



Mini review

A practical study of the proportion of non-unique neighbor-joining trees of microsatellite markers

Nàtalia Segura-Alabart ^{a,*}, Francesc Serratosa ^a, Alberto Fernández ^b

^a Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av. Països Catalans 26, Tarragona, 43007, Catalunya, Spain

^b Departament d'Enginyeria Química, Universitat Rovira i Virgili, Av. Països Catalans 26, Tarragona, 43007, Catalunya, Spain

ARTICLE INFO

Keywords:

Distance-based methods
Ties in proximity
Computational phylogenetics
SSR
STR

ABSTRACT

Neighbor-joining (NJ) is a distance-based method used to analyse and visualize data that generates a phylogenetic tree. It is found in a wide range of scientific research and is upstream of all fields of bioinformatics and computational biology. NJ's drawback is the generation of different binary phylogenetic trees if more than two clusters of elements have the same distance in any stage of the phylogenetic tree generation process. Therefore, articles that show only one phylogenetic tree, although NJ could generate several of them, are somewhat biased and hard to reproduce. The objective of this practical study is to quantify the extent of this problem in the literature in a particular context: the clustering of microsatellite markers with NJ. For this purpose, we calculate the proportion of publications in the Scopus database where the possibility of more than one binary phylogenetic tree exists. We have observed that 13% of articles can produce more than one binary phylogenetic tree; thus, results shown in these articles would have to be analysed or their conclusions revised. Moreover, this percentage is large enough to be considered a general problem in our scientific community. To exemplify the issue practically, we have selected an example where more than one binary phylogenetic tree solution is possible.

Contents

1. Introduction	2
2. Methods	2
2.1. Search strategy	2
2.2. Sample dataset	2
2.3. Non-unique binary phylogenetic trees	3
3. Results and discussion	3
3.1. Proportion of articles with ties	3
3.2. Publications per year	3
3.3. Publications per subject areas	3
3.4. Case study	3
4. Conclusions	4
CRediT authorship contribution statement	4
Declaration of competing interest	4
Acknowledgements	5
Appendix A. Supplementary data	5
References	5

* Corresponding author.

E-mail address: natalia.segura@urv.cat (N. Segura-Alabart).

1. Introduction

Microsatellites are molecular markers that can identify and characterize genetic variation between closely related genotypes [1]. They are also known as Short Tandem Repeats (STR) or Simple Sequence Repeats (SSR). Microsatellites are short DNA fragments of 2 to 6 base pairs that repeat in tandem and randomly in the genome [2]. They are commonly used because they are highly reproducible, codominant and multiallelic [3–5].

More specifically, microsatellite markers are used to measure the dissimilarity or distance between genotypes as a function of the proportion of shared alleles (e.g., 27 microsatellite markers are used in the case study analysed below to assess the genetic diversity). They are successful genetic tools due to the large number of alleles at a specific locus. Therefore, it often happens that different pairs of genotypes are separated by the same distance.

Phylogenetic trees are a graphical representation of the evolutionary relationships of the species under study. They are constructed using the matrix of pairwise distances extracted from the dataset using molecular markers. In Mathematics and Computer Science, it is well known the difficulty of reconstructing the correct phylogenetic tree corresponding to a given distance matrix; this reconstruction can only be computed efficiently when the distance matrix satisfies the four-point condition that characterizes a tree metric. The four-point condition requires that for any taxa i, j, k , and l the additive distances must satisfy the inequality $d(i, j) + d(k, l) \leq \max\{d(i, k) + d(j, l), d(i, l) + d(j, k)\}$ [6–9].

Neighbor-joining (NJ) is a distance-based method for reconstructing binary phylogenetic trees [10,11]. NJ is a greedy algorithm that combines the two closest clusters or elements into a parent cluster. When a given distance matrix satisfies the four-point condition, the NJ algorithm finds the correct tree for that distance matrix [11,12]. Even when a given distance matrix does not satisfy the four-point condition, NJ is considered to return a good approximate tree [13]. Therefore, the NJ algorithm is used by the scientific community to build phylogenetic trees, mainly due to its low computational cost of $O(N^3)$ and its good accuracy for small to medium sets [14].

