Farnia and Tahiri
Algorithms for Molecular Biology (2024) 19:22
https://doi.org/10.1186/s13015-024-00267-1

Algorithms for Molecular Biology

RESEARCH

Open Access

New generalized metric based on branch length distance to compare B cell lineage trees

Mahsa Farnia[1] and Nadia Tahiri[1*]

Abstract

The B cell lineage tree encapsulates the successive phases of B cell differentiation and maturation, transitioning from hematopoietic stem cells to mature, antibody-secreting cells within the immune system. Mathematically, this lineage can be conceptualized as an evolutionary tree, where each node represents a distinct stage in B cell development, and the edges reflect the differentiation pathways. To compare these lineage trees, a rigorous mathematical metric is essential. Analyzing B cell lineage trees mathematically and quantifying changes in lineage attributes over time necessitates a comparison methodology capable of accurately assessing and measuring these changes. Addressing the intricacies of multiple B cell lineage tree comparisons, this study introduces a novel metric that enhances the precision of comparative analysis. This metric is formulated on principles of metric theory and evolutionary biology, quantifying the dissimilarities between lineage trees by measuring branch length distance and weight. By providing a framework for systematically classifying lineage trees, this metric facilitates the development of predictive models that are crucial for the creation of targeted immunotherapy and vaccines. To validate the effectiveness of this new metric, synthetic datasets that mimic the complexity and variability of real B cell lineage structures are employed. We demonstrated the ability of the new metric method to accurately capture the evolutionary nuances of B cell lineages.

Keywords B cell lineage tree, Immunoinformatics, Generalized branch length distance, Immune repertoire

Introduction

Immunoglobulins (IG), instrumental in adaptive immunity, mediate antigen recognition and initiate sophisticated defense mechanisms against microorganisms [1]. B cells, essential components of the immune response, carry B cell receptors (BCRs) attached to their surfaces [2]. Somatic mutations [3] occurring in B cells are essential for creating a wide range of diverse naive B cell variants [4]. Somatic mutations [3] in B cells are essential for generating diverse naive B cell variants [4], which is vital for devising effective therapeutic approaches [5–8]. IG genes undergo somatic mutations, generating diverse functional genes and receptors for antigen recognition [9]. The diversity of receptors arises from distinct recombination events involving B cell segments [10].

The imperative to algorithmically discern the underlying principles governing B cell development and lineage diversification requires a well-defined framework. The complex dynamics inherent in this process demand a profound understanding of the temporal sequence and structural patterns of mutations in B cells, crucial for elucidating their ontogeny and response to antigens.

Let $\mathcal{T}$ be a set of observed BCR lineage trees [11–13] with identical somatic mutations. Each tree $T_i \in \mathcal{T}$ contains node labels that vary partially, including unmutated (naive) BCR IGH nodes. The utilization of lineage trees emerges as an indispensable computational tool for algorithmically unraveling the intricate evolutionary relationships embedded within B cell clones. A meticulous comparative analysis is composed to furnish nuanced

*Correspondence:
Nadia Tahiri
Nadia.Tahiri@USherbrooke.ca
[1] Department of Computer Science, University of Sherbrooke, 2500, boul. de l'Université, Sherbrooke, QC J1K 2R1, Canada

computational insights into various facets, including relatedness, clone diversity, antibody generation, memory B cell responses [14, 15], selection mechanisms, evolutionary patterns, and the key mechanisms governing B cell lineage development.

There is a notable gap in the scientific literature regarding the algorithmic exploration of comparing lineage trees. This paper introduces a novel quantitative metric grounded in the Jaccard index and Minkowski principles [16], establishing a theoretical foundation for this problem. Additionally, we propose algorithmic techniques optimized for efficiency to address this computational challenge rigorously. The introduced metric and algorithms collectively provide an analysis tool for B cell lineage trees, contributing to our understanding of the dynamic intricacies within the immune system.

Notation and preliminaries

In the subsequent section, we define key aspects of lineage trees, emphasizing their distinct features in comparison to phylogenetic trees [17] and clonal trees [18]. Topics include root identification, multifurcations resulting from B cell repertoire divergences, and the crucial role of cellular abundance in clonal expansion [19, 20], culminating in the introduction and application of a novel metric for lineage tree analysis.

Definitions

Different from phylogenetic trees, but similar to clonal trees lineage trees require a nuanced understanding for selecting appropriate metric methodologies [21] in scientific research. Identifying the root in the lineage tree is crucial, representing the unmutated sequence. Simultaneous divergences within the B cell repertoire result in multifurcations, generating zero-length branches or internal nodes of degree 2 [22]. This understanding refines scientific investigations, particularly in genetic lineage analysis.

Considering the potential coexistence of mutations, observed sequences may manifest as either *leaves* or *internal nodes* [20]. Therefore, a comprehensive analysis encompasses all tree nodes, underscoring the importance of considerations related to the *root* and *branch lengths*.

Cellular *abundance*, linked to clonal expansion based on antigen affinity, assumes a crucial role. Assessing distinct sequence variations (i.e., genotypes) aids in comprehending B cell evolution and clonal selection.

Considering these critical factors, we introduce a cutting-edge metric that adeptly incorporates node overlaps, branch lengths, Euclidean distance-based metrics, and lineage tree abundance. This innovative metric is applied for performance evaluation on a comprehensive dataset. Figure 1 provides a clear illustration of the intricate structure of a lineage tree, comprising a total of 20 nodes. This lineage tree includes a discerning naive node, 14 intricately detailed leaves, and 5 strategically positioned internal nodes. Notably, the size of each node reflects its abundance, while the lengths of the branches intricately portray the evolutionary period between consecutive nodes. A naive B cell (or naive node) generally signifies an unmutated progenitor cell that has not yet experienced somatic hypermutation or antigen-driven selection. To prevent complications arising from zero-length branches in visualizations, a naive node is assigned a distance. This convention ensures that the node is represented within the tree, even if it lacks subsequent branches or nodes.

Strong similarities with phylogenetic and clonal trees are easily observed. Therefore, in the next subsection, we have undertaken to comprehend these similarities as well as their differences.

Comparison between phylogenetic, lineage, and clonal trees

A systematic comparative examination of B cell lineage, clonal tumor and phylogenetic trees, elucidating shared attributes and distinctive features, is strongly needed to leverage advancements in phylogenetics within the field of immunology. Emphasizing the intricate interplay between somatic hypermutation and selection pressures within germinal centers, it underscores the imperative need for employing specialized methodologies to accurately reconstruct the evolutionary history of B cell lineage and tumor clonal trees. This analysis is presented in Table 1.

Table 1 illustrates the disparities (i.e., similarities and differences) between phylogenetic, lineage, and clonal trees. Phylogenetic trees focus on overall evolutionary relationships, representing genetic change through branch lengths and showcasing common ancestors. In contrast, lineage trees emphasize lineage splitting and evolution within specific groups, indicating the time or divergence through branch lengths and depicting lineage splitting events. The clonal theory of cancer, proposed by Nowell [18], views tumor development as an evolutionary process, which illustrates the history of somatic mutation acquisition. While phylogenetic trees may be rooted or unrooted [23], lineage trees and clonal trees are typically rooted, and they often exhibit multifurcations. Additionally, lineage trees and clonal trees consider both leaves and internal nodes. In lineage trees nodes incorporate abundance information, a feature not present in phylogenetic trees and clonal trees.

