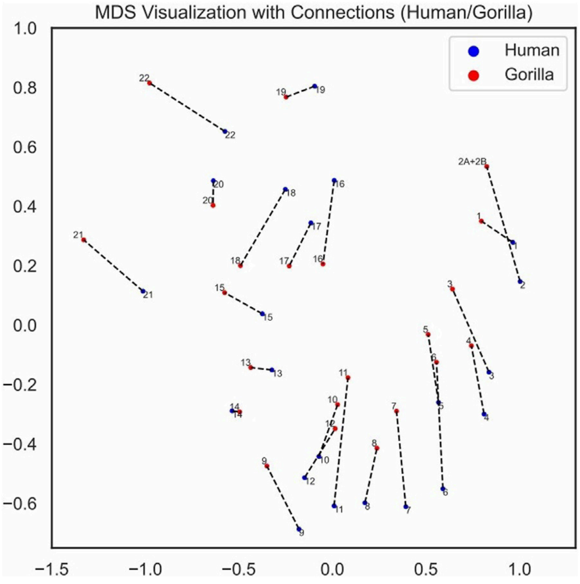

FIGURE 6
MDS visualization of chromosomes with pairing connections human/gorilla.

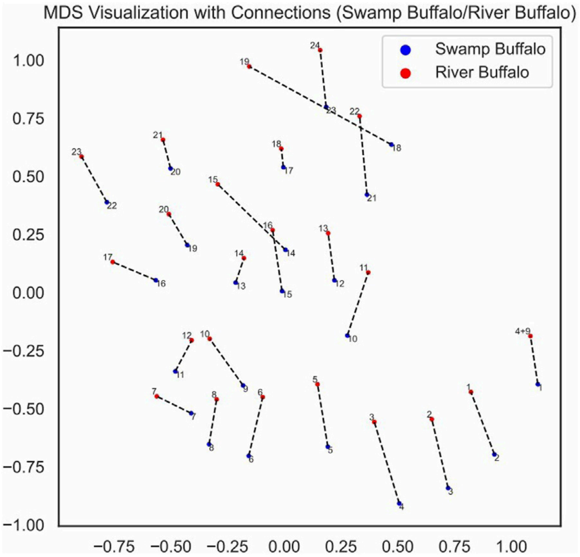

FIGURE 7
MDS visualization of chromosomes with pairing connections swamp buffalo/river buffalo.

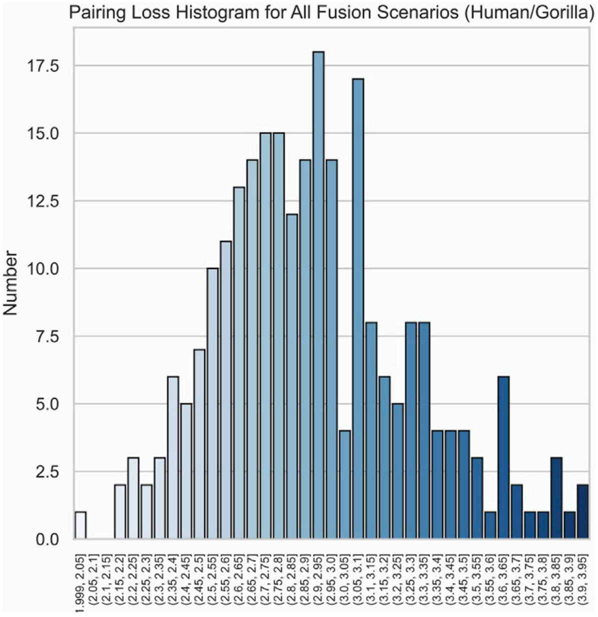

FIGURE 8
Histograms of pairing loss for all possible fusion scenarios human/gorilla.

pairing loss, the redder the corresponding square. In Figures 6, 7, we used multidimensional scaling (MDS) to project chromosomes as points onto a two-dimensional plane. The distances in the image can to some extent reflect the magnitude of the pairing loss. This representation offers greater intuitiveness, yet it's important to note that, since MDS involves the projection of high-dimensional information onto a two-dimensional plane, the distances in the graph may contain discrepancies compared to actual distances. Employing both of these methods effectively demonstrates our ability to accurately measure the differences between chromosomes using alignment-free features.