In the phylogenetic tree construction process, two or more clusters of elements might be separated by the same distance. In these cases, NJ could generate more than one binary phylogenetic tree, although it only returns one. Therefore, any inference drawn from a single binary phylogenetic tree may be incomplete and, consequently, questionable [15]. Note that the reproducibility of the results is hampered and biased towards one of several possible solutions [16,17]. This algorithmic property is known as the ties in proximity problem [18]. Ties can occur at any step of the clustering process, and they are boosted by the limited resolution (number of decimal places) used in the proximity matrix.

Several variants of the NJ algorithm can be used to avoid the ties in proximity problem. Live neighbor-joining [19] and extended neighbor-joining [20] both allow clusters with up to three elements. The multifurcating neighbor-joining (MFNJ) algorithm [21] consistently produces the same phylogenetic tree regardless of the genotype presentation order in the dataset, and it has the capability to group more than three elements simultaneously.

In this paper, we calculate the percentage of scientific publications employing the NJ algorithm in phylogenetic studies involving microsatellite markers that report only one possible solution, although several could be returned. We believe that the results and conclusions of these papers would have to be carefully analysed since selecting only one possible phylogenetic tree could induce to biased or incorrect conclusions. To properly explain the problem, we have selected a paper as an example where only one phylogenetic tree is reported although other ones could be shown. This example makes it simple to understand that different conclusions could be extracted from the problem at hand.

2. Methods

We conducted the examination of the articles in a three-step process. In the first section, we describe the search strategy implemented to acquire the population dataset of articles utilizing the NJ algorithm for microsatellite marker classification. Afterwards, we explain the procedure for generating the sample dataset. In the last section, we expound upon the analysis of ties within the phylogenetic trees and detail the methodology employed for quantifying the occurrences of non-unique binary phylogenetic trees across articles.

2.1. Search strategy

We analysed the scientific literature on microsatellite markers that used the NJ algorithm. We conducted a search for articles in the Scopus database up to the year 2022. The query used was similar to the one specified in [17]. The search query employed to obtain the articles was: (“microsatellite” OR “microsatellites” OR “simple sequence repeat” OR “simple sequence repeats” OR “SSR” OR “short tandem repeat” OR “short tandem repeats” OR “STR”) AND (“neighbor joining” OR “neighbour joining” OR “neighbor-joining” OR “neighbour-joining”), where AND and OR are the standard boolean operators. We expanded the search scope by including plural forms and multiple word variations. We limited the search to include only the terms from the query found in the title, abstract, or keywords. We gathered bibliometric data, including the document title, journal name, year of publication and subject area. In total, we retrieved 1245 articles ranging from the year 1994 to year 2022 (29 years). After eliminating three duplicated records, the dataset analysed in this study consisted of 1242 articles.

The search strategy utilized in this study might have unintentionally excluded some relevant articles. We acknowledge that this represents a conservative estimate of the overall count of publications featuring a NJ tree of microsatellite markers. This limitation arises from not incorporating articles published in journals beyond the Scopus Indexed Journal List. Moreover, there is a possibility that other articles within the Scopus database utilize an NJ tree with microsatellite markers, but they do not incorporate the specific terms we employed as search criteria within their titles, abstracts, or keywords.

2.2. Sample dataset

Fig. 1 describes the screening process of the articles included in the analysis. We downloaded the selected articles and randomized the dataset to prevent bias towards any of the collected bibliometric data, including specific years, subject areas, or alphabetical order. We worked with an initial subset of 353 articles, comprising roughly 30% of the dataset publications ($n = 1242$). We removed nine unavailable articles, leaving a subset of 344 publications. Within this subset, we specifically selected articles that contained either a phylogenetic tree and a matrix of proximity data (either similarities or distances), or a phylogenetic tree and a table detailing the genetic profiles of all genotypes, provided there were more than three genotypes. In those cases where only the genetic profile table was available, we calculated a matrix indicating the proportion of shared alleles using the *Adegenet* package [22] in R version 4.1.0 [23]. We excluded articles if the proximity data matrix, whether given or calculated, did not match the genotype information in the phylogenetic tree, if the phylogenetic tree was not generated using the NJ algorithm, or if the provided information was incomplete.