Somatic hypermutation, elucidated by Odegard et al. [24], aligns intricately with phylogenetic hotspots, as expounded by Tietje et al. [25], underscoring the nuanced interplay between sequence mutations and

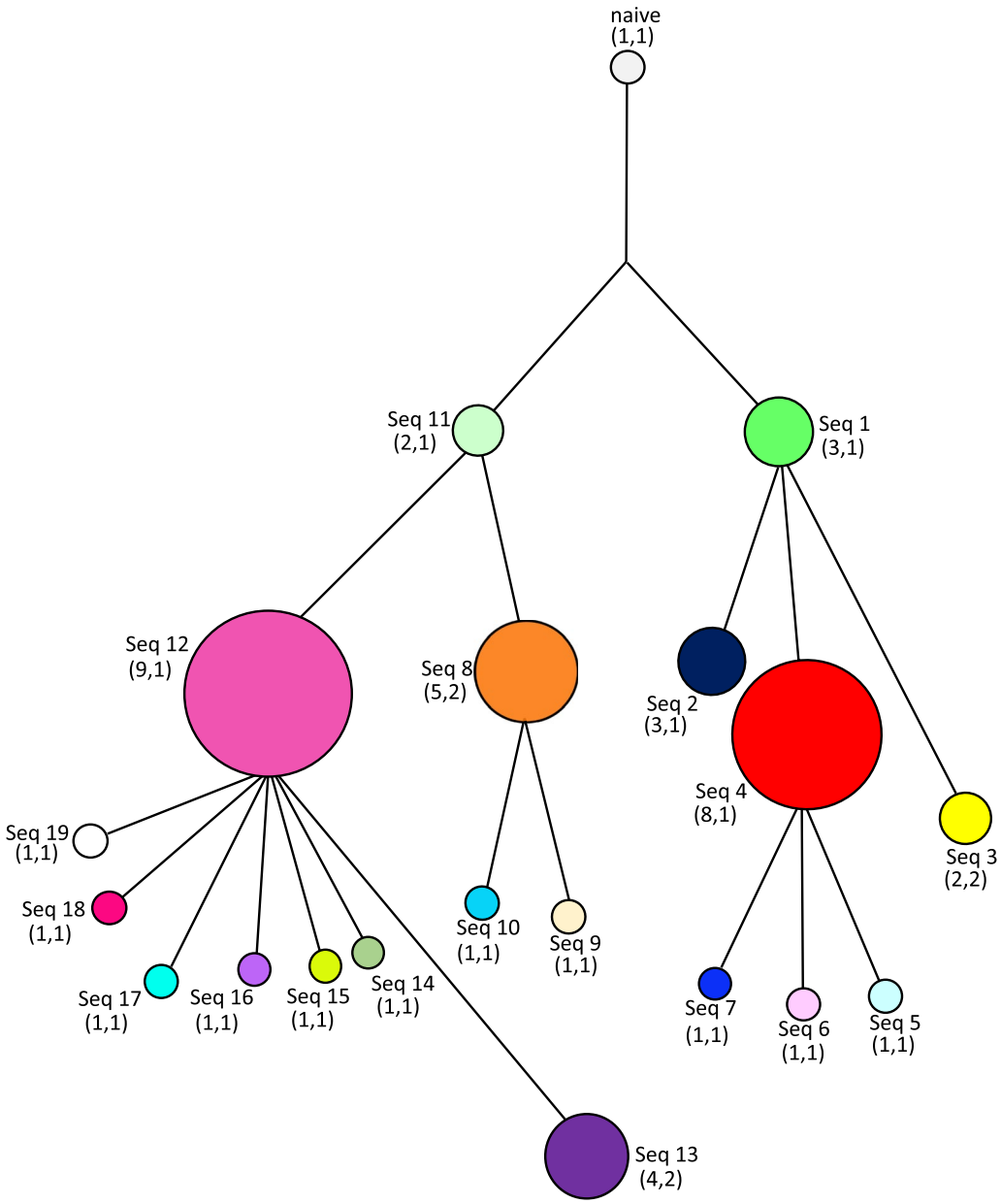

Fig. 1 Visual representation of a B cell lineage is presented, where each node corresponds to a sequence. The notation provided beneath each sequence corresponds to the (*w*, *d*) values associated with each node. The first value denoted by *w* is linked to the size of the node, reflecting the abundance of the corresponding sequence. The second value denoted by *d* represents the branch length associated with the node. The root is the node named naive, representing the unmutated naive B cell. Progressing through the branches of the lineage tree, the cells exhibit an increased affinity for the specific antigen. The distances between a node and its preceding one are directly associated with the *d* value assigned next to the node

their local contextual determinants. The discernible selection pressure operative within germinal centers [26, 27] serves to accentuate the convergence of B cell evolutionary processes with the foundational tenets of Darwinian evolution [28].

Notwithstanding these conceptual parallels with phylogenetic and clonal trees, the idiosyncratic features inherent in B cell lineage trees necessitate the application of specialized methodologies for the precise reconstruction of evolutionary lineage trees. The critical importance of abundance, reflecting both genotypic diversity and their

Table 1 Comparison of characteristics between phylogenetic trees, lineage trees, and clonal trees

Aspect	Phylogenetic trees	Lineage trees	Clonal trees
Structure	Evolutionary relationships among species	Cell lineage splitting and evolution	Tumor cell population evolution
Branch lengths	Often represent genetic change between species	Indication of time or divergence between updates of cells	Evolutionary time or the accumulation of genetic changes
Nodes	Common ancestors and relationships	Lineage splitting events or divergence	Distinct clones of tumor cells
Rooting	Rooted or unrooted	Typically rooted (naive cell)	Rooted (a single progenitor normal cell)
Focus	Overall evolutionary relationships among different species	Lineage evolution within a specific group of cells	Somatic mutation acquisition in tumor evolution
Degree of internal nodes	More than three	More than two (known as clonal expansion)	More than two
Sequences observed	Leaves	Leaves and internal nodes	Leaves and internal nodes
Multifurcations	Rare (biologically)	Common	Common
Abundance/weight	No	Yes	No

corresponding frequencies, emerges as a focal point for a nuanced understanding of B cell evolution and the dynamics of clonal selection.

B cell repertoires, marked by simultaneous divergences [29, 30], manifest as multifurcations or zero-length branches [31]. Considering potential mutations, sequences are categorized as leaves or internal nodes, including nodes with a degree of two.

The unmutated sequence of the naive B cell serves as the root in the lineage tree, establishing a distinctive feature when compared to traditional phylogenetics. Acknowledging the central role played by the unmutated sequence is foundational for meticulously tracking the progression of mutations and reconstructing the intricate trajectories of B cell lineages.

Related works

As the volume of biological data increases, the complexity of phylogenetic tree topologies also grows. Consequently, it becomes challenging, and sometimes infeasible, to discern differences between them. Many researchers have proposed methods for comparing phylogenetic trees that focus on their properties rather than their topologies. One new space of phylogenetic trees (wald space) is developed by [32]. Wald space is suitable for the statistical analysis of phylogeny, utilizing a geometry founded on more biologically principled assumptions than existing spaces is developed. A polynomial-time algorithm is developed to complete phylogenetic trees and calculate the distance between trees that are defined on different, yet overlapping, sets of taxa [33]. The principles of probabilistic phylogenetic distances are extended to calculate tree distances under models of continuous trait evolution along phylogeny [34]. Distance measurement using Monte Carlo methods, which relies on the probability distributions of genetic sequence data derived from phylogenetic trees, is a method considered for this analysis [35].

The development of dissimilarity measures for comparing clonal cancer trees has become a central focus among computational researchers. Recent advancements in this area include the GraPhyC method by Govek et al. [36], which employs a distance measure to derive consensus tumor histories, thereby enriching the toolkit for analyzing clonal evolution. More recently, DiNardo et al. [37] introduced the Common Ancestor Set (CASet) distance and the Distinctly Inherited Set Comparison (DISC) distance, both designed to account for subclonal mutations. In the same year, Llabrés et al. [38] proposed a distance metric for multi-labeled trees that extends the Robinson–Foulds distance. Building on this foundation, Jahn et al. [39] introduced Bourque distances, offering another generalization of the Robinson–Foulds metric. Finally, Khayatian et al. [40] further advanced the field by developing k-Robinson-Foulds dissimilarity measures specifically for labeled tree comparisons.

