Bull Math Biol
https://doi.org/10.1007/s11538-018-0444-0

SPECIAL ISSUE: ALGEBRAIC METHODS IN PHYLOGENETICS

Mathematical and Simulation-Based Analysis of the Behavior of Admixed Taxa in the Neighbor-Joining Algorithm

Jaehee Kim[1] · **Filippo Disanto**[2] · **Naama M. Kopelman**[3] · **Noah A. Rosenberg**[1]

Received: 27 October 2017 / Accepted: 10 May 2018

Abstract The neighbor-joining algorithm for phylogenetic inference (NJ) has been seen to have three specific properties when applied to distance matrices that contain an admixed taxon: (1) antecedence of clustering, in which the admixed taxon agglomerates with one of its source taxa before the two source taxa agglomerate with each other; (2) intermediacy of distances, in which the distance on an inferred NJ tree between an admixed taxon and either of its source taxa is smaller than the distance between the two source taxa; and (3) intermediacy of path lengths, in which the number of edges separating the admixed taxon and either of its source taxa is less than or equal to the number of edges between the source taxa. We examine the behavior of neighbor-joining on distance matrices containing an admixed group, investigating the occurrence of antecedence of clustering, intermediacy of distances, and intermediacy of path lengths. We first mathematically predict the frequency with which the properties are satisfied for a labeled unrooted binary tree selected uniformly at random in the absence of admixture. We then introduce a taxon constructed by a linear admixture of distances from two source taxa, examining three admixture scenarios by simulation: a model in which distance matrices are chosen at random, a model in which an admixed taxon is added to a set of taxa that reflect treelike evolution, and a model that introduces a perturbation of the treelike scenario. In contrast to previous

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11538-018-0444-0) contains supplementary material, which is available to authorized users.

✉ Jaehee Kim
jk2236@stanford.edu

1 Department of Biology, Stanford University, Stanford, CA 94305, USA

2 Department of Mathematics, University of Pisa, 56126 Pisa, Italy

3 Department of Computer Science, Faculty of Sciences, Holon Institute of Technology, 58109 Holon, Israel

Published online: 06 June 2018

conjectures, we observe that the three properties are sometimes violated by distance matrices that include an admixed taxon. However, we also find that they are satisfied more often than is expected by chance when the distance matrix contains an admixed taxon, especially when evolution among the non-admixed taxa is treelike. The results contribute to a deeper understanding of the nature of evolutionary trees constructed from data that do not necessarily reflect a treelike evolutionary process.

Keywords Admixture · Neighbor-joining · Phylogenetics · Rogue taxon

1 Introduction

The neighbor-joining (NJ) algorithm (Saitou and Nei 1987; Studier and Keppler 1988; Gascuel and Steel 2006) is a distance-based clustering method widely used in phylogenetic analysis. Given a pairwise distance matrix D, NJ forms a bifurcating tree by iteratively selecting and merging pairs of nodes in an order governed by a specified criterion, until the tree is fully resolved. When the input D is additive in the sense that its distances represent distances between taxon pairs in a generating tree whose branches are each assigned distance values (Buneman 1974; Steel 2016), NJ recovers the generating tree precisely (Atteson 1999; Bryant 2005; Steel 2016).

Distance matrices computed from data often fail to satisfy the additivity condition (Felsenstein 1984), however, particularly when populations evolve in a non-treelike manner—for example, by processes of hybridization or admixture in which certain groups are descended from pairs of source groups that have long been separated. In population-genetic studies that have used NJ to build trees that represent genetic relationships among populations, admixed taxa have been observed to behave in specific ways during the application of the NJ algorithm (Bowcock et al. 1991; Cavalli-Sforza et al. 1994; Mountain and Cavalli-Sforza 1994; Ruiz-Linares et al. 1995; Kopelman et al. 2013). In particular, NJ analyses of distance matrices in human populations tend to place admixed populations on relatively short external branches that lie in an intermediate position in the inferred tree in relation to source populations for the admixture. For example, using mean genetic distances calculated across multiple loci genome-wide, Kopelman et al. (2013) illustrated this phenomenon in mestizo populations descended from European and Native American populations in Latin America, finding that they appeared as pendant edges along the "spine" of an NJ tree connecting the European and Native American populations.

Kopelman et al. (2013) formalized the definitions of three properties often seen in trees constructed by NJ in the presence of admixture. First, during the agglomerative construction of the inferred tree, taxa admixed between two parental sources are often seen to agglomerate with a group containing one of those sources before those sources agglomerate with each other (*antecedence of clustering*, Fig. 1). Second, the distance of an admixed taxon on an inferred NJ tree to the more distant of its source populations is often seen to be smaller than the distance between the source populations themselves (*intermediacy of distances*, Fig. 2). Third, the number of edges separating an admixed taxon from the more distant of its source populations in terms of path length on the