Out of 344 articles analysed, only the 29% of them (100 articles) contained the necessary data to be reproducible (see Supplementary Data for the complete list and additional information). Therefore, we had a final sample data set of 100 articles published in over 50 distinct journals. These journals adhere to varying data archiving guidelines, ranging from none to mandatory submission with accompanying documentation. This lack of standardized and mandatory data archiving

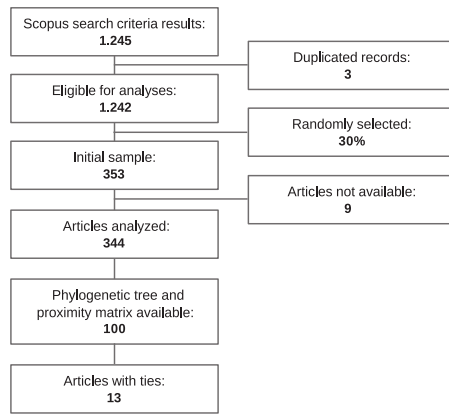


Fig. 1. Flowchart detailing the screening process of the articles included in the dataset.

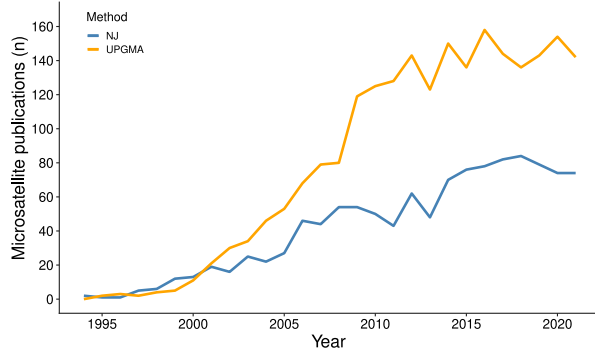


Fig. 2. Comparison of the number of scientific publications in the Scopus database that used the NJ and the UPGMA methods on microsatellite markers from 1994 to 2022. The UPGMA data was updated to match the year range used for NJ, with data originally extracted from [17].

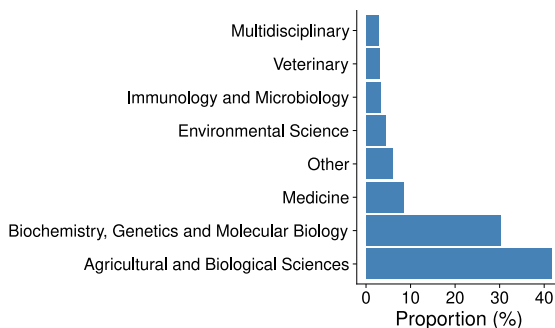


Fig. 3. Articles classified by subject areas. The category named “Other” groups together 14 subject areas that constitute less than 1% each.

policies across journals hinders data accessibility, reproducibility and restricts collaboration opportunities as well as scientific progress [15, 24,25]. The issue of inadequate of data availability is also addressed by [17], who report that only the 26% of the analysed articles featured correct data availability.

2.3. Non-unique binary phylogenetic trees

We used the MFNJ algorithm to analyse and count the presence of ties. As mentioned previously, MFNJ can group any number of elements at a time and thus it always returns a unique phylogenetic tree. We used the R packages *mphylo* to reconstruct the phylogenetic trees [26], and *ggtree* to plot trees [27].