Methods

Branch Length Distance known as *BLD* is one well-established metric for comparing phylogenetic trees based on branch length distances, [41–44]. It focuses on the differences in the lengths of branches, which is the base of our metric. By comparing branch lengths, the *BLD* metric provides a quantitative measure of how similar or different the phylogenetic trees are, making it a valuable tool for analyzing the evolutionary relationships represented in the phylogenetic trees.

$$BLD_{(T_1,T_2)} = \sum_{i=1}^{TN_{(T_1,T_2)}} (d_{T_1}(i) - d_{T_2}(i))^2, \quad (1)$$

where $d_{T_1}(i)$ and $d_{T_2}(i)$ are the vectors of the branch lengths between node i and the lowest common ancestor (LCA) of the node i in the phylogenetic trees T_1 and T_2, respectively. LCA of any pair of nodes i and j in a tree, denoted as LCA(i, j), is defined as the deepest common ancestor of both i and j. Each node is considered a descendant of itself, so if node i is directly connected to node j, then j is the LCA of i. Since *BLD* is tailored for phylogenetic trees, considering all the essential characteristics of lineage trees can lead to the development of a more appropriate metric. However, *BLD* (Eq. 1) has notable limitations, such as its failure to account for internal nodes, node abundance, and overlapping sets of leaves. To address these shortcomings, we propose an extension for *BLD*.

The study introduces an innovative metric methodology that provides a precise approach for comparing the optimal quantity of lineage trees. The metric approach is based on Minkowski and Jaccard principles.

Definition 1 (*Minkowski Distance*) Minkowski distance, characterized by order h, such that $h \in \mathbb{N}^+$, between two points $X = (x_1, ..., x_n)$ and $Y = (y_1, ..., y_n) \in \mathbb{R}^n$ is described as:

$$D(X, Y) = \sqrt[h]{\sum_{i=1}^{n} |x_i - y_i|^h}. \quad (2)$$

Definition 2 (*Jaccard Distance*) Jaccard distance quantifies the dissimilarity between sample sets and complements the Jaccard coefficient. It is defined as:

$$d_j(A, B) = 1 - J(A, B) = 1 - \frac{A \cap B}{A \cup B}, \quad (3)$$

where J(A,B) is Jaccard index, and A and B are two sets.

Evaluating whether a mathematical structure or space is a metric involves verifying that it satisfies certain properties. In this context, a specific definition is presented to enhance comprehension and clarity.

Definition 3 (*Metric properties*) A function $d : X \times X \to \mathbb{R}^+$ is a *metric* on the set X if it verifies, for any x, y, z in X, the following properties:

1. Identity/separation: $d(x, y) = 0$ if and only if $x = y$
2. Symmetry: $d(x, y) = d(y, x)$
3. Positive: $d(x, y) \geq 0$
4. Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$

Theorem 1 *The Minkowski distance is a metric, satisfying identity, non-negativity, symmetry, and the triangle inequality properties.*

Our primary objective is to expand the scope of our calculations to encompass all specified criteria. The Euclidean distance [45] emerges as a robust metric for measuring distances between data points within a dataset, delineating the straight-line distance between points and providing an intuitive measure of their similarity or dissimilarity. This metric functions as a specific instance within the paradigm of Minkowski distance, becoming apparent with the parameter h adjusted to 2.

In the context of lineage trees, Euclidean distance takes on a distinctive characterization, defined as follows:

$$D_{(T_1,T_2)} = \frac{1}{TN_{(T_1,T_2)}} \sqrt{\sum_{i=1}^{TN_{(T_1,T_2)}-1} \sum_{j=i+1}^{TN_{(T_1,T_2)}} \left| d_{T_1}(i,j) - d_{T_2}(i,j) \right|^2}, \quad (4)$$

where T_1 and T_2 represent lineage tree 1 and lineage tree 2, respectively, and $TN_{(T_1,T_2)}$ is the total number of nodes in T_1 and T_2. The distances between nodes i and j are denoted as $d_{T_1}(i,j)$ and $d_{T_2}(i,j)$, representing the spatial separation between nodes i and j in lineage trees T_1 and T_2 respectively.

Two lineage trees must have common nodes to be comparable. If node i is missing in lineage tree T_1 but present in T_2, a ghost node i with a weight and branch length of zero is added to T_1. This ensures that the branch length distances involving node i in T_2 are considered when calculating the distance between these two lineage trees. Including such absent nodes helps maintain their impact across all the lineage trees under comparison.

In addition to the difference in distances between all pairs of nodes $D_{(T_1,T_2)}$, the difference in abundances for each node between two trees $W_{(T_1,T_2)}$ is added. Manhattan distance, derived from the Minkowski metric with parameter h fixed at 1, serves well for this task as it highlights discrepancies in dimensions, contrasting with Euclidean distance which measures the straightforward distance between points. Utilizing both distances offers a detailed understanding of the data. The Manhattan distance emphasizes subtle differences in node attributes, whereas the Euclidean distance provides a comprehensive view of the structural variations between family trees, thereby enhancing the overall analysis and interpretation.

$$W_{(T_1,T_2)} = \frac{1}{TN_{(T_1,T_2)}} \sum_{i=1}^{TN_{(T_1,T_2)}} \left| w_{T_1}(i) - w_{T_2}(i) \right|, \quad (5)$$

where $w_{T_1}(i)$ and $w_{T_2}(i)$ represent the weights (i.e., abundances) of node i in T_1 and T_2, respectively.

Another criterion has been effectively integrated and fine-tuned to function as a penalty [46, 47] between two lineage trees, offering a more advantageous assessment by considering the ratio of uncommon nodes to the total number of nodes.

$$P_{(T_1,T_2)} = 1 - \frac{CN_{(T_1,T_2)}}{TN_{(T_1,T_2)}}, \quad (6)$$

where $CN_{(T_1,T_2)}$ is the number of common nodes between T_1 and T_2.

As the presented method extends beyond the traditional *BLD*, it is referred to as the Generalized Branch Length Distance (*GBLD*).

The *GBLD* between T_1 and T_2 is defined as follows:

$$GBLD_{(T_1,T_2)} = P_{(T_1,T_2)} + W_{(T_1,T_2)} + D_{(T_1,T_2)}. \quad (7)$$

Remark 1 Lineage trees share structural similarities with rooted phylogenetic trees, justifying the application of a criterion that requires a minimum number of shared nodes for comparative analysis. Equation 7 applies to compare the lineage trees sharing a minimum of three common nodes, i.e., $3 \leq CN_{(T_1,T_2)} \leq TN_{(T_1,T_2)}$.

Initiating the process of clustering, we have systematically embraced a methodical approach. Our commitment involves intricately decomposing the task into well-defined preliminary objectives, with a significant emphasis on lineage trees that exhibit congruence in node composition. The primary goal is to precisely compute the *GBLD* distance between these two lineage trees, adhering rigorously to established scientific standards. Subsequently, our focus smoothly transitions towards the systematic construction of clusters, incorporating multiple trees concurrently. This is executed with a mindful awareness of statistical significance and employing robust methodologies. The final stage of our systematic exploration entails meticulous preprocessing to refine the clustering process. This ensures not only its widespread applicability to lineage trees with varying sets of nodes but also maintains the highest standards of scientific rigor and accuracy throughout the entire procedure.

