A Hierarchical Haplotype-Based Distance for Genome-Wide Phylogenetic Reconstruction

Pasquale Marino[1], Roberto Pagliarini[1,*]

[1]University of Udine, Department of Mathematics, Computer Science, and Physics, Via delle Scienze, 206, Udine, 33100, Italy

Abstract

Phylogenetic reconstruction is a central problem in computational biology, typically addressed through sequence alignments and distance- or character-based inference methods. Recent advances in sequencing technologies have enabled the generation of large-scale haplotype-resolved datasets, motivating the development of phylogenetic approaches capable of exploiting allele-specific information. In this work, we propose a hierarchical distance model for genome-wide phylogenetic reconstruction based on allele-specific haplotypes. The model defines genetic distances through successive aggregation steps at the allele, genotype, chromosome, and genome levels, producing a unified distance matrix suitable for distance-based phylogenetic inference. Experiments on a population-scale *Vitis vinifera* dataset show that the proposed framework produces stable genome-wide phylogenies while preserving chromosome-specific evolutionary signals. Reconstructions obtained with Neighbor-Joining and Balanced Minimum Evolution are largely consistent, indicating that the proposed distance captures robust phylogenetic information. Our results suggest that hierarchical aggregation of haplotype-level information provides a promising framework for genome-wide phylogenetic analysis, and offers a promising foundation for future studies on large population-scale genomic datasets.

Keywords

Phylogenetic reconstruction, Haplotype analysis, Genome-wide phylogeny, Distance-based methods, Computational genomics

1. Introduction

Phylogenetic trees provide a compact representation of evolutionary relationships among biological taxa, the official groups used to classify living organisms, and constitute one of the most important tools in evolutionary biology and bioinformatics [1]. Traditional phylogenetic reconstruction methods typically rely on sequence alignments and pairwise genetic distances computed directly from nucleotide sequences [2]. However, this two-step approach creates an information bottleneck: errors in sequence alignment propagate directly to the tree topology [3, 4]. The increasing availability of genome-wide (GW) datasets has introduced new challenges and opportunities. In particular, phased haplotypes and allele-specific information offer a richer representation of genetic variability than single consensus sequences [5, 6, 7, 8].

In this work, we present a hierarchical distance model for phylogenetic reconstruction from allele-specific haplotypes. The proposed framework aggregates information from individual alleles up to complete genomes and enables the construction of phylogenetic trees using standard distance-based methods. The main contributions are: *i)* a hierarchical haplotype-based distance model; *ii)* a GW phylogenetic reconstruction framework; *iii)* an experimental evaluation on a population-scale *Vitis vinifera* dataset. The primary objective of this work is to introduce and evaluate a hierarchical framework for aggregating haplotype-level information into GW distances, rather than to benchmark alternative nucleotide-level distance measures. Since the proposed approach is orthogonal to the underlying allele-level distance, different distance functions could be incorporated within the same hierarchical framework. The present study therefore focuses on assessing the internal consistency and robustness of

ICTCS 2026: Italian Conference on Theoretical Computer Science, September 7–-9, 2026, Udine, Italy

*Corresponding author.

marino.pasquale@spes.uniud.it (P. Marino); roberto.pagliarini@uniud.it (R. Pagliarini)

http://bioinf.dimi.uniud.it/authors/roberto-pagliarini/ (R. Pagliarini)

0000-0001-9672-4326 (R. Pagliarini)

the proposed aggregation strategy, that can be employed within standard distance-based reconstruction methods.

The paper follows this structure: in section 2 we briefly review the state of the art in phylogenetic pipelines; section 3 goes into the details of our hierarchical distance models and analyzes its computational complexity. In section 4, we describe our dataset and analyze the experimental results. Finally, in section 5 we draw conclusions and discuss possible improvements of the method, which we plan to further investigate in the future.

2. Related Work

Modern phylogenetic pipelines generally consist of three main stages: sequence alignment, genetic distance estimation, and tree inference [2]. Different methods have been proposed for each stage, resulting in a broad spectrum of phylogenetic approaches that differ in terms of accuracy, computational complexity, and biological assumptions.

Sequence alignment is the first step in phylogenetic analysis. Its objective is to arrange biological sequences in a way that highlights homologous positions and evolutionary relationships among nucleotides or amino acids. Since phylogenetic inference strongly depends on the quality of the alignment, errors introduced at this stage may propagate to the final tree reconstruction [9]. Sequence alignment can be classified into pairwise alignment (PA) and multiple sequence alignment (MSA). PA compares two sequences at a time, whereas MSA compares three or more sequences simultaneously [10]. Considering MSA methods, they can be broadly classified into progressive, dynamic-programming-based, and consistency-based approaches. Progressive methods represent one of the most widely adopted families of MSA algorithms. They typically start by computing pairwise alignments among sequences, then construct a guide tree, and progressively align sequences according to the tree topology. Well-known examples include ClustaL W [11], Clustal Omega [12], and MAFFT [13]. These methods provide a good trade-off between computational efficiency and alignment quality, making them suitable for large datasets [14]. Dynamic-programming-based approaches compute optimal alignments through recursive formulations. A classical example is the Needleman–Wunsch algorithm, which performs global sequence alignment by filling a dynamic programming matrix and subsequently reconstructing the optimal alignment through backtracking [15]. Although highly accurate, dynamic programming approaches become computationally expensive when extended to multiple sequences. Consistency-based methods improve alignment quality by incorporating information obtained from multiple pairwise alignments before constructing the final MSA. T-Coffee is one of the most representative examples of this category [16]. These approaches generally produce more accurate alignments than purely progressive methods, especially when sequences are moderately divergent, at the cost of increased computational complexity.

Once the sequences have been aligned, phylogenetic relationships can be inferred directly from the sequence characters. Character-based methods analyze each position of the alignment and attempt to identify the tree that best explains the observed evolutionary changes. Maximum Parsimony (MP) is one of the earliest approaches proposed for phylogenetic reconstruction [17]. The underlying principle is that the preferred evolutionary scenario is the one that requires the smallest number of character changes. Given an alignment, MP searches for the tree minimizing the total number of mutations necessary to explain the observed sequences. One of the main advantages of MP is that it does not require an explicit evolutionary model. However, its performance may deteriorate in the presence of a high evolutionary divergence, where multiple substitutions at the same site become common. Maximum Likelihood (ML) addresses this limitation by introducing explicit probabilistic models of sequence evolution. In ML-based approaches, different tree topologies are evaluated according to the likelihood of observing the available data under a selected evolutionary model. Commonly used substitution models include Jukes–Cantor [18], Tamura–Nei [19], and General Time Reversible [20]. Although ML methods are generally more accurate than MP, they are computationally demanding because of the large search space of possible tree topologies and model parameters. Distance-based

approaches reconstruct phylogenetic trees from pairwise genetic distances rather than directly from sequence characters. These methods are often preferred when large datasets are analyzed because they generally have lower computational costs than character-based alternatives. The effectiveness of distance-based phylogenetic reconstruction depends on the definition of an appropriate distance measure. Among the most common choices are the Hamming distance, the p-distance, and model-based corrections such as the Jukes–Cantor distance. The Hamming distance counts the number of positions at which two sequences differ, whereas the p-distance corresponds to its normalized version. The Jukes–Cantor model introduces a correction for multiple substitutions occurring at the same site, providing a more realistic estimate of evolutionary divergence [21]. Neighbor-Joining (NJ) is one of the most influential distance-based methods [22]. It iteratively identifies pairs of taxa that minimize a specific optimization criterion and joins them through newly introduced internal nodes. The algorithm produces unrooted phylogenetic trees and simultaneously estimates branch lengths. Due to its balance between computational efficiency and reconstruction quality, NJ remains one of the most widely used methods in molecular phylogenetics. Distance-based phylogenetic reconstruction includes several algorithmic families. Agglomerative clustering methods such as UPGMA [23] and WPGMA [24] iteratively merge clusters according to average-distance criteria and assume a molecular clock. NJ is also an agglomerative distance-based algorithm, but it relies on a different optimization criterion and does not require the molecular clock assumption. Although UPGMA and WPGMA are historically important and remain useful in settings where the molecular clock assumption is approximately satisfied, they are less commonly employed for genome-scale phylogenetic reconstruction because violations of this assumption may substantially affect reconstruction accuracy. Minimum Evolution (ME) methods extend the distance-based framework by selecting the topology minimizing the total tree length [25]. Balanced Minimum Evolution (BME) further improves this idea through weighted estimates of branch lengths and more efficient optimization procedures [26]. Previous studies have shown that BME often achieves competitive reconstruction accuracy while maintaining favorable computational properties.