3. Results and discussion

This section is divided as follows. First, we quantify the amount of publications having at least one tie in their clustering process to reconstruct the corresponding phylogenetic tree. Then, we examine the yearly number of articles published containing a NJ tree of microsatellite markers, and we compare it with another agglomerative clustering algorithm such as UPGMA. We continue showing the topics that the articles are classified into. Finally, we show an example of a case study that yields two different binary phylogenetic trees.

3.1. Proportion of articles with ties

In order to count the publications with ties in their phylogenetic trees, we utilized the proximity data matrix extracted from every article in our sample dataset. Subsequently, we calculated their corresponding phylogenetic tree employing the NJ method and used the MFNJ algorithm to assess and quantify the occurrence of ties. We found that there was more than one possible binary phylogenetic tree in 13 out of the 100 articles analysed. That is the 13% of the articles, with a 95% confidence interval (CI) ranging from 6% to 20%. Extending this percentage to the entire population, it yields an estimate of 161 articles (95% CI 74 – 248 articles) that may have alternative solutions in the form of different binary phylogenetic trees. The presence of multiple potential trees in a significant proportion of analysed articles raises important considerations for downstream analyses. In such instances, relying on a single binary phylogenetic tree might be misleading.

3.2. Publications per year

Articles in our population dataset ranged from the year 1994 to the year 2022 (Fig. 2). We compare, under the same search strategy, the number of scientific publications utilizing both the NJ method and the UPGMA method, analysed in [17], because they are both agglomerative clustering methods. Although we observed a greater use of the UPGMA method over the years compared to the NJ method, both clustering methods have shown a steady increase since 2000 and reached a plateau in the number of publications in the last ten years. Approximately 70 articles have been published yearly for the NJ method and 100 articles for the UPGMA method. In the case of the NJ method, 2018 was the year with the highest number of publications ($n = 84$). The increase in publications may be due to two factors: the rise of next-generation sequencing, which brought high-throughput data and lowered sequencing costs, and the improved availability of software for distance-based methods.

3.3. Publications per subject areas

The 1242 articles belong to 21 distinct subject areas. The predominant fields are Agricultural and Biological Sciences, comprising 42%, and Biochemistry, Genetics, and Molecular Biology, constituting 30% (Fig. 3). Together, these two subjects account for 72% of the total article count. It was anticipated that a majority of the articles would pertain to life sciences or analogous subject areas, given the prevalent use of STR and SSR as tools within these fields. Subject areas with minimal representation, each contributing less than 1% to the total article count, have been grouped into a category labelled “Other” in Fig. 3. We can find examples in areas that differ significantly from the previous ones, such as Chemical Engineering [28], Neuroscience [29] and Computer Science [30].

3.4. Case study

We have selected a single case from the 100 articles analysed in our sample dataset to demonstrate how it is possible to generate distinct