Theorem 2 *The function $GBLD_{(T_1,T_2)}$ satisfies the fundamental properties of a metric.*

Proof The metric properties of a method render it an effective tool for measuring the dissimilarity between two lineage trees. The function $GBLD_{(T_1,T_2)}$ demonstrates adherence to these essential characteristics. A comprehensive analysis follows:

- Non-negativity ($GBLD_{(T_1,T_2)} \geq 0$): The dissimilarity function, $GBLD_{(T_1,T_2)}$, manifests non-negativity, indicating that the dissimilarity between two trees is always a non-negative value. In the context of the dissimilarity functions, ε denotes a small positive value with dissimilarity based on specific conditions regarding distance functions. These symbols succinctly convey essential information about dissimilarity quantification in the mathematical expressions. This property is expressed formally in the probabilistic and weight functions. Given that the number of common nodes is always less than or equal to the total number of nodes ($TN_{(T_1,T_2)} \geq CN_{(T_1,T_2)}$), it follows that the penalty is always non-negative, ranging from 0 to $1 - \varepsilon$. It is reliable to compare two lineage trees in which the number of common nodes is equal to the total number of nodes ($CN_{(T_1,T_2)} = TN_{(T_1,T_2)}$). In this case, there are two possibilities as follows:

$$W_{(T_1,T_2)} = \begin{cases} 0, \text{ if } w_{T_1}(i) = w_{T_2}(i),\ i \in \{1, \ldots, TN_{(T_1,T_2)}\}. \\ \varepsilon, \text{ otherwise.} \end{cases} \quad (8)$$

  If $TN_{(T_1,T_2)} > CN_{(T_1,T_2)}$, the possibility of $W_{(T_1,T_2)} \geq \varepsilon$, is added to the aforementioned possibilities. Within this framework, the potential scenarios concerning distance, taking into account the likelihood of shared nodes and the total node count, as previously outlined in our discussion on node weights are presented.

$$D_{(T_1,T_2)} = \begin{cases} 0, \text{ if } d_{T_1}(i,j) = d_{T_2}(i,j). \\ \varepsilon, \text{ otherwise.} \end{cases} \quad (9)$$

- Identity of Indiscernible ($GBLD_{(T_1,T_2)} = 0$, if and only if $T_1 = T_2$): The identity of indiscernible property asserts that the dissimilarity between a lineage tree and itself is always zero. This property holds true for $GBLD_{(T_1,T_2)}$, as the components within the weight, and distance functions result in zero dissimilarity, i.e., $P_{(T_1,T_2)} = 0, W_{(T_1,T_2)} = 0$, and $D_{(T_1,T_2)} = 0$.

- Symmetry ($GBLD_{(T_1,T_2)} = GBLD_{(T_2,T_1)}$): Symmetry dictates that the dissimilarity between two lineage trees is the same, regardless of their order. This symmetry property is formally expressed in the probabilistic, weight, and distance functions:

$$P_{(T_1,T_2)} = 1 - \frac{CN_{(T_1,T_2)}}{TN_{(T_1,T_2)}} = 1 - \frac{CN_{(T_2,T_1)}}{TN_{(T_2,T_1)}} = P_{(T_2,T_1)}. \tag{10}$$

$$W_{(T_1,T_2)} = \begin{cases} 0, \text{ if } w_{T_1}(i) = w_{T_2}(i). \\ \varepsilon, \text{ otherwise.} \end{cases} \tag{11}$$

$$D_{(T_1,T_2)} = \begin{cases} 0, \text{ if } d_{T_1}(i,j) = d_{T_2}(i,j). \\ \varepsilon, \text{ otherwise.} \end{cases} \tag{12}$$

- Triangle Inequality ($GBLD_{(T_1,T_2)} + GBLD_{(T_2,T_3)} \geq GBLD_{(T_1,T_3)}$): The triangle inequality asserts that the sum of dissimilarities between two pairs of lineage trees is always greater than or equal to the dissimilarity between the first and third lineage trees. The method of *GBLD* is constructed by combining Jaccard, Manhattan, and Euclidean distances. The penalty (Jaccard distance) possesses the property of triangular inequality [47]. For any three points (w_{T_1}, w_{T_2} and w_{T_3} in this study), Manhattan distance satisfies the triangle inequality:

$$\begin{aligned} W_{(T_1,T_3)} &= \sum_{i=1}^{n} \left| w_{T_1}(i) - w_{T_3}(i) \right| \\ &= \sum_{i=1}^{n} \left| w_{T_1}(i) - w_{T_3}(i) \pm w_{T_2}(i) \right| \\ &\leq \sum_{i=1}^{n} \left(\left| w_{T_1}(i) - w_{T_2}(i) \right| + \left| w_{T_2}(i) - w_{T_3}(i) \right| \right) \\ &\leq \underbrace{\sum_{i=1}^{n} \left(\left| w_{T_1}(i) - w_{T_2}(i) \right| \right)}_{W_{(T_1,T_2)}} + \underbrace{\sum_{i=1}^{n} \left(\left| w_{T_2}(i) - w_{T_3}(i) \right| \right)}_{W_{(T_2,T_3)}} \\ &\leq W_{(T_1,T_2)} + W_{(T_2,T_3)}. \end{aligned} \tag{13}$$

This inequality is valid because the absolute difference between two points along a coordinate axis remains less than or equal to the sum of the absolute differences between the points along that axis when considering a third point. To prove the triangularity characteristic of Euclidean distance, consider the triangle formed by the distances of trees (d_{T_1}, d_{T_2}, and d_{T_3}). Let $\overrightarrow{d_{T_1,T_2}}$ represents the distance vector from T_1 to T_2 and $\overrightarrow{d_{T_2,T_3}}$ represents the distance vector from T_2 to T_3. In line with the triangle inequality principle for vectors:

$$\left\| \overrightarrow{d_{T_1,T_2}} + \overrightarrow{d_{T_2,T_3}} \right\| \leq \left\| \overrightarrow{d_{T_1,T_2}} \right\| + \left\| \overrightarrow{d_{T_2,T_3}} \right\|. \tag{14}$$

This inequality can be extended to each dimension.

$$\begin{aligned} \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| d_{T_1}(i,j) - d_{T_3}(i,j) \right|^2} \leq & \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| d_{T_1}(i,j) - d_{T_2}(i,j) \right|^2} \\ & + \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| d_{T_2}(i,j) - d_{T_3}(i,j) \right|^2}. \end{aligned} \tag{15}$$