3.2 Chromosome fusion recognition for synthetic fusion scenarios

In order to further demonstrate the effectiveness of our method, we conducted recognition in synthetic fusion scenarios. Specifically, we transformed the swamp buffalo/river buffalo dataset to generate a large number of fusion scenarios for testing. Initially, we fused chromosome 4 and chromosome nine of river buffalo in the correct manner. Subsequently, we artificially fused two random chromosomes from swamp buffalo, resulting in a new scenario where river buffalo still has one more chromosome than swamp buffalo.

Given that swamp buffalo has 23 autosomes, we obtained 253 possible fusion scenarios. (Different orientations and fusion orders are algorithmically equivalent.) We conducted chromosome fusion recognition on each of these 253 scenarios, successfully identifying the fusion pairing and accurately correlating the remaining chromosomes 252 times. In other words, our method achieved an accuracy rate of 99.6%, demonstrating its high performance.

3.3 Effectiveness and efficiency of the algorithm

Due to the fact that a chromosome has two complementary strands, when sequencing closely related species, we cannot ensure that the corresponding chromosome strands are the same. This is precisely the case with the swamp buffalo/river buffalo dataset. Therefore, our algorithm incorporates the consideration of

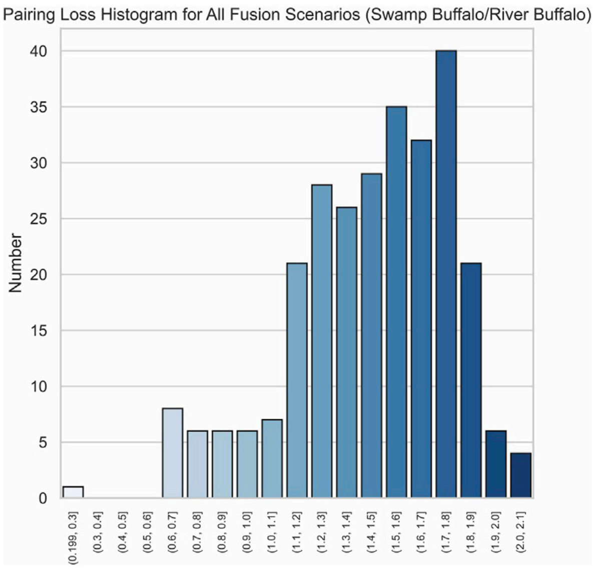

FIGURE 9
Histograms of pairing loss for all possible fusion scenarios swamp buffalo/river buffalo.

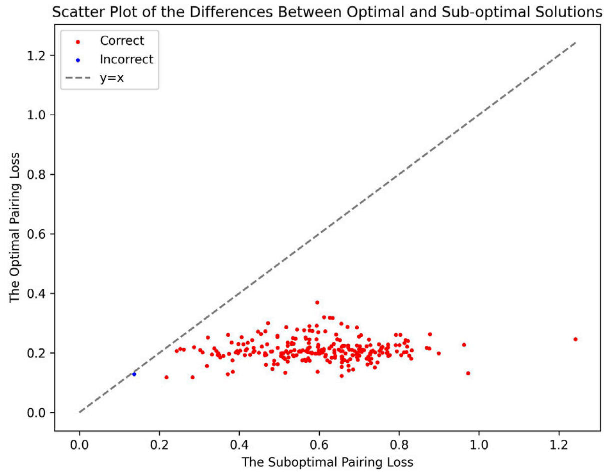

FIGURE 10
Scatter plot of the differences between optimal and sub-optimal solutions.

complementary strands. When computing the pairing loss, we opt for the strand with the minimal loss to address this issue.