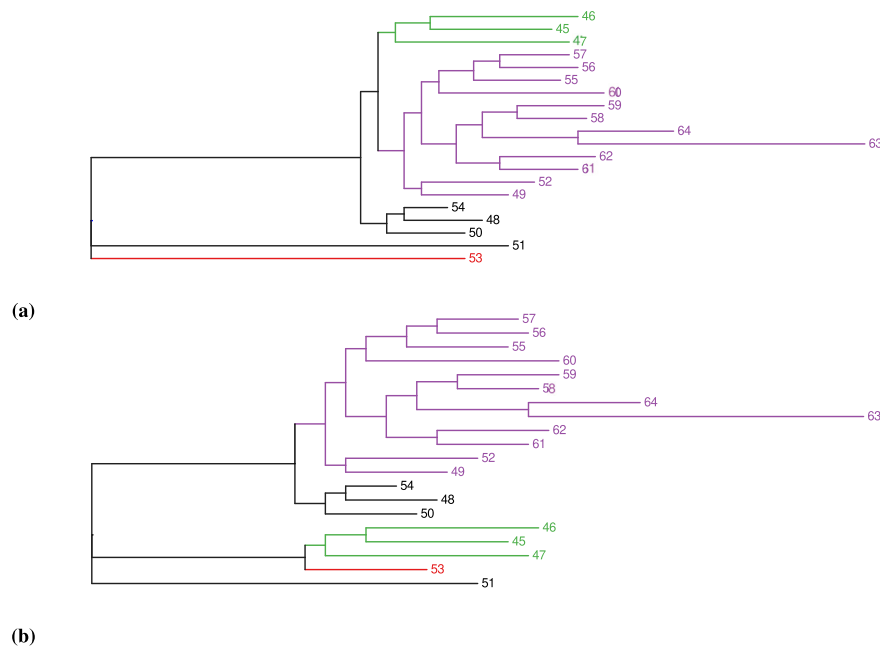


Fig. 4. Phylogenetic trees of the genetic distances between 20 cultivars of upland cotton [31] using the NJ method. Distinct colours are employed to represent tied elements across the phylogenetic trees. Red for cultivar 53, purple for subcluster 1, and green for subcluster 2. Subcluster 1 consists of cultivars 49, 52, 55, 56, 57, 58, 59, 60, 61, 62, 63 and 64 while subcluster 2 includes cultivars 45, 46 and 47. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

binary phylogenetic trees from the same data using the NJ algorithm. This example illustrates a scenario that produces two distinct binary phylogenetic trees.

The authors of this study [31] analysed the genetic diversity and population structure of the cultivated upland cotton (*Gossypium hirsutum* L.) (Fig. 4). The study includes 20 taxa in total, assessing 27 microsatellite markers. The values in the distance matrix range from 0 to 0.73, with a precision of two decimal places. There is a tie between cultivar 53 (depicted in red in Fig. 4) and two clusters. The first cluster, referred to as subcluster 1, is portrayed in purple in Fig. 4, while the second cluster, referred to as subcluster 2, is depicted in green. The cultivar names corresponding to the numbered identifiers are as follows: 45-BRS PEROBA, 46-BRS 7H, 47-ITA90, 48-BRS 8H, 49-BRS ARAÇÁ, 50-BRS PRECOCE, 51-BRS SUCUPIRA, 52-BRS 336, 53-BRS IPÊ, 54-BRS 286, 55-BRS CAMAÇARI, 56-ITA96, 57-BRS 335, 58-BRS ANTARES, 59-BRS 201, 60-BRS FACUAL, 61-BRS PRECOCE, 62-BRS CEDRO, 63-GIBANGA and 64-IMA CD05-8221.

Fig. 4a represents one of the possible binary phylogenetic trees for the same data, where subclusters 1 and 2 are grouped together first. Fig. 4b showcases another binary phylogenetic tree for the same data, wherein subcluster 2 is first grouped with cultivar 53.

A notable difference between the two phylogenetic trees is the addition of cultivar 53. In Fig. 4a, it is clustered at the last step, while in Fig. 4b, it is grouped with subcluster 2 in the middle of the phylogenetic tree.

4. Conclusions

NJ is an algorithm for reconstructing binary phylogenetic trees, which clusters two elements in each step, independently of whether there are more than two elements with the same distance. Thus, it is unable to return multiple binary phylogenetic trees despite the possibility of their existence, as it solely produces a single phylogenetic tree. Results and conclusions in published scientific papers that display a single phylogenetic tree generated by the NJ method could be biased or limited. For this reason, we have quantified how frequently these articles presented only one phylogenetic tree, although more than one was possible.

Our analysis shows that the 13% of the articles (95% CI 6%–20%) possess at least one alternative binary phylogenetic tree to the published one. This corresponds to approximately 161 articles with at least one tie (95% CI 74 – 248 articles) in our dataset of 1242 articles. The potential implications of this result need to be seriously considered, as up to a fifth of the articles considered could be affected by the issue of ties in proximity. The presence of articles featuring NJ binary phylogenetic trees that lack uniqueness may impact not only the direct conclusions drawn from these publications but also have indirect ramifications on subsequent research built upon these original findings.