Although sequence alignment remains a central component of many phylogenetic pipelines, several studies have highlighted its limitations, particularly when highly divergent sequences are analyzed. Alignment-free methods attempt to overcome these limitations by representing biological sequences through numerical or statistical features rather than explicit alignments. A common strategy consists of transforming sequences into vectors of k-mer frequencies and computing similarity measures directly in the resulting feature space. These approaches often provide significant computational advantages and have become increasingly attractive for large-scale genomic analyses [27, 28].

3. The hierarchical distance model

Unlike approaches based on concatenated sequences, the proposed framework reconstructs phylogenetic relationships through a hierarchical aggregation of genetic information. From aligned allele sequences, pairwise distances are first computed at the allele level. These measures are then combined to obtain gene-level comparisons between individuals. The resulting gene-level estimates are subsequently aggregated at the chromosome level and finally integrated into a GW dissimilarity matrix. This matrix is then used to infer phylogenetic trees using the NJ algorithm.

Let $P = \{p_1, p_2, \dots, p_m\}$ be a population of m diploid individuals involving a gene set $\mathscr{G} = \{g_1, g_2, \dots, g_n\}$. For any $g \in \mathscr{G}$, we denote by $\mathscr{A}^g = \{A_1^g, A_2^g, \dots, A_{a(g)}^g\}$ the set of observed distinct alleles of g. For each $g \in \mathscr{G}$, let $S^g = \{1, 2, \dots, s(g)\}$ be the set of variable nucleotide positions (SNP loci) characterizing the gene. Each $A_k^g \in \mathscr{A}^g$ is defined as a sequence of length $s(g)$ over the nucleotide alphabet $\Sigma = \{A, C, G, T\}$. Thus, an A_k^g is represented as $A_k^g = (v_{k,1}^g, v_{k,2}^g, \dots, v_{k,s(g)}^g)$, where $v_{k,j}^g \in \Sigma$ is the specific nucleotide variant of A_k^g at the j-th SNP locus. The distance between sequences of two alleles A_i^g and A_j^g is defined as:

$$\delta_{\mathscr{A}}\left(A_i^g, A_j^g\right) = \frac{1}{s(g)} \sum_{r=1}^{s(g)} \left[v_{i,r}^g \neq v_{j,r}^g\right] \tag{1}$$

corresponding to the proportion of positions in which the two aligned allele sequences differ.

Let $(A_i^g, A_j^g)_k$ denote the genotype of individual k at gene g, with $(A_i^g, A_j^g)_k = (A_j^g, A_i^g)_k$. We define the distance between two genotypes $(A_i^g, A_j^g)_k$ and $(A_x^g, A_y^g)_m$ based on the distances between the alleles composing their haplotypes. It is computed as the minimum over the two possible allele matchings, that is:

$$\delta_{\mathscr{H}}\left((A_i^g, A_j^g)_k, (A_x^g, A_y^g)_m\right) = \min\left\{\delta_{\mathscr{A}}(A_i^g, A_x^g) + \delta_{\mathscr{A}}(A_j^g, A_y^g), \delta_{\mathscr{A}}(A_i^g, A_y^g) + \delta_{\mathscr{A}}(A_j^g, A_x^g)\right\}. \tag{2}$$

The minimum is required because genotypes are represented as unordered pairs of alleles, and therefore both possible matches must be considered. However, in some datasets, a haplotype can be associated with multiple alternative allele assignments due to multiallelic loci, structural variation, or uncertainty in variant calling. Under these conditions, population data may contain several alternative allele assignments for the same haplotype block [29]. In this case, all compatible genotype configurations are generated and the distance between two individuals is computed as the average distance across all genotype pairs. Formally, let

$$\mathscr{P}_k^g = \left(\{A_{i_1}^g, A_{i_2}^g, \dots, A_{i_{n_i}}^g\}, \{A_{j_1}^g, A_{j_2}^g, \dots, A_{j_{n_j}}^g\}\right)_k$$

represent an individual whose two haplotypes are associated with multiple alternative allele assignments for the gene g. The set of all possible genotypes is defined as:

$$\mathscr{G}(\mathscr{P}_k^g) = \{A_{i_1}^g, A_{i_2}^g, \dots, A_{i_{n_i}}^g\} \times \{A_{j_1}^g, A_{j_2}^g, \dots, A_{j_{n_j}}^g\},$$

and:

$$\delta_{\mathscr{M}}(\mathscr{P}_k^g, \mathscr{P}_m^g) = \frac{1}{|\mathscr{G}(\mathscr{P}_k^g)||\mathscr{G}(\mathscr{P}_m^g)|} \sum_{a_k \in \mathscr{G}(\mathscr{P}_k^g)} \sum_{a_m \in \mathscr{G}(\mathscr{P}_m^g)} \delta_{\mathscr{H}}(a_k, a_m). \tag{3}$$

Equation 3 generalizes the comparison to the case in which each haplotype is associated with multiple alternative allele assignments. Ambiguous haplotype assignments arise when multiple allele combinations are compatible with the observed sequencing data, for example because of multiallelic loci or uncertainty in allele assignment. Rather than selecting a single match, it estimates the expected genotype dissimilarity by averaging the distances over all compatible genotype configurations.

Once gene-level distances have been computed, chromosome distances are obtained by averaging all genotype distances belonging to the same chromosome. Let $\mathscr{C}_i$ denote the set of genes that belong to the chromosome c_i. The chromosome distance between individuals p_k and p_m, for chromosome c_i, is defined as:

$$\delta_{\mathscr{C}}(p_k, p_m, i) = \frac{1}{|\mathscr{C}_i|} \sum_{g \in \mathscr{C}_i} \delta_{\mathscr{M}}(\mathscr{P}_k^g, \mathscr{P}_m^g). \tag{4}$$

The use of equal weights is a deliberate design choice aimed at preventing the final GW distance from being dominated by genes with a larger number of polymorphic sites or by chromosomes containing more annotated genes. Since the objective of the proposed framework is to preserve the hierarchical organization of genomic information, each biological entity contributes equally to the aggregation process. Alternative weighting strategies, for example based on gene length, SNP density, or functional relevance, could also be considered, but would introduce additional biological assumptions that are beyond the scope of the present work.

Finally, to obtain a GW representation of genetic relationships, Equation 4 is aggregated across all t chromosomes in a genome. The GW distance between two individuals is defined as:

$$\delta_{\mathscr{I}}(p_k, p_m) = \frac{1}{t} \sum_{i=1}^{t} \delta_{\mathscr{C}}(p_k, p_m, i). \tag{5}$$

Equation 5 represents the overall genetic distance between all pairs of individuals in the population P and provides the basis for phylogenetic reconstruction by means of the NJ algorithm [22].