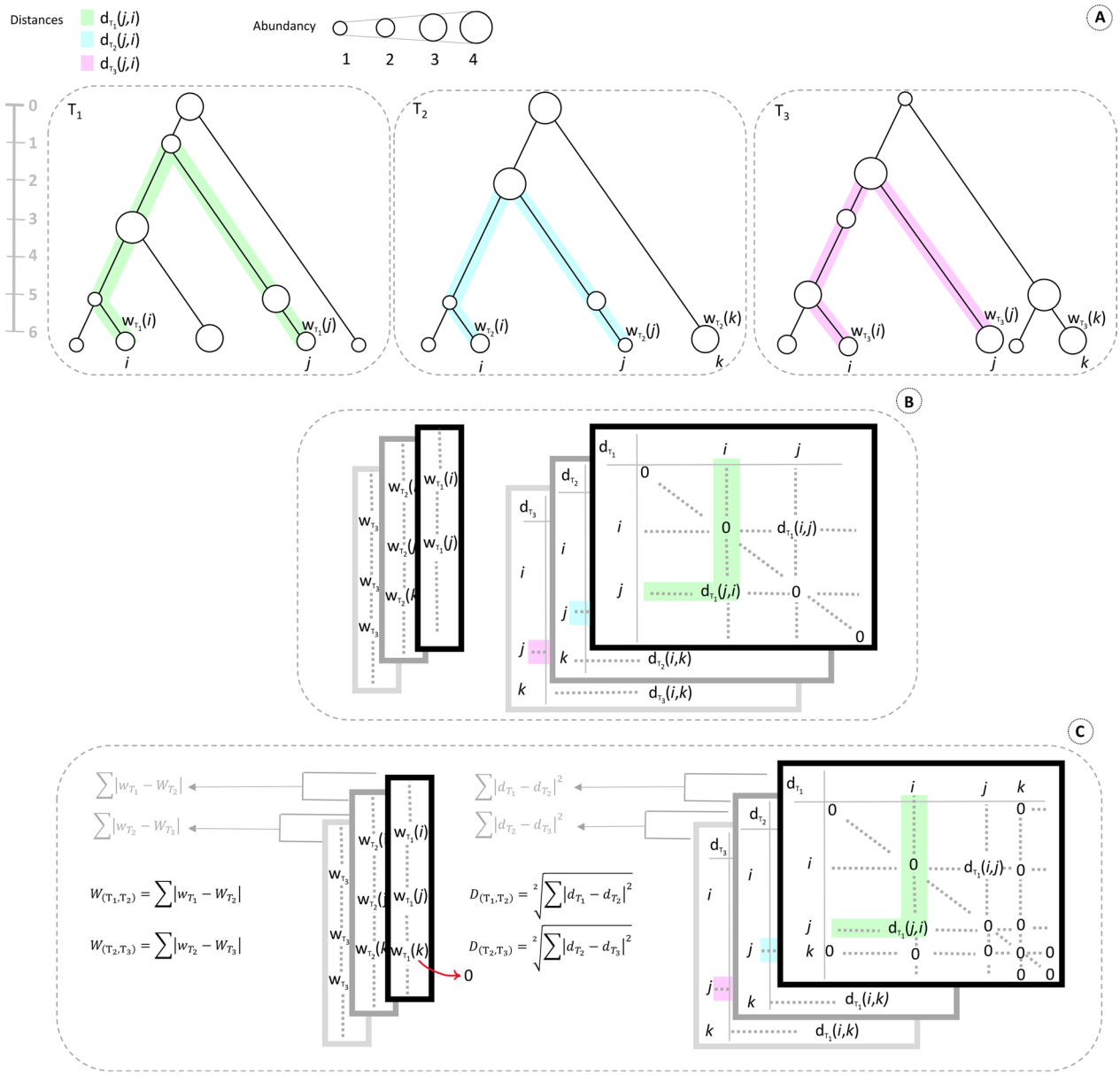

Fig. 2 A graphical representation of the *GBLD* metric approach. **A** It depicts three lineage trees exhibiting variations in both the weights and lengths of branches. The weights of the nodes are symbolized by circles, wherein larger circles correspond to higher node weights, while smaller circles signify lower weights. Different paths connecting nodes *i* and *j* in trees T_1, T_2, and T_3 are depicted using distinct colors; notably, the path in tree T_1 is green, in T_2 it is blue, and in T_3 it is pink. **B** It represents the extracted data from lineage trees in the form of weight vectors and branch length matrices. **C** It displays the status of missing nodes in the context of comparing lineage trees. Vectors and matrices, upon manipulation that involves the incorporation of absent nodes, are prepared for pairwise subtraction

Thus, we have demonstrated that:

$$D_{(T_1,T_3)} \leq D_{(T_1,T_2)} + D_{(T_2,T_3)}. \tag{16}$$

All the metric properties are preserved after normalization of $W_{(T_1,T_2)}$ and $D_{(T_1,T_2)}$ because being metric depends on the intrinsic properties of the function itself, which remain unchanged by scaling or normalization [48]. Combining the distances measured by three metrics, each obeying the properties of a metric, preserves the essential characteristics of distance measurement, ensuring the resulting sum $(P_{(T_1,T_2)} + W_{(T_1,T_2)} + D_{(T_1,T_2)})$ remains a metric [49].

Given the fulfillment of all requisite properties within *GBLD*, it is a metric. □

Remark 2 The *GBLD* score between two distinct lineage trees ranges from ε to infinity (i.e., $\varepsilon \cong GBLD_{(T_1,T_2)} < \infty$). The *GBLD* score closer to ε indicates high similarity between the two lineage trees, and vice versa.

Illustration of the *GBLD* metric methodology

The theoretical aspects of the *GBLD* metric approach are presented in the previous sections to elucidate its nature as a metric method. This section includes a visual representation of the *GBLD* method, aiming to elucidate its mechanism and enhance comprehension of this innovative approach.

With this objective in mind, Fig. 2 presents the mathematical depiction. The first row of Fig. 2 (2A) illustrates three lineage trees with varying weights and branch lengths for comparative analysis. In these three lineage trees, a pathway connecting nodes i and j is distinguished by employing three distinct colors. The second row of Fig. 2 (2B) illustrates the mathematical representations derived from the extracted data of the topology of lineage trees. Each vector on the left side indicates the weights assigned to each node in each lineage tree. On the right, each matrix displays the distances between each pair of nodes in each tree.

As stated in Remark 1, a minimum of three common nodes is required in two lineage trees for them to be comparable. However, there is no restriction on the number of different nodes between them. In the process of comparing two lineage trees, any missing nodes are incorporated into the respective lineage trees where they are absent to facilitate a comprehensive analysis. In this context, node k exemplifies a present node in lineage trees T_2 and T_3, yet it is not found in T_1. The last row of Fig. 2 (2C) shows the modified lineage trees to include all nodes. The lineage tree T_1 acquires a hypothetical node k. Since the node k is not present in T_1, it is assumed that the weight of node k and the distances between node k and the rest of the nodes in T_1 is zero. This assumption is made to prevent the exclusion of the branch length distance between node k and other nodes in the other lineage trees when calculating the *GBLD* score. Incorporating absent nodes not only preserves the influence of these nodes in other trees that contain them but also plays a vital role in facilitating the preparation for the calculation and alignment of weight vectors and branch length matrices to achieve uniformity in the dimension.

Algorithm

In order to give a better understanding of the *GBLD* method, Algorithm 1, containing three algorithms, is provided to show the most prominent steps in the calculation of *GBLD* score. Three inner algorithms of Algorithm 1 are the components of Eq. 7. The main objective of this Algorithm is using the data of lineage trees and evaluate the *GBLD* score regardless of the form of the data. Although it is possible to use the original lineage trees and calculate the differences in weights and distances between nodes through a recursive process or by using Newick formats, it is preferred to use precalculated distance matrices and weight vectors to avoid losing the focus of Algorithm 1. The details of this process are illustrated in the previous subsection, specifically in Fig. 2.

Algorithm 1 Calculate *GBLD* score

Require:
$P_{(T_1,T_2)}$: Penalty between two lineage trees T_1 and T_2
$W_{(T_1,T_2)}$: Weight differentiation between two lineage trees T_1 and T_2
$D_{(T_1,T_2)}$: Distance differentiation between two lineage trees T_1 and T_2
Ensure: $GBLD$ score between two lineage trees T_1 and T_2

function GBLDSCORE($P_{(T_1,T_2)}, W_{(T_1,T_2)}, D_{(T_1,T_2)}$)
  $P_{(T_1,T_2)} \leftarrow$ PENALITY(N_{T_1}, N_{T_2}) ▷ Algo. S1*
  $W_{(T_1,T_2)} \leftarrow$ PAIRWEIGHTS($N_{T_1}, N_{T_2}, W_{T_1}, W_{T_2}$) ▷ Algo. S2*
  $D_{(T_1,T_2)} \leftarrow$ PAIRDISTANCE($D_{T_1}, D_{T_2}, N_{T_1}, N_{T_2}$) ▷ Algo. S3*
  $GBLD_{(T_1,T_2)} \leftarrow P_{(T_1,T_2)} + W_{(T_1,T_2)} + D_{(T_1,T_2)}$ ▷ Eq. 7
  return $GBLD_{(T_1,T_2)}$
end function