Usually, researchers frequently rely on the initial or default outcome produced by a clustering method to continue their research. This predisposition derives from the fact that not all researchers are experts in clustering algorithms and, as a result, they may not know of the possibility of multiple phylogenetic trees and do not explore alternative solutions. For this reason, the adaptation of the phylogenetic tree generation process towards a single phylogenetic tree that deals with ties in proximity is crucial in such cases. Consequently, we recommend the utilization of variants of the NJ method, such as the MFNJ method, capable of generating n-ary phylogenetic trees. Complementing the analysis with a bootstrap can further aid in identifying ties, offering a more comprehensive approach to address the challenges associated with clustering methods and the potential existence of multiple phylogenetic trees.

CRediT authorship contribution statement

Natàlia Segura-Alabart: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Francesc Serratosa:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Alberto Fernández:** Writing – review & editing, Supervision, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Universitat Rovira i Virgili, Spain under the Martí Franquès fellowship program [2020PMF-PIPF-45] and the PFR program [2021PFR-URV-100]; Agència de Gestió d'Ajuts Universitaris i de Recerca, Spain [2021SGR-00111]; and Ministerio de Ciencia e Innovación, Spain [PID2021-124139NB-C22, TED2021-129851B-I00].

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.csbr.2024.100022>.

References

- [1] Ismail N, Rafii M, Mahmud T, Hanafi M, Miah G. Molecular markers: a potential resource for ginger genetic diversity studies. *Mol Biol Rep* 2016;43:1347–58. <http://dx.doi.org/10.1007/s11033-016-4070-3>.
- [2] Tautz D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res* 1989;17(16):6463–71. <http://dx.doi.org/10.1093/nar/17.16.6463>.
- [3] Brondani R, Brondani C, Tarchini R, Grattapaglia D. Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. *Theor Appl Genet* 1998;97:816–27. <http://dx.doi.org/10.1007/s001220050961>.
- [4] Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 2004;5(6):435–45. <http://dx.doi.org/10.1038/nrg1348>.
- [5] Vieira M, Santini L, Diniz A, Munhoz C. Microsatellite markers: what they mean and why they are so useful. *Genet Mol Biol* 2016;39(3):312–28. <http://dx.doi.org/10.1590/1678-4685-GMB-2016-0027>.
- [6] Buneman P. A note on the metric properties of trees. *J Combin Theory Ser B* 1974;17(1):48–50. [http://dx.doi.org/10.1016/0095-8956\(74\)90047-1](http://dx.doi.org/10.1016/0095-8956(74)90047-1).
- [7] Steel M. The complexity of reconstructing trees from qualitative characters and subtrees. *J Classification* 1992;9:91–116. <http://dx.doi.org/10.1007/BF02618470>.
- [8] Jiang T, Kearney P, Li M. A polynomial time approximation scheme for inferring evolutionary trees from quartet topologies and its application. *SIAM J Comput* 2001;30(6):1942–61. <http://dx.doi.org/10.1137/S0097539799361683>.
- [9] Cilibrasi R, Vitányi P. Clustering by compression. *IEEE Trans Inform Theory* 2005;51(4):1523–45.
- [10] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4(4):406–25. <http://dx.doi.org/10.1093/oxfordjournals.molbev.a040454>.
- [11] Studier J, Keppler K. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol* 1988;5(6):729–31. <http://dx.doi.org/10.1093/oxfordjournals.molbev.a040527>.
- [12] Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.
- [13] Atteson K. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica* 1999;25:251–78. <http://dx.doi.org/10.1007/PL00008277>.