3.1. Computational Complexity

The computational cost of the hierarchical distance model can be analyzed considering each level of aggregation separately. Let $L = \sum_{g=1}^{n} s(g)$ be the total number of polymorphic positions considered across the genome. The distance $\delta_{\mathscr{A}}$ between the sequence of two alleles is calculated on the aligned SNP positions of a gene by means of Equation 1. Therefore, the cost of evaluating a single allele distance for gene g is $O(s(g))$. For diploid individuals, Equation 2 requires the evaluation of the two possible allele matches. Since each matching involves two allele-distance computations, the complexity of a genotype comparison remains $O(s(g))$. In the presence of multi-allelic haplotypes, Equation 3, let $r_{p_i}^{g}$ and $r_{p_j}^{g}$ represent the numbers of compatible genotype configurations for individuals p_i and p_j, respectively. Considering a single gene g, the complexity of comparing two genotypes becomes $O\left(r_{p_i}^{g} r_{p_j}^{g} s(g)\right)$. The aggregation steps defined by Equation 4 and Equation 5 consist only of arithmetic averages. Consequently, their computational cost is linear in the number of distances being combined and is negligible with respect to the cost of genotype comparisons. To compute the complete pairwise distance matrix used for phylogenetic reconstruction, distances must be evaluated for all pairs of individuals. Since there are $\binom{m}{2} = O(m^2)$ such pairs, the overall complexity of constructing the genome-wide distance matrix is $O\left(m^2 \sum_{g=1}^{n} r_{p_i}^{g} r_{p_j}^{g} s(g)\right)$. In the common diploid setting without multi-allelic ambiguity, this simplifies to $O(m^2 L)$, that is, quadratic in the number of individuals and linear in the total number of SNP positions. Once the distance matrix has been computed, phylogenetic reconstruction is performed using the NJ algorithm, which requires $O(m^3)$ and $O(m^2)$ memory. Therefore, the overall complexity of the complete framework is $O(m^2 L + m^3)$. Since genome-wide datasets typically satisfy $L \gg m$, the distance computation stage generally dominates the total running time, whereas the tree reconstruction phase becomes comparatively less expensive. Therefore, the framework scales linearly with the number of polymorphic sites and quadratically with the number of individuals.

4. Experimental Results

4.1. Dataset

We apply our hierarchical distance model to a real dataset derived from 98 cultivars representative of the variability present in *Vitis vinifera* developed in [30]. It contains information on allele-specific haplotypes derived from RNA-seq data for each individual. This dataset includes SNP information for genes located on all 19 nuclear chromosomes. For each gene, two complementary data structures are available. The first contains aligned allele sequences, whereas the second stores haplotype assignments for each individual. Haplotypes may consist of either a single allele or a set of alternative alleles. A small number of genes contained incomplete haplotype assignments. Whenever one haplotype was unavailable and the other was present, the corresponding individual has been treated as heterozygous. Since incomplete assignments affected only a limited number of genes, their impact on the overall GW distance is expected to be negligible.

4.2. Genome-Wide Phylogenetic Reconstruction

A total of 20 phylogenetic trees have been reconstructed by employing NJ: one for each chromosome and a GW tree obtained by aggregating information across all chromosomes. Figure 1 shows the GW phylogenetic tree. It represents the genetic relationships among grapevine cultivars, providing an interpretative framework for understanding the evolutionary dynamics, domestication processes, and geographical spread of cultivated grapevines. Analysis of the tree allows the identification of several major clades, reflecting both the evolutionary history of the species and the influence of human selection.

A first group, located in the upper portion of the tree, includes cultivars originating from the Caucasus region as well as widely distributed international varieties such as *Cabernet Franc*, *Cabernet Sauvignon*, and *Merlot*. The presence of Caucasian cultivars in basal or otherwise well-represented positions

supports the widely accepted hypothesis that the Caucasus constitutes one of the primary centers of grapevine domestication [31]. Furthermore, the close phylogenetic proximity between *Cabernet Sauvignon* and *Cabernet Franc* reflects well-established genetic evidence indicating that the former originated from a cross between *Cabernet Franc* and *Sauvignon Blanc* [32]. This example demonstrates how the phylogenetic tree is capable of accurately representing known genealogical relationships that have been experimentally validated.

A second group, located in the central portion of the tree, includes typically Mediterranean varieties such as *Moscato Bianco*, *Vermentino*, *Primitivo*, and *Teran*. This clade appears more heterogeneous than the previous one, suggesting an evolutionary history characterized by multiple dispersal events and local selection processes. The distribution of these cultivars is consistent with the historical routes of viticulture expansion throughout the Mediterranean basin, particularly through Phoenician and Greek colonization [33]. In this context, the observed genetic diversity can be interpreted as the result of environmental adaptation and differentiated agricultural practices.

Particularly noteworthy is the clade grouping numerous Italian grapevine cultivars, including *Sangiovese*, *Trebbiano*, *Verdicchio*, *Aglianico*, and *Garganega*. These individuals form a relatively compact cluster, indicative of a shared genetic background and strong regional structuring. This evidence is supported by recent studies showing that Italian grapevine cultivars exhibit significant genetic homogeneity, resulting from both local selection processes and clonal propagation [34]. The presence of this cluster further suggests that, following the introduction of grapevines into Italy, a phase of internal diversification occurred, driven by environmental and anthropogenic factors.

The analysis of relationships among taxa also allows the identification of numerous sister-group relationships, namely pairs or groups of varieties sharing a recent common ancestor. In addition to the previously mentioned case of *Cabernet Sauvignon* and *Cabernet Franc*, close relationships can be observed among several cultivars belonging to the same geographical context. This pattern reflects the central role of human selection in shaping grapevine genetic diversity through practices such as hybridization, selection, and clonal propagation. As highlighted by [35], the identification of sister groups represents one of the principal tools for interpreting evolutionary relationships within a phylogenetic tree.

Branch lengths provide additional information regarding the degree of genetic divergence among cultivars. Some taxa are characterized by particularly long branches, indicating greater genetic distances relative to others. This phenomenon may be attributed to several factors, including ancient evolutionary separations, geographic isolation, or events of genetic introgression. In grapevine, the complexity of its evolutionary history is further enhanced by the practice of vegetative propagation, which has contributed to the long-term maintenance of distinct yet closely related genotypes [36].

From an evolutionary perspective, the tree supports a model in which domesticated grapevine originated in an eastern region, most likely the Caucasus, and subsequently spread progressively toward Europe and the Mediterranean basin. During this process, the species underwent strong anthropogenic influence, leading to the formation of distinct regional groups. The combination of geographical diffusion and human selection has therefore shaped the current genetic structure of cultivated grapevine populations [37]. The phylogenetic tree analyzed here provides a coherent representation of the genetic diversity of *Vitis vinifera* cultivars, highlighting the presence of distinct clades associated with specific geographical regions and viticultural traditions. The observed relationships are largely consistent with current historical and genetic knowledge, supporting the hypothesis of an eastern origin of grapevine followed by its spread into Europe and subsequent regional differentiation.

4.3. Chromosome-Level Tree Comparison

With the aim to compare phylogenetic trees, two complementary aspects can be considered: evolutionary similarity and topological similarity. Evolutionary similarity evaluates whether the inferred branch lengths and pairwise evolutionary relationships among taxa are preserved, whereas topological similarity focuses exclusively on the branching structure of the trees. To investigate how phylogenetic information is distributed across the genome, all chromosome-specific NJ trees have been compared with each other and with the GW reconstruction.