*Note: S denotes supplementary material file

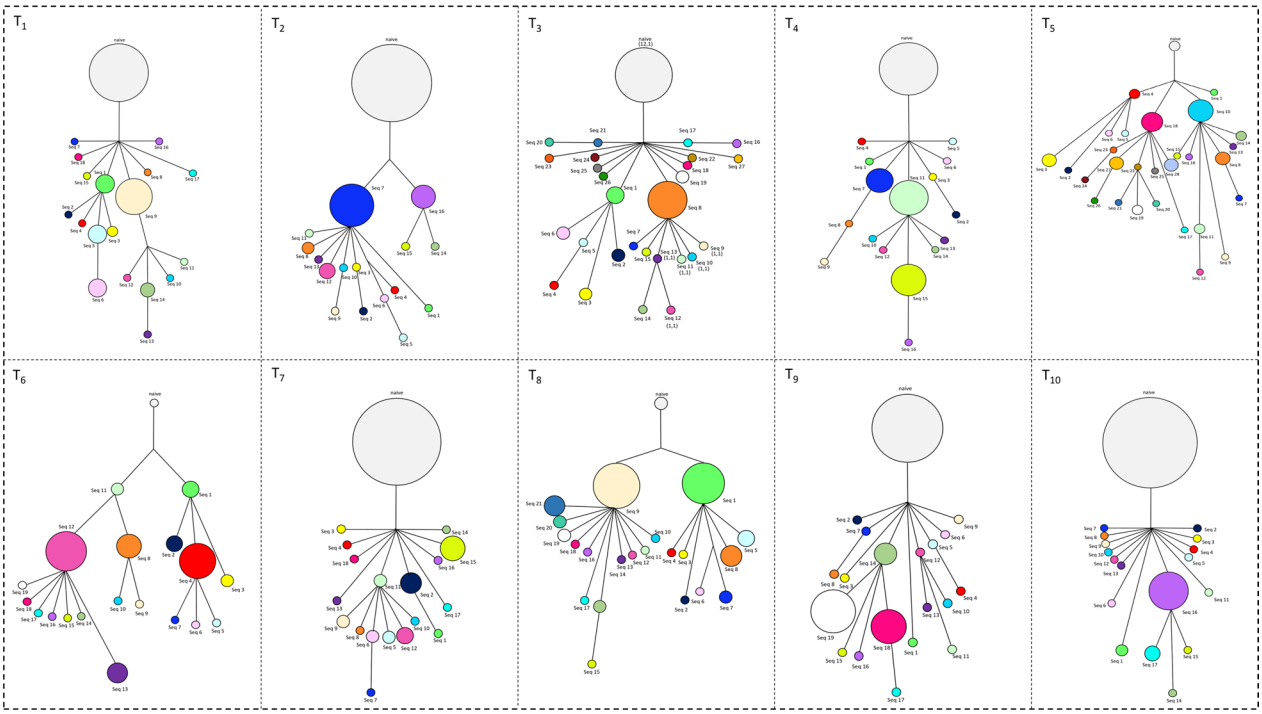

Fig. 3 The topological structure of the simulated dataset. It exhibits ten lineage trees, which feature various common nodes

Lemma 3 *Algorithm 1 computes the solution of the GBLD problem in $\mathcal{O}(n^2)$ time, where n and m are the sizes of the two lineage trees T_1 and T_2, and $n > m$.*

Proof The overall time complexity for Algorithm S1, iterating over N_{T_1} and checking membership in N_{T_2} is $\mathcal{O}(n)$. On the other hand, iterating over the nodes set in the Algorithm S2, having at most $n + m$ elements takes $\mathcal{O}(n + m)$ time. Algorithm S3 containing two nested loops, takes $\mathcal{O}(n + m)^2$ time which is the largest time compared to the other algorithms. Combining the complexities of all the algorithms, the overall time complexity of Algorithm 1 is dominated by Algorithm S3. If $n > m$, time complexity of Algorithm 1 is $\mathcal{O}(n^2)$. □

Materials

Simulated dataset

In this section, we provide a detailed explanation of our data collection and processing approach, with a specific focus on generating simulated datasets. We subsequently validate this approach through the application of mathematical methods.

Simulated data are derived from the GCtree simulator [50]. GCtree simulator aims to produce data similar to experimental BCR sequencing results. Simulation parameters are calibrated using summary statistics on sequences and abundances to match characteristics of real data, without directly involving tree inference in the calibration process.

The *GBLD* metric methodology involves a meticulous examination of three attributes within lineage trees, aimed at precisely quantifying their similarities and dissimilarities. Consequently, we intentionally manipulated the magnitudes of these characteristics within lineage trees to systematically assess the robustness of this methodology and rigorously scrutinize its accuracy.

To rigorously evaluate the performance of *GBLD*, one simulated dataset was generated. The configurations of this dataset which is comprised of 10 lineage trees are illustrated in Fig. 3. Examining this figure reveals variations in the characteristics of the lineage trees. The lineage trees depicted in Fig. 3 possess a spectrum of common and uncommon, which impart a distinct coloration pattern to each tree. In comparing one lineage tree to another, certain nodes within these trees may have the same weight and branch length distances. However, this lineage tree might show different levels of variation in weight and/or branch lengths when compared with another lineage tree.

Providing detailed insights into the dataset facilitates the establishment of a perspective for interpreting the

final *GBLD* matrix. In this dataset, all lineage trees, with the exceptions of T_5, T_6, and T_8, feature a naive node with significant weight. Lineage trees labeled T_3 and T_5, containing 27 and 28 nodes respectively, exhibit a higher node count compared to others; however, these two lineage trees demonstrate fewer similarities in terms of the weights and branch lengths. Trees T_2, T_7, and T_{10} possess nearly common nodes, with the weights of these nodes being very similar. Trees T_1,T_4, and T_9 share nearly the same nodes as T_2, T_7, and T_{10}, yet the weights of the nodes within each group exhibit slight variations.

Real dataset

To validate the *GBLD* metric, one public database, from [51], comprising over 37 million unique BCR sequences of three healthy adult donors is utilized, which is significantly more comprehensive than any existing resource. Each adult donor contains BCR sequences of naive (N) and memory (M) B cell repertoires. A sample of 150 BCR sequences was selected from the dataset, which was composed of nucleotide sequences with detailed information on their frequency, genetic composition, and formation via V(D)J recombination. Subsequently, the sequences (in fasta format) were aligned using MUSCLE [52] of BioPython version 1.84 [53] to identify regions of similarity. Then, distance matrices between sequences were calculated using the class 'Bio.Phylo.TreeConstruction. DistanceCalculator('identity')' [53] from the BioPython. Finally, a phylogenetic tree (in Newick format) was constructed using the Neighbor-Joining (NJ) method [54]. The default values were used for all parameters.

The degree of overlapping nodes between two independent replicates of the naive repertoire of donor 1 (D1-Na and D1-Nb) is almost higher than in other comparisons. Additionally, there is a significant overlap between the naive and memory repertoires of each donor. Despite the large size and diversity of the B cell repertoire, the overall overlap between donors is small, with only 11 common nodes identified.

Results of this study [51] demonstrate that the similarities between two independent replicates of the naive repertoire of donor 1 are almost greater than those between different donors. Furthermore, the naive and memory B cell populations within each donor are more similar to each other than to those of different donors, aligning with the results of the *GBLD* method.

Results and discussion

Simulated dataset

The *GBLD* metric method is applied to the provided simulated dataset. The following steps elucidate how the features of two lineage trees under comparison are incorporated into the *GBLD* metric method. Firstly, the weights of all nodes in both lineage trees are included in Eq. 5. Then, the branch length distances of each pair of nodes are integrated into Eq. 4. The total and common number of nodes are counted and placed in the penalty index specified by Eq. 6. Finally, the Eq. 7 gives the final *GBLD* score between two lineage trees.