We validated the effectiveness of our algorithm through two real fusion scenarios and 253 synthetic fusion scenarios. Among all scenarios, we encountered errors in only one synthetic case, showcasing an impressive level of accuracy.

The effectiveness of our method is further highlighted by the pronounced characteristics of the correct fusion scenarios. In real fusion, we observed that in the case of human/gorilla, the pairing loss for the optimal fusion is 2.03, whereas the pairing losses for the other 252 possible fusions range from 2.16 to 3.92. Notably, the average difference in pairing loss for non-optimal fusions is 7.0 × 10^{-3}, significantly smaller than the 0.13 difference between the optimal and sub-optimal fusions (Figure 8). Similarly, for the swamp buffalo/river buffalo dataset, the optimal fusion has a pairing loss of 0.26, while the pairing losses for the other 275 possible fusions range from 0.61 to 2.06. The average difference in pairing loss for non-optimal fusion is 5.3 × 10^{-3}, again significantly smaller than the 0.35 difference between the optimal and sub-optimal fusions (Figure 9). In Figure 10, we can also observe that in synthetic fusion scenarios, the optimal pairing loss are consistently much smaller than the sub-optimal solutions. The only exception where the optimal and sub-optimal pairing loss are relatively close is the case of the error, as previously mentioned.

From the above results, we can draw two conclusions. Firstly, in both real fusion scenarios and synthetic fusion scenarios, the computed optimal loss is significantly smaller than the sub-optimal loss. This reflects the robustness of the algorithm, meaning that perturbing the original data won't immediately change the optimal solution. This underscores the recognizability of the features associated with the correct fusion event. Secondly, the gap between the optimal and sub-optimal solutions can reflect the reliability of the results. Generally, correct identification is usually associated with a significant difference between these two values. Conversely, if they are very close, it may indicate potential issues with the results. Additionally, it can be observed that, compared to the recognition between the two types of buffalo, the identification of human/gorilla is relatively less reliable. This is attributed to the presence of another significant chromosomal structural variation in this example, namely, the exchange between chromosome five and chromosome 17 (Figure 2).

This algorithm does not require alignment and is therefore faster than previous methods. In real fusion scenarios, conducting synteny analysis for human/gorilla and swamp buffalo/river buffalo using MCScan took 1,052 and 993 s, respectively. Using our algorithm, computing *k*-mer natural vectors took 182 and 185 s respectively, and determining the most probable fusion scenarios with the algorithm took 271 and 293 s respectively (CPU 3.10GHz, 8C16T). (We use parallel computing to calculate each natural vector independently.) In synthetic fusion scenarios, we do not need to calculate natural vectors separately, and the average time spent on determining the most probable fusion scenarios is 265 s.

It's worth mentioning that, in fact, we can disregard the time required for computing *k*-mer natural vectors. We can precompute the natural vectors and simply read them when comparing with other organisms. This is because if there are M organisms, the calculation of natural vectors only needs to be performed $O(M)$ times. However, the comparisons require $O(M^2)$ operations. Therefore, it is reasonable to focus solely on the fusion identification time. This fact can also be observed in synthetic scenarios, where natural vectors are all precomputed. Furthermore, even when including the time spent computing *k*-mer natural vectors, our algorithm remains faster.

Another noteworthy point is that in MCScan, our analysis is limited to experimentally determined CDS sequences, which account for only about 10% of the entire genome. This analysis relies heavily on manual experimental annotation and utilizes incomplete information. If one needs to segment the entire chromosome, the time required would significantly increase. In contrast, our method does not require annotation and allows for a rapid analysis of the entire chromosome.

4 Conclusion

In this study, we propose an alignment-free algorithm based on natural vectors and the Kuhn-Munkres algorithm to address the challenge of chromosome fusion recognition. Our approach offers a fresh perspective on understanding chromosome fusion phenomena. Previously, most alignment-free methods struggled to tackle the intricate issue of chromosome internal structures, while our method demonstrates significant improvements.