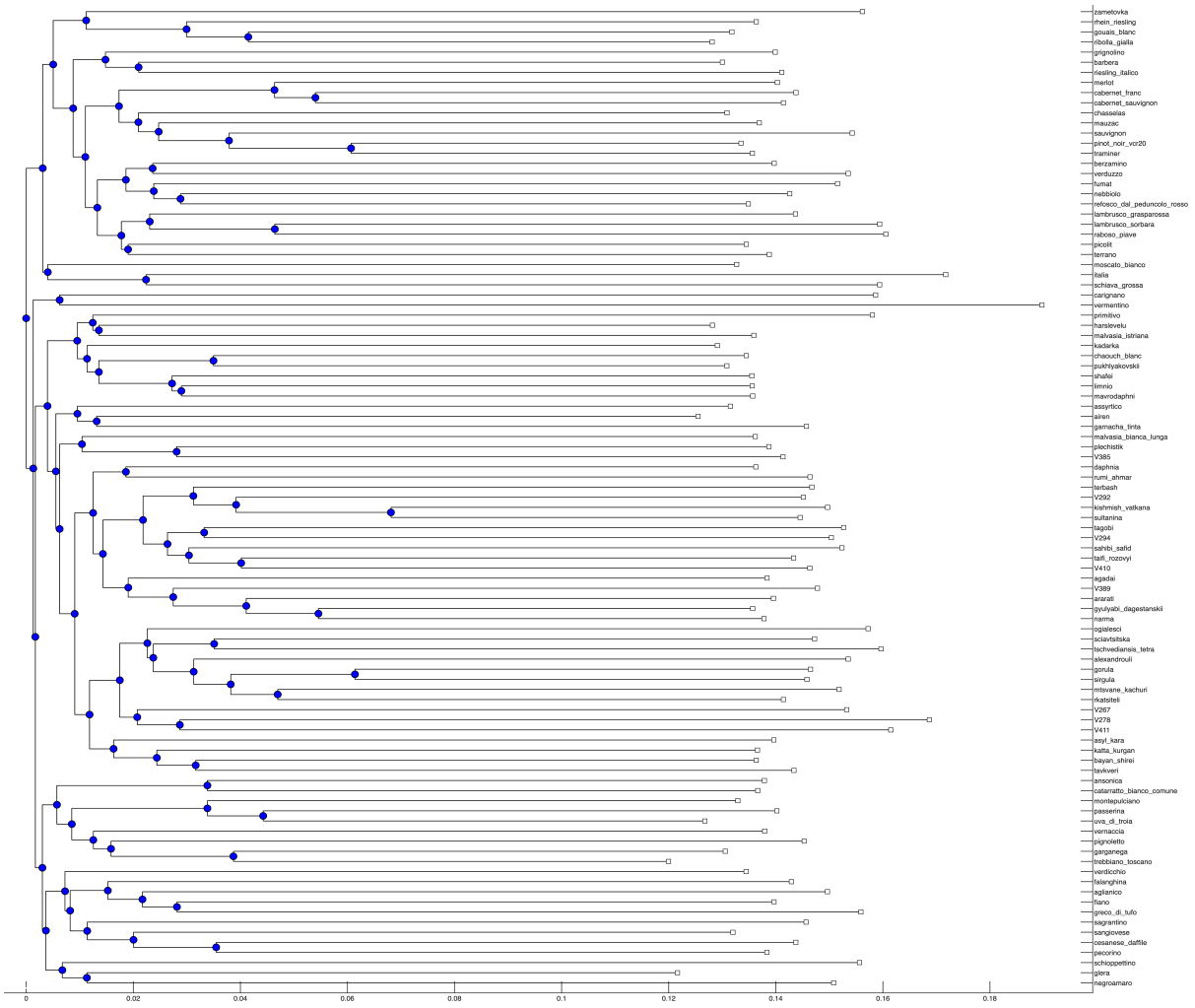

Figure 1: Phylogenetic tree inferred through NJ by employing the developed genome-wide distance.

Evolutionary distances between trees were evaluated and the entire process consisted of two stages. Initially, a matrix was calculated for each tree, where every entry represents the branch-length distance between two leaves. Subsequently, evolutionary distances were measured through the Frobenius norm applied to the difference between all pairs of matrices obtained in the prior stage, without normalizing any matrix. Evolutionary similarity evaluates whether the inferred branch lengths and pairwise evolutionary relationships among taxa are preserved, whereas topological similarity focuses exclusively on the branching structure of the trees. The resulting distances are shown in Figure 2. The analysis reveals substantial variability among chromosome-specific phylogenies. Pairwise evolutionary distances range between 2 and 7.5, indicating that different chromosomes may capture distinct evolutionary signals. In particular, chromosome 9 exhibits some of the largest deviations from the remaining trees, especially when compared with chromosomes 6 and 17. Conversely, the phylogeny reconstructed from chromosome 1 shows one of the smallest evolutionary distances from the GW tree, suggesting that it captures a representative fraction of the overall genetic signal.

Topological similarities have been further evaluated through the Jaccard–Robinson–Foulds (JRF) metric [38] (Figure 3). While the Frobenius norm compares pairwise leaf-distance matrices and is therefore sensitive to differences in branch lengths, the JRF metric ignores these lengths and quantifies only the similarity of the tree topology through the agreement of internal bipartitions. Although all chromosome-level trees differ substantially from each other, the observed dissimilarities remain within a relatively narrow interval. The average topological dissimilarity is approximately 64%, while values never exceed 70%. Interestingly, chromosome 7 recovers nearly half of the clades observed

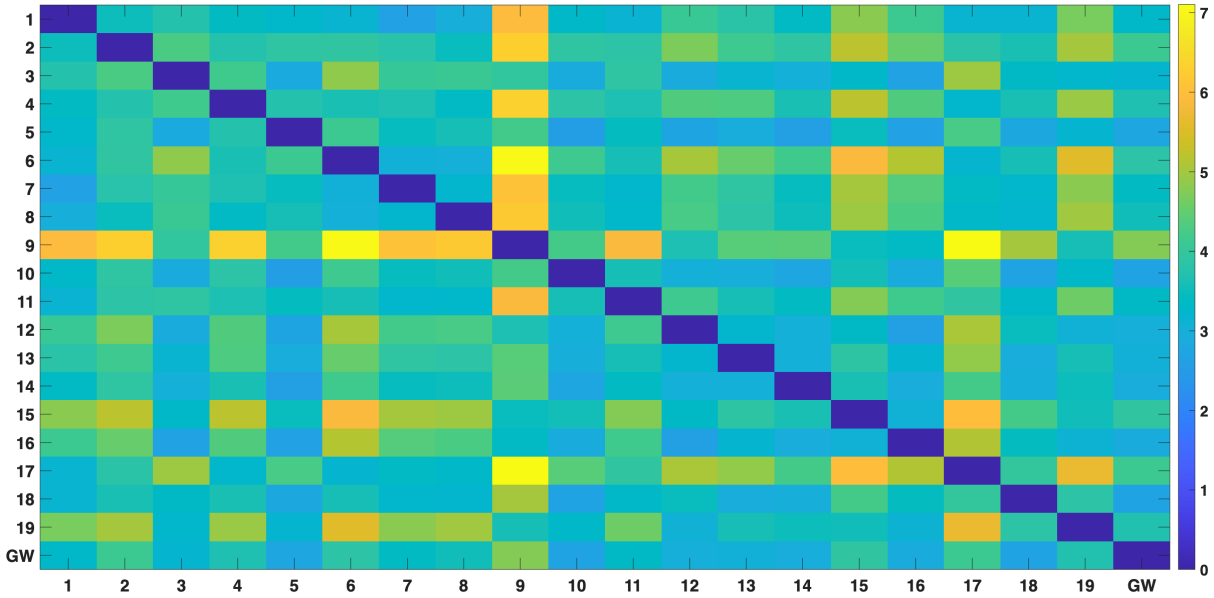

Figure 2: Evolutionary distances between pairs of trees constructed with NJ. Numbers represent chromosome-specific trees, while GW indicates genome-wide one.