- [14] Mailund T, Pedersen C. QuickJoin—fast neighbour-joining tree reconstruction. *Bioinformatics* 2004;20(17):3261–2. <http://dx.doi.org/10.1093/bioinformatics/bth359>.
- [15] McTavish E, Drew B, Redelings B, Cranston K. How and why to build a unified tree of life. *BioEssays* 2017;39(11). <http://dx.doi.org/10.1002/bies.201700114>.
- [16] Podani J. On the sensitivity of ordination and classification methods to variation in the input order of data. *J Veg Sci* 1997;8(1):153–6. <http://dx.doi.org/10.2307/3237253>.
- [17] Segura-Alabart N, Serratos F, Gómez S, Fernández A. Nonunique UPGMA clusterings of microsatellite markers. *Brief Bioinform* 2022;23(5):bbac312. <http://dx.doi.org/10.1093/bib/bbac312>.
- [18] Backeljau T, De Bruyn L, De Wolf H, Jordaens K, Van Dongen S, Winnepeninckx B. Multiple UPGMA and neighbor-joining trees and the performance of some computer packages. *Mol Biol Evol* 1996;13(2):309–13. <http://dx.doi.org/10.1093/oxfordjournals.molbev.a025590>.
- [19] Telles G, Araújo G, Walter M, Brígido M, Almeida N. Live neighbor-joining. *BMC Bioinformatics* 2018;19:172. <http://dx.doi.org/10.1186/s12859-018-2162-x>.
- [20] Hong Y, Guo M, Wang J. ENJ algorithm can construct triple phylogenetic trees. *Mol Ther Nucleic Acids* 2021;23(5):286–93. <http://dx.doi.org/10.1016/j.omtn.2020.11.004>.
- [21] Fernández A, Segura-Alabart N, Serratos F. The multifurcating neighbor-joining algorithm for reconstructing polytomic phylogenetic trees. *J Mol Evol* 2023;91:773–9. <http://dx.doi.org/10.1007/s00239-023-10134-z>.
- [22] Jombart T, Ahmed I. ADEGENET 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 2011;27(21):3070–1. <http://dx.doi.org/10.1093/bioinformatics/btr521>.
- [23] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2023, R version 4.2.0. URL <https://www.R-project.org/>.
- [24] Vines T, Andrew R, Bock D, Franklin M, Gilbert K, Kane N, et al. Mandated data archiving greatly improves access to research data. *FASEB J* 2013;27(4):1304–8. <http://dx.doi.org/10.1096/fj.12-218164>.
- [25] Roche D, Lanfear R, Binning S, Haff T, Schwanz L, Cain K, et al. Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol* 2014;12(1):e1001779. <http://dx.doi.org/10.1371/journal.pbio.1001779>.
- [26] Fernández A. mphylo: multifurcated phylogenetic trees in R. 2023, URL <https://github.com/albyfs/mphylo>. [Accessed 23 December 2023].
- [27] Yu G. Using ggtree to visualize data on tree-like structures. *Curr Protoc Bioinform* 2020;69:e96. <http://dx.doi.org/10.1002/cpbi.96>.
- [28] Ditta A, Zhou Z, Cai X, Wang X, Okubazghi K, Shehzad M, et al. Assessment of genetic diversity, population structure, and evolutionary relationship of uncharacterized genes in a novel germplasm collection of diploid and allotetraploid *Gossypium* accessions using EST and genomic SSR markers. *Int J Mol Sci* 2018;19(8):2401. <http://dx.doi.org/10.3390/ijms19082401>.
- [29] Li Z, Gichira A, Wang Q, Chen J. Genetic diversity and population structure of the endangered basal angiosperm *Brasenia schreberi* (Cabombaceae) in China. *PeerJ* 2018;6:e5296. <http://dx.doi.org/10.7717/peerj.5296>.
- [30] Chapal-Ilani N, Maruvka Y, Spiro A, Reizel Y, Adar R, Shlush L, et al. Comparing algorithms that reconstruct cell lineage trees utilizing information on microsatellite mutations. *PLoS Comput Biol* 2013;9(11):1–17. <http://dx.doi.org/10.1371/journal.pcbi.1003297>.
- [31] Moiana LD, Filho PSV, Gonçalves-Vidigal MC, de Carvalho LP. Genetic diversity and population structure of upland cotton Brazilian cultivars (*Gossypium hirsutum* L. race *latifolium* H.) using SSR markers. *Aust J Crop Sci* 2015;9(2):143–52.