The ultimate outcomes of comparing the lineage trees are consolidated in the subsequent symmetric matrix.

$$GBLD = \begin{bmatrix} 0.0 & 0.78 & 1.04 & 0.77 & 1.13 & 0.75 & 0.53 & 0.76 & 0.49 & 0.67 \\ 0.78 & 0.0 & 1.13 & 0.44 & 1.23 & 0.88 & 0.7 & 0.96 & 0.74 & 0.64 \\ 1.04 & 1.13 & 0.0 & 1.14 & 0.49 & 1.07 & 0.98 & 0.92 & 1.01 & 1.01 \\ 0.77 & 0.44 & 1.14 & 0.0 & 1.24 & 0.93 & 0.69 & 0.98 & 0.77 & 0.72 \\ 1.13 & 1.23 & 0.49 & 1.24 & 0.0 & 1.14 & 1.11 & 0.98 & 1.04 & 1.12 \\ 0.75 & 0.88 & 1.07 & 0.93 & 1.14 & 0.0 & 0.71 & 0.89 & 0.69 & 0.81 \\ 0.53 & 0.7 & 0.98 & 0.69 & 1.11 & 0.71 & 0.0 & 0.81 & 0.59 & 0.52 \\ 0.76 & 0.96 & 0.92 & 0.98 & 0.98 & 0.89 & 0.81 & 0.0 & 0.84 & 0.66 \\ 0.49 & 0.74 & 1.01 & 0.77 & 1.04 & 0.69 & 0.59 & 0.84 & 0.0 & 0.89 \\ 0.67 & 0.64 & 1.01 & 0.72 & 1.12 & 0.81 & 0.52 & 0.66 & 0.89 & 0.0 \end{bmatrix}$$

To optimally partition lineage trees based on their *GBLD* scores, the DBSCAN [55] method proves to be effective. This method is advantageous as it facilitates the identification of outliers and supports the formation of a single cluster. DBSCAN works by grouping points that are closely packed together, marking them as *core* points if they have a sufficient number of neighbors within a specified radius, or as *border* points if

Table 2 *GBLD* score for three healthy individuals

	D1-M	D1-Na	D1-Nb	D2-M	D2-N	D3-M	D3-N
D1-M	0	1.35	1.6	1.99	2.18	2.1	2.17
D1-Na	1.35	0	0.78	2.06	2.27	2.2	2.31
D1-Nb	1.6	0.78	0	2.33	2.49	2.4	2.51
D2-M	1.99	2.06	2.33	0	0.98	2.18	2.27
D2-N	2.18	2.27	2.49	0.98	0	2.36	2.46
D3-M	2.1	2.2	2.4	2.18	2.36	0	0.68
D3-N	2.17	2.31	2.51	2.27	2.46	0.68	0

It displays the *GBLD* score obtained from applying the *GBLD* metric to the B cell sequences of three individuals

they are close to core points but do not have enough neighbors to be considered core points. Points that lie alone in low-density regions are marked as *outliers*. This characteristic makes it particularly suited for scenarios where the cluster structure is not strictly spherical and when dealing with noise in the dataset.

Based on the DBSCAN the optimal scenario occurs when there are two clusters.

The following outlines the optimal partition for the dataset.

- Cluster 1: $T_1, T_2, T_4, T_6, T_7, T_8, T_9, T_{10}$
- Cluster 2: T_3, T_5

As stated in *Remark* 2, a *GBLD* score approaching zero indicates a higher degree of similarity, whereas a higher score implies greater dissimilarity.

The smallest *GBLD* scores are associated with T_2 and T_4 (0.44), and T_1 and T_9 (0.49), which are almost one third the magnitude of the largest scores, observed between T_4 and T_5 (1.24), and T_2 and T_5 (1.23). This indicates that the degree of similarity between the lineage trees of T_2, T_4 and T_1, T_9 exceeds that of others.

The *GBLD* score between lineage trees T_3 and T_5 is (0.49), which shows a high similarity between them in comparison with other lineage trees. Additionally, the *GBLD* scores of lineage trees T_3 and T_5 with other lineage trees are not small enough and have a negligible impact on the final partitioning.

Real dataset

Evaluation of pairwise overlap between repertoires in the public database of memory and naive B cell receptor sequences [51] highlights the necessity of the penalty index in the *GBLD* metric. It is also noted that shared sequences (those present in two or three subjects) tend to have higher maximum occupancy. *GBLD* metric effectively incorporates the examination of shared sequences with high weights, and high common weights do not significantly increase the *GBLD* score. Applying the *GBLD* metric to the sampled data results in the following *GBLD* matrix.

According to Table 2, the lowest values correspond to the repetitions of naive cells for donor 1 ($GBLD_{(D1-Na,D1-Nb)} = 0.78$), and the comparisons between naive and memory cells within each individual ($GBLD_{(D1-M,D1-Na)} = 1.35$, $GBLD_{(D1-M,D1-Nb)} = 1.6$, $GBLD_{(D2-M,D2-N)} = 0.98$, and $GBLD_{(D3-M,D3-N)} = 0.68$).

Comparing these values with the other values in the GBLD matrix corroborates the findings of the previous study [51], and validates the accuracy of the *GBLD* metric.

Revisiting the section on the design of the simulated dataset and analyzing the *GBLD* matrix, it becomes apparent that lineage trees with a greater number of common nodes and higher similarity in weights and branch lengths generally exhibit lower *GBLD* scores. While the detection of these similarities is relatively simple in smaller datasets, it would be more difficult in larger, more complex datasets. Nevertheless, the *GBLD* metric method provides a proficient approach for identifying these similarities, bypassing the challenges posed by complex topologies.

In our investigation, some common nodes in lineage trees were observed that shared the same length and weight but were positioned differently. This discrepancy in position resulted in varying branch length distances between this node and others in the lineage trees, leading to the oversight of certain topology details. As a prospective initiative, it is beneficial to explore an index associated with the topologies of compared lineage trees. This new index can raise the *GBLD* score, improving the accuracy of the *GBLD* metric method and upholding its metric integrity.

Conclusion

Our research focuses on introducing an innovative technique that enables a comprehensive evaluation of lineage tree attributes. The primary objective is to achieve optimal partitioning while maintaining the metric properties of the proposed method. Our metric approach is diligent in incorporating the most crucial features of lineage trees, ensuring a nuanced analysis. To validate our method, we adopt the *DBSCAN* algorithm, which offers a robust framework for determining the optimal number of clusters and outliers.

To fully comprehend the findings of our study, it is crucial to have a deep understanding of lineage tree topologies. The *GBLD* score is a major metric that provides valuable insights into how two lineage trees are similar. This, in turn, helps us make more accurate predictions about how B cells react to viruses. Such knowledge can significantly improve the precision of immunotherapy and vaccine development by offering a more nuanced understanding of B cell behavior.

We discovered that variations in the structures of lineage trees have a significant impact on the *GBLD* score. This finding emphasizes the importance of this particular feature. In upcoming research, we plan on incorporating this observation into our metric framework. We believe that doing so will lead to more accurate assessments of lineage tree dynamics. These advancements will offer valuable insights to the scientific community, particularly in comprehending and regulating B cell immune responses.

To improve our novel metric, we need to carefully manage nodes both internally and externally to avoid potential biases in the metric. To address this, we are considering the possibility of adding leaves on both sides of the lineage trees, which would include all nodes in the dataset. This approach is inspired by the RF(+) method [56]. We will introduce branches and nodes that are absent in one tree but present in the other, aligning with established methodologies in the field during preprocessing.

Acknowledgements
The authors would like to thank the Department of Computer Science, University of Sherbrooke, Quebec, Canada for providing the necessary resources to conduct this research. The authors also thank the reviewers and the Editor for their valuable comments on this paper.

Author contributions
Both authors contributed equally for the development of this article.

Funding
This work was supported by the Natural Sciences and Engineering Research Council of Canada (Grant no. RGPIN-2022-04322), Fonds de recherche du QuÃ©bec - Nature and technologies (Grant no. 326911), and the University of Sherbrooke grant.

Availability of data and materials
The datasets generated and analyzed during the current study along with the *GBLD* metric programs (C++ source code and Python scripts) are freely available at: https://github.com/tahiri-lab/ClonalTreeClustering.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no cmpeting interests.