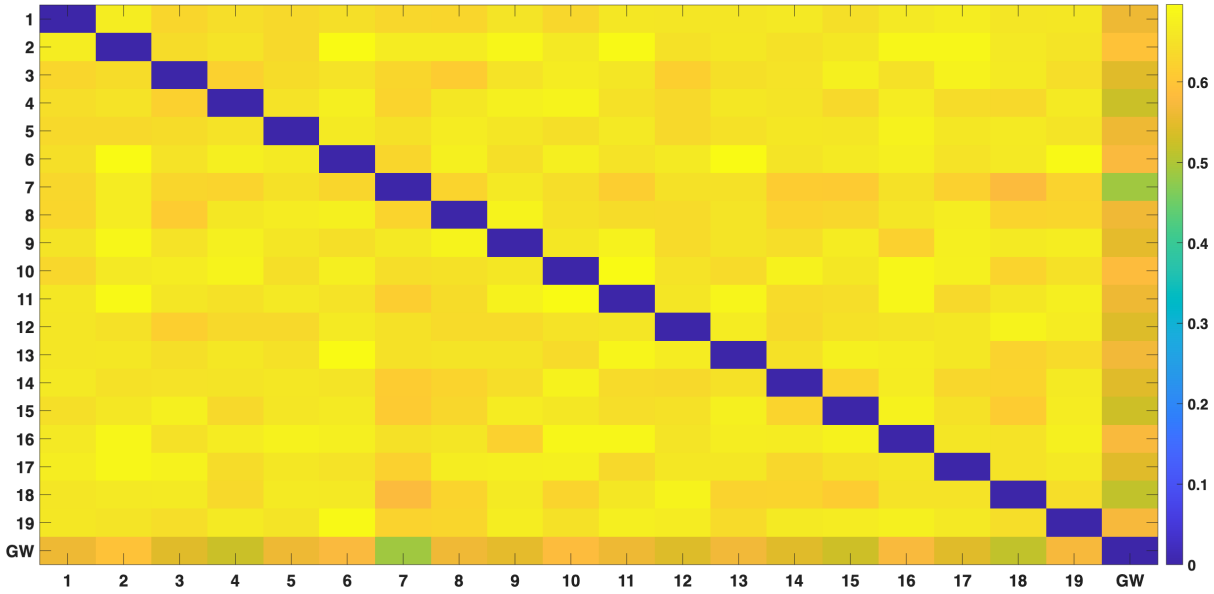

Figure 3: Topological similarity between pairs of trees constructed with NJ. Numbers represent chromosome-specific trees, while GW indicates genome-wide one.

in the GW reconstruction, suggesting that some chromosomes capture a substantial fraction of the global phylogenetic signal. Nevertheless, no single chromosome reproduces the complete genome-wide topology, supporting the need for hierarchical aggregation. Overall, the observed variability may indicate that single chromosomes capture only partial evolutionary histories, whereas the proposed hierarchical aggregation integrates these complementary signals into a more robust genome-wide representation.

4.4. Comparison between Neighbor-Joining and Balanced Minimum Evolution

To evaluate the robustness of the proposed distance model with respect to the reconstruction algorithm, each NJ tree has been compared with the corresponding BME tree obtained from the same distance matrix. Figure 4 reports the evolutionary differences between corresponding NJ and BME reconstructions.

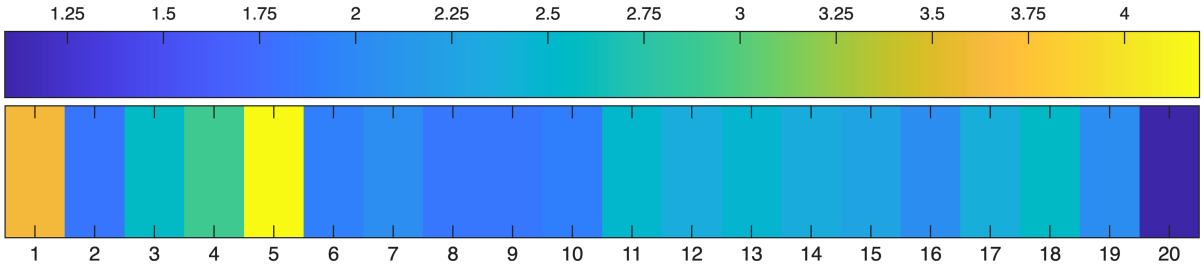

Figure 4: Comparison of evolutionary distances between trees built with NJ and BME. 20 represents GW tree.

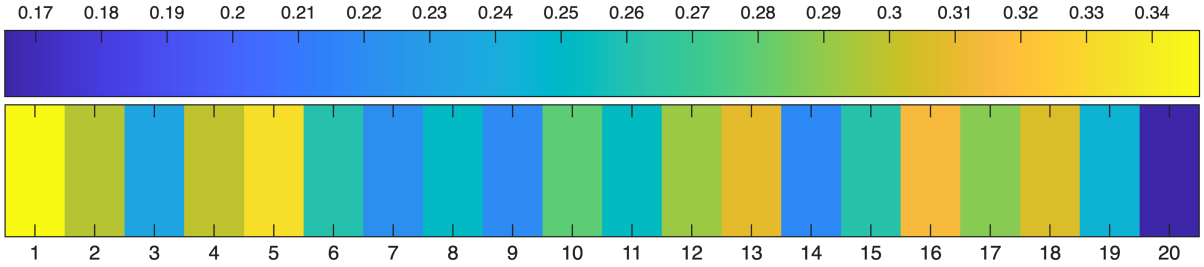

Figure 5: Comparison of topological similarity between trees constructed with NJ and BME. 20 represents GW tree.

Overall, the observed distances remain relatively low. Most chromosome-level comparisons exhibit values close to 2, indicating that both methods infer similar evolutionary relationships among taxa. The largest differences are observed for chromosomes 1 and 5, whose distances exceed 3.5.

Topological comparisons are shown in Figure 5. The topological dissimilarity ranges approximately between 16% and 35%, with an average value of about 27%. The highest agreement is obtained for the GW trees, whose dissimilarity drops below 17%. These results suggest that NJ and BME produce comparable evolutionary reconstructions despite moderate differences in topology. In particular, genome-wide distances appear to mitigate methodological differences and yield more stable phylogenetic structures. The consistency observed across two different reconstruction algorithms suggests that the proposed distance, rather than the inference algorithm itself, is the primary source of the recovered phylogenetic signal.

4.5. Chromosome Relationships and Multidimensional Scaling

We then tried to infer whether distinct chromosomes encode consistent and recurrent evolutionary histories. We analyzed the two complementary distance matrices, Figure 2 and Figure 3, by means of hierarchical clustering, to explore relationships among chromosome-specific phylogenies and to identify groups of chromosomes sharing similar evolutionary signals. If similar groups of chromosomes emerge independently from both topological and evolutionary distance measures, the resulting structure may be regarded as robust. Figure 6 shows the dendrogram obtained from the JRF topological distances. The most evident result is that the GW tree clusters first with chromosome 7. This observation is fully consistent with the results reported previously, where chromosome 7 was found to recover approximately half of the clades present in the genome-wide phylogeny. The hierarchical clustering therefore confirms that chromosome 7 is the most similar chromosome to the genome-wide reconstruction in terms of topology. Furthermore, chromosomes 18 and 15 appear in the same cluster as GW and chromosome 7, suggesting the existence of a group of chromosomes sharing similar topological structures. A different pattern emerges when considering the dendrogram derived from Frobenius distances (Figure 7). In this case, the genome-wide tree is no longer dominated by chromosome 7. Instead, chromosome 18 appears substantially closer to the genome-wide reconstruction. Taken together, these results indicate that topological and evolutionary tree-distance measures emphasize different properties of the reconstructed phylogenies. While this observation may reflect complementary aspects of the inferred evolutionary