Our method has two main advantages. Firstly, our algorithm demonstrates a significant speed advantage, being about four times faster than synteny-based methods for datasets we consider. This allows for efficient data processing while maintaining high accuracy. Secondly, it considers whole chromosomes instead of segments, eliminating the need for manual selection of segment boundaries and additional annotation data, making the algorithm more automated.

However, our method still has limitations. It is primarily designed for fusion recognition and cannot detect other non-fusion structural variations, such as repeats. Additionally, in situations where multiple fusions occur, the speed advantages may diminish. In future studies, we aim to incorporate heuristic search designs to further enhance the algorithm's speed, especially in scenarios involving multiple fusions, while maintaining the accuracy of the algorithm. Additionally, we aspire to identify alignment-free features for more localized chromosome structural variations.

Author's Note

Dedicated to Dr. Henry Laufer on the occasion of his 80th birthday.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/.

Author contributions

HY: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft. SY: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research is supported by National Natural Science Foundation of China (NSFC) grant (12171275) and Tsinghua University Education Foundation fund (042202008).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1364951/full#supplementary-material

References

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *Basic local aligment search Tool.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

Calabrese, P., Chakravarty, S., and Vision, T. (2003). Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinforma. Oxf. Engl.* 19 (Suppl. 1), i74–i80. doi:10.1093/bioinformatics/btg1008

Cameron, D., Baber, J., Shale, C., Valle-Inclan, J., Besselink, N., Hoeck, A., et al. (2021). Gridss2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol.* 22, 202. doi:10.1186/s13059-021-02423-x

Cicconardi, F., Lewis, J., Martin, S., Reed, R., Danko, C., and Montgomery, S. (2021). Chromosome fusion affects genetic diversity and evolutionary turnover of functional loci but consistently depends on chromosome size. *Mol. Biol. Evol.* 38, 4449–4462. doi:10.1093/molbev/msab185

Cover, T. M., and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi:10.1109/tit.1967.1053964

Deng, M., Yu, C., Liang, Q., He, R. L., and Yau, S. S.-T. (2011). A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS ONE* 6, e17293. doi:10.1371/journal.pone.0017293

Edmonds, J., and Karp, R. (2003). Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM* 19, 248–264. doi:10.1145/321694.321699

Feulner, P., and De-Kayne, R. (2017). Genome evolution, structural rearrangements and speciation. *J. Evol. Biol.* 30, 1488–1490. doi:10.1111/jeb.13101

Guerrero, R., and Kirkpatrick, M. (2014). Local adaptation and the evolution of chromosome fusions. *Evolution* 68, 2747–2756. doi:10.1111/evo.12481

Haas, B., Delcher, A., Wortman, J., and Salzberg, S. (2005). Dagchainer: a tool for mining segmental genome duplications and synteny. *Bioinforma. Oxf. Engl.* 20, 3643–3646. doi:10.1093/bioinformatics/bth397

Hauffe, H., and Searle, J. (1998). Chromosomal heterozygosity and fertility in house mice (mus musculus domesticus) from northern Italy. *Genetics* 150, 1143–1154. doi:10.1093/genetics/150.3.1143

Iannuzzi, A., Parma, P., and Iannuzzi, L. (2021). The cytogenetics of the water buffalo: A review. *Animals open access J. MDPI* 11, 3109. doi:10.3390/ani11113109

Ijdo, J., Baldini, A., Ward, D., Reeders, S., and Wells, R. (1991). Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. U. S. A.* 88, 9051–9055. doi:10.1073/pnas.88.20.9051

Jun, S.-R., Sims, G., Wu, G., and Kim, S.-H. (2009). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. *Proc. Natl. Acad. Sci. U. S. A.* 107, 133–138. doi:10.1073/pnas.0913033107