Received: 17 June 2024 Accepted: 24 September 2024
Published online: 05 October 2024

References

1. Schwab I, Nimmerjahn F. Intravenous immunoglobulin therapy: how does IgG modulate the immune system? Nat Rev Immunol. 2013;13(3):176–89.
2. Lefranc MP. Immunoglobulin and T cell receptor genes: IMGT® and the birth and rise of immunoinformatics. Front Immunol. 2014;5:22.
3. Zhang L, Vijg J. Somatic mutagenesis in mammals and its implications for human disease and aging. Annu Rev Genet. 2018;52:397–419.
4. Ruprecht CR, Lanzavecchia A. Toll-like receptor stimulation as a third signal required for activation of human naive B cells. Eur J Immunol. 2006;36(4):810–6.
5. de Bourcy CF, Angel CJL, Vollmers C, Dekker CL, Davis MM, Quake SR. Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. Proc Natl Acad Sci. 2017;114(5):1105–10.
6. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381):306–13.
7. Hoehn KB, Fowler A, Lunter G, Pybus OG. The diversity and molecular evolution of B-cell receptors during infection. Mol Biol Evol. 2016;33(5):1147–57.
8. Nouri N, Kleinstein SH. Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data. PLoS Comput Biol. 2020;16(6): e1007977.
9. Li A, Rue M, Zhou J, Wang H, Goldwasser MA, Neuberg D, et al. Utilization of Ig heavy chain variable, diversity, and joining gene segments in children with B-lineage acute lymphoblastic leukemia: implications for the mechanisms of VDJ recombination and for pathogenesis. Blood. 2004;103(12):4602–9.
10. Alt FW, Oltz EM, Young F, Gorman J, Taccioli G, Chen J. VDJ recombination. Immunol Today. 1992;13(8):306–14.
11. Tabibian-Keissar H, Zuckerman NS, Barak M, Dunn-Walters DK, Steiman-Shimony A, Chowers Y, et al. B-cell clonal diversification and gut-lymph node trafficking in ulcerative colitis revealed using lineage tree analysis. Eur J Immunol. 2008;38(9):2600–9.
12. Uduman M, Shlomchik MJ, Vigneault F, Church GM, Kleinstein SH. Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. J Immunol. 2014;192(3):867–74.
13. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R. IgTree: creating immunoglobulin variable region gene lineage trees. J Immunol Methods. 2008;338(1–2):67–74.
14. Kurosaki T, Kometani K, Ise W. Memory B cells. Nat Rev Immunol. 2015;15(3):149–59.
15. Seifert M, Küppers R. Human memory B cells. Leukemia. 2016;30(12):2283–92.
16. Walter S. Minkowski, mathematicians, and the mathematical theory of relativity. Expand Worlds Gen Relativity. 1999;7:45–86.
17. Woese CR. Interpreting the universal phylogenetic tree. Proc Natl Acad Sci. 2000;97(15):8392–6.
18. Nowell PC. The clonal evolution of tumor cell populations: acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. Science. 1976;194(4260):23–8.
19. DeWitt WS III, Mesin L, Victora GD, Minin VN, Matsen FA IV. Using genotype abundance to improve phylogenetic inference. Mol Biol Evol. 2018;35(5):1253–65.
20. Abdollahi N, Jeusset L, de Septenville A, Davi F, Bernardes JS. Reconstructing B cell lineage trees with minimum spanning tree and genotype abundances. BMC Bioinform. 2023;24(1):70.
21. Buneman P. A note on the metric properties of trees. J Combin Theory Ser B. 1974;17(1):48–50.
22. Davidsen K, Matsen FA IV. Benchmarking tree and ancestral sequence inference for B cell receptor sequences. Front Immunol. 2018;9:2451.
23. Górecki P, Eulenstein O. A Robinson-Foulds measure to compare unrooted trees with rooted trees. In: International symposium on bioinformatics research and applications. Springer; 2012. p. 115–26.
24. Odegard VH, Schatz DG. Targeting of somatic hypermutation. Nat Rev Immunol. 2006;6(8):573–83.
25. Tietje M, Antonelli A, Forest F, Govaerts R, Smith SA, Sun M, et al. Global hotspots of plant phylogenetic diversity. N Phytol. 2023;240(4):1636–46.
26. Hamza N, Hershberg U, Kallenberg CG, Vissink A, Spijkervet FK, Bootsma H, et al. Ig gene analysis reveals altered selective pressures on Ig-producing cells in parotid glands of primary Sjögren's syndrome patients. J Immunol. 2015;194(2):514–21.
27. Chan TD, Brink R. Affinity-based selection and the germinal center response. Immunol Rev. 2012;247(1):11–23.
28. Mesin L, Ersching J, Victora GD. Germinal center B cell dynamics. Immunity. 2016;45(3):471–82.
29. Riedel R, Addo R, Ferreira-Gomes M, Heinz GA, Heinrich F, Kummer J, et al. Discrete populations of isotype-switched memory B lymphocytes are maintained in murine spleen and bone marrow. Nat Commun. 2020;11(1):2570.
30. Hershberg U, Luning Prak ET. The analysis of clonal expansions in normal and autoimmune B cell repertoires. Philos Trans R Soc B Biol Sci. 2015;370(1676):20140239.
31. Hoehn KB, Kleinstein SH. B cell phylogenetics in the single cell era. Trends Immunol. 2023;45:62–74.
32. Garba MK, Nye TM, Lueg J, Huckemann SF. Information geometry for phylogenetic trees. J Math Biol. 2021;82:1–39.

33. Koshkarov A, Tahiri N. Novel algorithm for comparing phylogenetic trees with different but overlapping taxa. Symmetry. 2024;16(7):790.
34. Adams RH, Blackmon H, DeGiorgio M. Of traits and trees: probabilistic distances under continuous trait models for dissecting the interplay among phylogeny, model, and data. Syst Biol. 2021;70(4):660–80.
35. Garba MK, Nye TM, Boys RJ. Probabilistic distances between trees. Syst Biol. 2018;67(2):320–7.
36. Govek K, Sikes C, Oesper L. A consensus approach to infer tumor evolutionary histories. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics; 2018. p. 63–72.
37. DiNardo Z, Tomlinson K, Ritz A, Oesper L. Distance measures for tumor evolutionary trees. Bioinformatics. 2020;36(7):2090–7.
38. Llabrés M, Rosselló F, Valiente G. A generalized Robinson-Foulds distance for clonal trees, mutation trees, and phylogenetic trees and networks. In: Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics; 2020. p. 1–10.
39. Jahn K, Beerenwinkel N, Zhang L. The Bourque distances for mutation trees of cancers. Alg Mol Biol. 2021;16(1):9.
40. Khayatian E, Valiente G, Zhang L. The k-Robinson-Foulds dissimilarity measures for comparison of labeled trees. J Comput Biol. 2024;31(4):328–44.
41. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol. 1994;11(3):459–68.
42. Semple C, Steel M, et al. Phylogenetics, vol. 24. Oxford: Oxford University Press on Demand; 2003.
43. Soria-Carrasco V, Talavera G, Igea J, Castresana J. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics. 2007;23(21):2954–6.
44. Duchêne DA, Tong KJ, Foster CS, Duchêne S, Lanfear R, Ho SY. Linking branch lengths across sets of loci provides the highest statistical support for phylogenetic inference. Mol Biol Evol. 2020;37(4):1202–10.
45. Danielsson PE. Euclidean distance mapping. Comput Graph Image Process. 1980;14(3):227–48.
46. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat. 1901;37:547–79.
47. Kosub S. A note on the triangle inequality for the Jaccard distance. Pattern Recogn Lett. 2019;120:36–8.
48. Yianilos PN. Normalized forms for two common metrics. NEC Res Inst: Rep; 2002. p. 91–082.
49. Doboš J. Metric preserving functions. Štroffek Košice; 1998.
50. DeWitt I, William S, Mesin L, Victora GD, Minin VN, Matsen I, Frederick A. Using genotype abundance to improve phylogenetic inference. Mol Biol Evol. 2018;35(5):1253–65.
51. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. A public database of memory and naive B-cell receptor sequences. PLoS ONE. 2016;11(8): e0160853.
52. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform. 2004;5:1–19.
53. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422.
54. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.
55. Schubert E, Sander J, Ester M, Kriegel HP, Xu X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans Database Syst TODS. 2017;42(3):1–21.
56. Cotton JA, Wilkinson M. Majority-rule supertrees. Syst Biol. 2007;56(3):445–52.

Publisher's Note