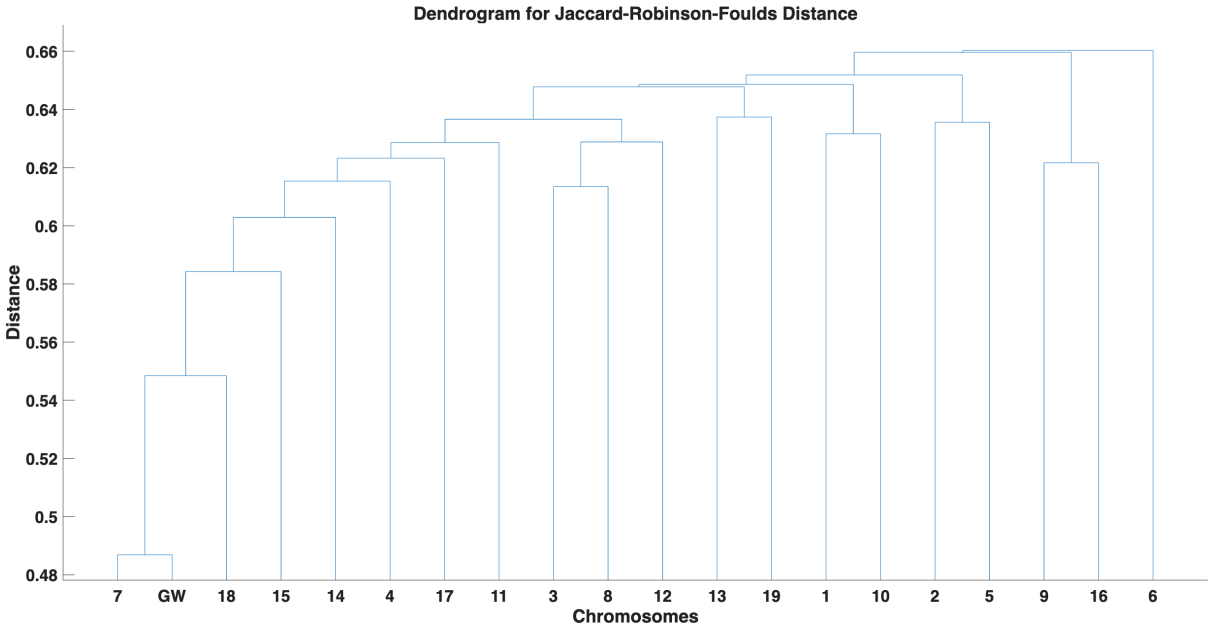

Figure 6: Hierarchical clustering of chromosome-specific phylogenies based on JRF distances. The dendrogram highlights groups of chromosomes sharing similar phylogenetic topologies.

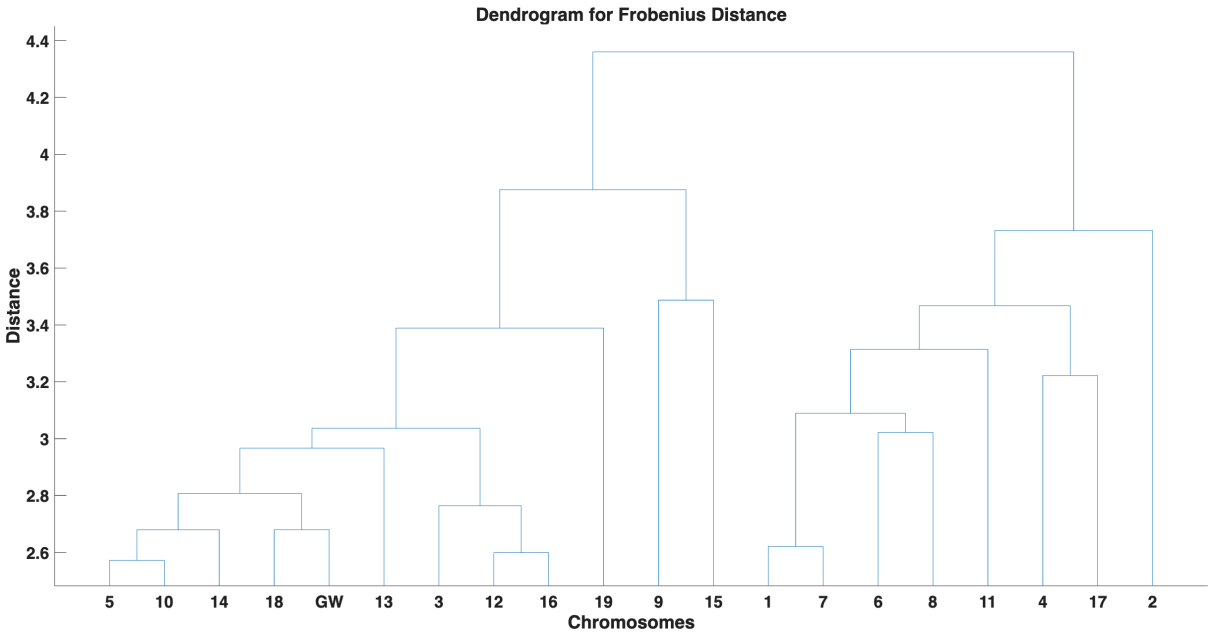

Figure 7: Hierarchical clustering of chromosome-specific phylogenies based on Frobenius distances computed from pairwise leaf-distance matrices.

relationships, alternative explanations cannot be excluded, including the different sensitivities of the adopted tree-distance measures and the intrinsic variability of chromosome-specific phylogenetic reconstructions. Chromosome 7 appears particularly effective at preserving clade structure, whereas chromosome 18 provides a closer approximation of the evolutionary distances inferred from the complete genome.

To further investigate the global organization of the phylogenetic space, Non-metric Multidimensional Scaling (MDS) has been applied to the tree-distance matrix [39]. MDS projects the trees into a low-dimensional space while preserving the relative ordering of pairwise distances, thus providing an intuitive visualization of higher-dimensional relationships. Therefore, MDS has been employed solely as a visualization tool to provide an intuitive representation of the relationships among chromosome-

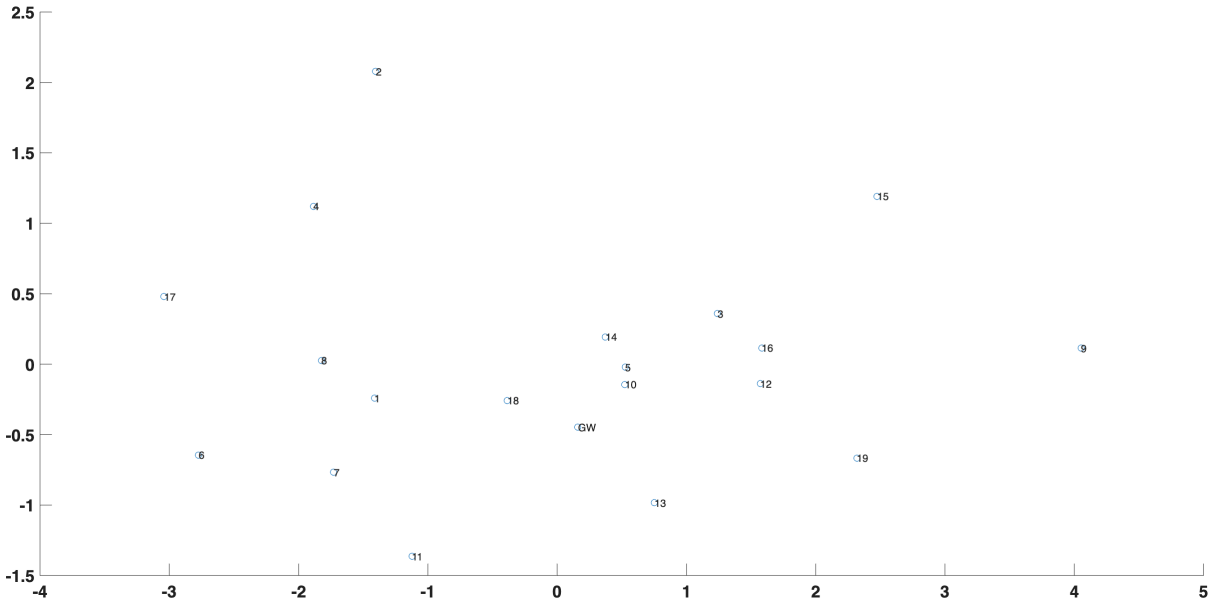

Figure 8: MDS projection of chromosome-specific and genome-wide phylogenetic trees. Distances between points reflect similarities among phylogenies in the original tree-distance space.