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. (NRL)* 2, 83–97. doi:10.1002/nav.3800020109

Kuhn, H. W. (1956). Variants of the Hungarian method for assignment problems. *Nav. Res. Logist. (NRL)* 3, 253–258. doi:10.1002/nav.3800030404

Layer, C. C. Q. A. R., Ryan, M., and Hall, I. M. (2014). Lumpy: a probabilistic framework for structural variant discovery. *Genome Biol.* 15, R84. doi:10.1186/gb-2014-15-6-r84

Liu, H., Yin, H., Li, G., Li, J., and Wang, X. (2021). Aperture: alignment-free detection of structural variations and viral integrations in circulating tumor dna. *Briefings Bioinforma.* 22, bbab290. doi:10.1093/bib/bbab290

Mead, A. (1992). Review of the development of multidimensional scaling methods. *Statistician* 41, 27. doi:10.2307/2348634

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *J. Soc. Industrial Appl. Math.* 5, 32–38. doi:10.1137/0105003

Poot, M., and Hochstenbach, R. (2021). Prevalence and phenotypic impact of robertsonian translocations. *Mol. Syndromol.* 12, 1–11. doi:10.1159/000512676

Qi, J., Wang, B., and Hao, B. (2004). Whole proteome prokaryote phylogeny without sequence alignment: a k -string composition approach. *J. Mol. Evol.* 58, 1–11. doi:10.1007/s00239-003-2493-7

Sinha, A. U., and Meller, J. (2007). Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinforma.* 8, 82. doi:10.1186/1471-2105-8-82

Sun, N., Pei, S., He, L., Yin, C., He, R. L., and Yau, S. S.-T. (2021). Geometric construction of viral genome space and its applications. *Comput. Struct. Biotechnol. J.* 19, 4226–4234. doi:10.1016/j.csbj.2021.07.028

Tang, H., Bowers, J., Wang, X., Ming, R., Alam, M., and Paterson, A. (2008a). Synteny and collinearity in plant genomes. *Sci. (New York, N.Y.)* 320, 486–488. doi:10.1126/science.1153917

Tang, H., Wang, X., Bowers, J., Ming, R., Alam, M., and Paterson, A. (2008b). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18, 1944–1954. doi:10.1101/gr.080978.108

Tian, K., Zhao, X., and Yau, S. (2018). Convex hull analysis of evolutionary and phylogenetic relationships between biological groups. *J. Theor. Biol.* 456, 34–40. doi:10.1016/j.jtbi.2018.07.035

Tomizawa, N. (1971). On some techniques useful for solution of transportation network problems. *Networks* 1, 173–194. doi:10.1002/net.3230010206

Trombetta, B., and Cruciani, F. (2017). Y chromosome palindromes and gene conversion. *Hum. Genet.* 136, 605–619. doi:10.1007/s00439-017-1777-8

Vara, C., Paytuví Gallart, A., Cuartero, Y., Álvarez González, L., Marin, L., García, F., et al. (2021). The impact of chromosomal fusions on 3d genome folding and recombination in the germ line. *Nat. Commun.* 12, 2981. doi:10.1038/s41467-021-23270-1

Wang, Y., Tang, H., Debarry, J., Tan, X., Li, J., Wang, X., et al. (2012). Mcscanx: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids Res.* 40, e49. doi:10.1093/nar/gkr1293

Wen, J., Chan, R., Yau, S.-T., He, R., and Yau, S. (2014a). K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* 546, 25–34. doi:10.1016/j.gene.2014.05.043

Wen, J., Chan, R. H.-F., Yau, S.-C., He, R. L., and Yau, S. S.-T. (2014b). K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* 546 (1), 25–34. doi:10.1016/j.gene.2014.05.043

Yunis, J., and Om, P. (1982). The origin of man: a chromosomal pictorial legacy. *Sci. (New York, N.Y.)* 215, 1525–1530. doi:10.1126/science.7063861