specific phylogenies. Three major patterns can be observed (Figure 8). First, a central group composed of GW together with chromosomes 3, 5, 10, 12, 14, 16, and 18 occupies the densest region of the projection. The presence of multiple chromosomes surrounding the GW tree suggests that the genome-wide phylogeny is not dominated by a single chromosome. Instead, it appears to emerge as a consensus signal resulting from the combined contribution of several chromosomes. Second, a number of chromosomes occupy peripheral positions, including chromosomes 2, 9, 11, 15, and 17. Their separation from the central cluster may indicate the presence of distinctive phylogenetic signals that are not fully represented by the genome-wide reconstruction. Third, the GW tree itself is located near the center of the configuration rather than at its boundary. If a single chromosome were primarily responsible for the genome-wide structure, one would expect the GW point to overlap closely with that chromosome. Instead, its central position supports the interpretation of the genome-wide phylogeny as an aggregate representation of multiple chromosomal signals. Interestingly, some of the chromosomes occupying peripheral positions have previously been associated with traits related to domestication, environmental adaptation, fruit quality, and pathogen resistance. While the present study does not directly test functional hypotheses, the observed distribution suggests that selective pressures may have affected different chromosomal regions in distinct ways. From a broader perspective, the MDS configuration reveals a structured phylogenetic landscape rather than a homogeneous cloud of chromosome-specific trees. The observed structure is compatible with the hypothesis that different chromosomes preserve distinct evolutionary signals. However, further analyses on additional datasets and species would be required to assess whether this pattern reflects biological processes rather than methodological effects. From a methodological perspective, the central position of the genome-wide tree further supports the proposed hierarchical aggregation strategy. Rather than reflecting the evolutionary history encoded by a single chromosome, the GW reconstruction emerges as a consensus representation integrating complementary chromosomal signals.

Taken together, these complementary analyses indicate that the proposed hierarchical distance produces phylogenetic reconstructions that are consistent both in terms of evolutionary distances and tree topology, while highlighting chromosome-specific variability. Although the experimental results are encouraging, they are currently limited to a single population-scale dataset. Additional evaluations on datasets from different species and genomic contexts will be necessary to assess the generality of the proposed framework.

5. Conclusions

Most existing phylogenetic approaches rely either on aligned nucleotide sequences or on classical genetic distance measures computed directly from sequence data. In contrast, the proposed framework operates on allele-specific haplotypes inferred from population-scale genomic data. Rather than defining distances directly between sequences, we introduced a hierarchical distance model that aggregates information at multiple biological levels, namely alleles, genes, chromosomes, and complete genomes. This allows genome-wide phylogenetic reconstruction while preserving haplotype-level information and enables a quantitative comparison of chromosome-specific and GW evolutionary signals.

Experimental results on the analyzed *Vitis vinifera* dataset indicate that the proposed framework is capable of producing biologically meaningful phylogenetic reconstructions while preserving chromosome-specific evolutionary information. The analysis revealed substantial variability among chromosome-level phylogenies and greater stability for GW reconstructions, suggesting that different genomic regions contribute complementary evolutionary signals. Furthermore, the comparison between NJ and BME demonstrated that the proposed distance model yields consistent evolutionary patterns across different reconstruction algorithms.

From a computational perspective, the framework provides a flexible and scalable approach for exploiting haplotype-resolved genomic data in phylogenetic studies. Unlike conventional distance measures, which are computed directly between complete sequences, the proposed framework preserves the hierarchical organization of genomic information throughout the entire reconstruction process. This makes it possible to analyze evolutionary relationships at multiple biological scales while naturally extending from allele-level variation to GW phylogenies. This aggregation is performed at biologically meaningful levels, rather than after concatenating all sequence information into a single representation. It preserves the contribution of each biological level and enables chromosome-specific analyses that would not be possible using a single concatenated sequence. This design also makes it possible to inspect evolutionary relationships at intermediate biological scales, enabling chromosome-level analyses alongside GW reconstruction.

An additional factor that may influence chromosome-specific phylogenetic reconstructions is chromosome size, or more generally the amount of available genetic information. Although this aspect was beyond the scope of the present study, investigating possible relationships between chromosome characteristics and phylogenetic behavior represents an interesting direction for future work.

Although these results are encouraging, they should be interpreted in the context of the analyzed dataset. A broader experimental assessment involving additional species and population-scale datasets will be necessary to further evaluate the general applicability and robustness of the proposed approach. Moreover, we would like to point out that the experimental evaluation was designed to assess the behavior and robustness of the proposed hierarchical aggregation framework. While the obtained phylogenetic reconstructions are biologically consistent, the study does not aim to provide a comprehensive benchmark against existing SNP- or haplotype-based distance measures. Such a comparative evaluation would further clarify the strengths and limitations of the proposed framework and represents an interesting direction for future research. Finally, future work will investigate alternative aggregation strategies and weighted distance formulations. An additional direction concerns the integration of functional and expression-related information, with the objective of combining evolutionary reconstruction and genotype-specific functional characterization within a unified computational framework.

Acknowledgments

The work of Roberto Pagliarini is partially supported by the Complementary Operational Programme of the Autonomous Region Friuli-Venezia Giulia 2014-2020, CUP G23C25000880002 – OPERATION CODE 2025/13846 – PROJECT CODE 2025/13846/15.
Roberto Pagliarini is member of the GNCS group of INdAM.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. De Bruyn, D. P. Martin, P. Lefeuvre, Phylogenetic Reconstruction Methods: An Overview, Humana Press, Totowa, NJ, 2014, pp. 257–277. URL: https://doi.org/10.1007/978-1-62703-767-9_13. doi:10.1007/978-1-62703-767-9_13.

[2] Y. Zou, Z. Zhang, Y. Zeng, H. Hu, Y. Hao, S. Huang, B. Li, Common methods for phylogenetic tree construction and their implementation in r., Bioengineering (Basel) 11 (2024). doi:10.3390/bioengineering11050480.

[3] W. Cao, L.-Y. Wu, X.-Y. Xia, X. Chen, Z.-X. Wang, X.-M. Pan, A sequence-based evolutionary distance method for phylogenetic analysis of highly divergent proteins, Scientific Reports 13 (2023) 20304. URL: https://doi.org/10.1038/s41598-023-47496-9. doi:10.1038/s41598-023-47496-9.

[4] W.-F. Yang, Z.-G. Yu, V. Anh, Whole genome/proteome based phylogeny reconstruction for prokaryotes using higher order markov model and chaos game representation, Molecular Phylogenetics and Evolution 96 (2016) 102–111. URL: https://www.sciencedirect.com/science/article/pii/S1055790315003905. doi:https://doi.org/10.1016/j.ympev.2015.12.011.

[5] I. A. Diaz, T. Ostovar, J. Chen, E. A. de Dios, R. Piscatella, R. S. Perez-Alfaro, O. Zayed, S. Saddoris, R. J. Schmitz, S. R. Wessler, J. E. Stajich, D. K. Seymour, A haplotype-resolved chromatin landscape connects cis-regulatory variants to trait variation in citrus., BMC Genomics 26 (2025) 978. doi:10.1186/s12864-025-12137-0.

[6] J.-Y. Guk, M.-J. Jang, J.-W. Choi, Y. M. Lee, S. Kim, De novo phasing resolves haplotype sequences in complex plant genomes, Plant Biotechnology Journal 20 (2022) 1031–1041. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/pbi.13815. doi:https://doi.org/10.1111/pbi.13815. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/pbi.13815.

[7] R. Pagliarini, F. Marroni, C. Piazza, G. Gabelli, G. Magris, G. Di Gaspero, M. Morgante, A. Policriti, Towards a computational approach to quantification of allele specific expression at population level, in: I. Rojas, F. Ortuño, F. Rojas, L. J. Herrera, O. Valenzuela (Eds.), Bioinformatics and Biomedical Engineering, Springer Nature Switzerland, Cham, 2024, pp. 127–139.

[8] R. Pagliarini, F. Nascimben, A. Policriti, Self-organizing maps for allele specific expression data reconstruction and identification of anomalous genomic regions, Frontiers in Bioinformatics Volume 6 - 2026 (2026). URL: https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2026.1810835. doi:10.3389/fbinf.2026.1810835.

[9] G. Munjal, M. Hanmandlu, S. Srivastava, Phylogenetics Algorithms and Applications, 2019, pp. 187–194. doi:10.1007/978-981-13-5934-7_17.

[10] S. Wang, M. Gribskov, Multiple Sequence Alignment, Springer Nature Singapore, Singapore, 2025, pp. 63–72. URL: https://doi.org/10.1007/978-981-96-7949-2_8. doi:10.1007/978-981-96-7949-2_8.

[11] J. D. Thompson, D. G. Higgins, T. J. Gibson, Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice., Nucleic Acids Res 22 (1994) 4673–4680. doi:10.1093/nar/22.22.4673.

[12] F. Sievers, D. G. Higgins, Clustal omega for making accurate alignments of many protein sequences., Protein Sci 27 (2018) 135–145. doi:10.1002/pro.3290.

[13] K. Katoh, D. M. Standley, Mafft multiple sequence alignment software version 7: improvements in performance and usability., Mol Biol Evol 30 (2013) 772–780. doi:10.1093/molbev/mst010.

[14] J. Sabonsolin, D. Lao, A comprehensive systematic literature review of multiple sequence alignment algorithms, Discover Computing 29 (2026) 33. URL: https://doi.org/10.1007/s10791-026-09911-3. doi:10.1007/s10791-026-09911-3.

[15] S. B. Needleman, C. D. Wunsch, A general method applicable to the search for similarities in

the amino acid sequence of two proteins, Journal of Molecular Biology 48 (1970) 443–453. URL: https://www.sciencedirect.com/science/article/pii/0022283670900574. doi:https://doi.org/10.1016/0022-2836(70)90057-4.

[16] C. Notredame, D. G. Higgins, J. Heringa, T-coffee: A novel method for fast and accurate multiple sequence alignment., J Mol Biol 302 (2000) 205–217. doi:10.1006/jmbi.2000.4042.

[17] X. Xia, Maximum Parsimony Method in Phylogenetics, Springer International Publishing, Cham, 2018, pp. 327–341. URL: https://doi.org/10.1007/978-3-319-90684-3_14. doi:10.1007/978-3-319-90684-3_14.

[18] T. H. Jukes, C. R. Cantor, Chapter 24 - evolution of protein molecules, in: H. Munro (Ed.), Mammalian Protein Metabolism, Academic Press, 1969, pp. 21–132. URL: https://www.sciencedirect.com/science/article/pii/B9781483232119500097. doi:https://doi.org/10.1016/B978-1-4832-3211-9.50009-7.

[19] M. K. Gupta, R. Vadde, Next-generation development and application of codon model in evolution., Front Genet 14 (2023) 1091575. doi:10.3389/fgene.2023.1091575.

[20] S. Tavaré, Some probabilistic and statistical problems in the analysis of dna sequences, 1986. URL: https://api.semanticscholar.org/CorpusID:82212051.

[21] X. Xia, Distance-Based Phylogenetic Methods, Springer International Publishing, Cham, 2018, pp. 343–379. URL: https://doi.org/10.1007/978-3-319-90684-3_15. doi:10.1007/978-3-319-90684-3_15.

[22] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees., Mol Biol Evol 4 (1987) 406–425. doi:10.1093/oxfordjournals.molbev.a040454.

[23] I. Gronau, S. Moran, Optimal implementations of upgma and other common clustering algorithms, Information Processing Letters 104 (2007) 205–210. URL: https://www.sciencedirect.com/science/article/pii/S0020019007001767. doi:https://doi.org/10.1016/j.ipl.2007.07.002.

[24] R. R. Sokal, C. D. Michener, A statistical method for evaluating systematic relationships, University of Kansas science bulletin 38 (1958) 1409–1438. URL: https://api.semanticscholar.org/CorpusID:61950873.

[25] A. Rzhetsky, M. Nei, Theoretical foundation of the minimum-evolution method of phylogenetic inference, Molecular Biology and Evolution (1993).

[26] R. Desper, O. Gascuel, Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle, Journal of computational biology : a journal of computational molecular cell biology 9 (2002) 687–705. doi:10.1089/106652702761034136.

[27] S. Vinga, Editorial: Alignment-free methods in computational biology., Brief Bioinform 15 (2014) 341–342. doi:10.1093/bib/bbu005.

[28] A. Zielezinski, S. Vinga, J. Almeida, W. M. Karlowski, Alignment-free sequence comparison: benefits, applications, and tools., Genome Biol 18 (2017) 186. doi:10.1186/s13059-017-1319-7.

[29] I. M. Campbell, T. Gambin, S. Jhangiani, M. L. Grove, N. Veeraraghavan, D. M. Muzny, C. A. Shaw, R. A. Gibbs, E. Boerwinkle, F. Yu, J. R. Lupski, Multiallelic positions in the human genome: Challenges for genetic analyses., Hum Mutat 37 (2016) 231–234. doi:10.1002/humu.22944.

[30] R. Pagliarini, F. Nascimben, A. Policriti, A two-phase clustering procedure based on allele specific expression, BMC Bioinformatics 27 (2026) 76. URL: https://doi.org/10.1186/s12859-026-06398-z. doi:10.1186/s12859-026-06398-z.

[31] P. McGovern, M. Jalabadze, S. Batiuk, M. P. Callahan, K. E. Smith, G. R. Hall, E. Kvavadze, D. Maghradze, N. Rusishvili, L. Bouby, O. Failla, G. Cola, L. Mariani, E. Boaretto, R. Bacilieri, P. This, N. Wales, D. Lordkipanidze, Early neolithic wine of georgia in the south caucasus, Proceedings of the National Academy of Sciences 114 (2017) E10309–E10318. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1714728114. doi:10.1073/pnas.1714728114. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1714728114.

[32] J. Bowers, C. Meredith, Bowers je, meredith cp. the parentage of a classic wine grape, cabernet sauvignon. nat genet 16: 84-87, Nature genetics 16 (1997) 84–7. doi:10.1038/ng0597-84.

[33] P. This, T. Lacombe, M. R. Thomas, Historical origins and genetic diversity of wine grapes, Trends in Genetics 22 (2006) 511–519. URL: https://www.sciencedirect.com/science/article/pii/

S0168952506002253. doi:https://doi.org/10.1016/j.tig.2006.07.008.
[34] G. De Lorenzis, F. Mercati, C. Bergamini, M. Cardone, A. Lupini, A. Mauceri, A. Caputo, L. Abbate, M. Barbagallo, D. Antonacci, F. Sunseri, L. Brancadoro, Snp genotyping elucidates the genetic diversity of magna graecia grapevine germplasm and its historical origin and dissemination, BMC Plant Biology 19 (2019). doi:10.1186/s12870-018-1576-y.
[35] J. Felsenstein, Inferring phylogenies, Sinauer Associates, 2003.
[36] M. K. Aradhya, G. S. Dangl, B. H. Prins, J.-M. Boursiquot, M. A. Walker, C. P. Meredith, C. J. Simon, Genetic structure and differentiation in cultivated grape, vitis vinifera l., Genetical Research 81 (2003) 179–192. doi:10.1017/S0016672303006177.
[37] S. Myles, A. R. Boyko, C. L. Owens, P. J. Brown, F. Grassi, M. K. Aradhya, B. Prins, A. Reynolds, J.-M. Chia, D. Ware, C. D. Bustamante, E. S. Buckler, Genetic structure and domestication history of the grape, Proceedings of the National Academy of Sciences 108 (2011) 3530–3535. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1009363108. doi:10.1073/pnas.1009363108. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1009363108.
[38] M. Llabrés, F. Rosselló, G. Valiente, The generalized robinson-foulds distance for phylogenetic trees., J Comput Biol 28 (2021) 1181–1195. doi:10.1089/cmb.2021.0342.
[39] M. C. Hout, M. H. Papesh, S. D. Goldinger, Multidimensional scaling., Wiley Interdiscip Rev Cogn Sci 4 (2013) 93–103. doi:10.1002/wcs.1203.

A. Online Resources

The code associated with this manuscript and all the details about the inferred phylogenetic tree can be found at the following link: https://github.com/pasqualemarino/VitisViniferaPhytree.