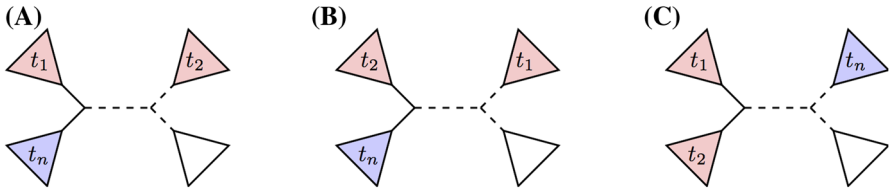

Fig. 1 Antecedence of clustering (Property 1). Taxa t_1 and t_2 are source taxa, and t_n is a taxon admixed with sources t_1 and t_2. In the example trees shown, triangles represent subtrees. Clades containing the source taxa are colored in red, and a clade containing the admixed taxon t_n appears in blue. The unlabeled triangle contains one or more taxa but does not contain t_1, t_2, or t_n. If we denote a clade containing taxon t_i by C_{t_i}, three choices exist for the pair of clades that agglomerate first in neighbor-joining, among (C_{t_1}, C_{t_2}), (C_{t_1}, C_{t_n}), (C_{t_2}, C_{t_n}); this pair is connected by solid rather than dashed edges. **a** (C_{t_1}, C_{t_n}). **b** (C_{t_2}, C_{t_n}). **c** (C_{t_1}, C_{t_2}). Cases (**a**) and (**b**) satisfy Property 1, *antecedence of clustering*. Case (**c**), in which the source taxa t_1 and t_2 agglomerate before either agglomerates with t_n, violates Property 1

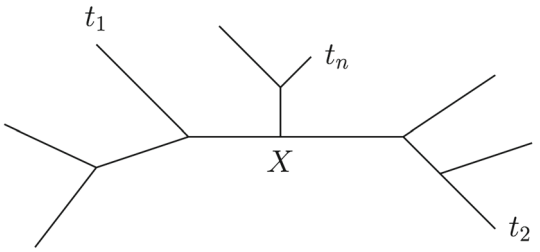

Fig. 2 Example tree in which t_1 and t_2 are regarded as source taxa and t_n is treated as an admixed taxon. X is the unique node that lies at the intersection of paths $t_1 - t_2$, $t_1 - t_n$, and $t_2 - t_n$. The tree shown satisfies Property 2, *intermediacy of distances*, as $d_{t_1,t_n} < d_{t_1,t_2}$ and $d_{t_2,t_n} < d_{t_1,t_2}$. Because it has $b_{t_1,t_n} = 4$, $b_{t_2,t_n} = 5$, and $b_{t_1,t_2} = 5$, it also satisfies Property 3, *intermediacy of path lengths*: $b_{t_1,t_n} \leq b_{t_1,t_2}$ and $b_{t_2,t_n} \leq b_{t_1,t_2}$

inferred NJ tree is often seen to be smaller than the number of edges separating the source populations from each other (*intermediacy of path lengths*, Fig. 2).

Using computations that mechanistically considered the application of NJ to distance matrices, Kopelman et al. (2013) studied the characteristics of NJ trees that incorporate an admixed taxon whose distances are formed from the linear mixture of distances involving two source taxa. They investigated two special cases of distance matrices: an additive distance matrix with an arbitrary number of taxa and a general 4-taxon distance matrix. For the former case, they demonstrated that intermediacy of distances and intermediacy of path lengths are necessarily observed. For the latter, they showed that antecedence of clustering also holds in addition to the other two properties. They conjectured that the three properties are general features of the application of NJ on admixed distance matrices.

In this paper, we extend the mechanistic computation of Kopelman et al. and examine their conjectures in more general settings. We find that, contrary to the conjectures, the properties do not always hold for nonadditive distance matrices with $n \geq 5$ taxa. On the basis of simulations, however, we also find that the properties hold much more often than is expected by chance under a null model in which admixed taxa are absent. We explore the dependence on model parameters of the extent to which the properties

hold. The results provide further information useful for interpreting the outcome of NJ in circumstances that contain admixed populations.

2 Null Model Without Admixture

Before evaluating by simulation the extent to which the properties of Kopelman et al. hold in cases with admixture, we first predict their frequencies using a null model without admixture. Consider a tree with n taxa, labeled $t_1, t_2, \ldots, t_n$. We choose a taxon t_n to play the role of an admixed taxon, with source populations t_1 and t_2. Let d_{ij} be the distance between leaves t_i and t_j of the tree for some distance measure d, and let b_{ij} be the number of edges of the tree that lie on the path from t_i to t_j.

We formally state the properties of Kopelman et al. involving an admixed taxon t_n and source taxa t_1 and t_2: (1) For antecedence of clustering, the admixed taxon must cluster with one of the sources before the sources cluster with each other; (2) for intermediacy of distances, $d_{t_1,t_n} < d_{t_1,t_2}$ and $d_{t_2,t_n} < d_{t_1,t_2}$ must hold; and (3) for intermediacy of path lengths, $b_{t_1,t_n} \leq b_{t_1,t_2}$ and $b_{t_2,t_n} \leq b_{t_1,t_2}$ must hold.

At each step of NJ agglomeration of taxa, the algorithm evaluates a quantity q_{ij} for each pair (i, j) of taxa, or previously agglomerated clusters of taxa, represented as internal nodes of the constructed tree. The pair with the minimal value of the quantity is then merged ("Appendix A"). When two or more pairs of clusters have the same minimal q_{ij}, the tie is broken at random. When the tie involves clades containing source taxa and an admixed taxon and could thus affect whether Property 1 is attained, however, we adopt the convention of Kopelman et al. and agglomerate a pair of clades satisfying Property 1.

Note that we are concerned here with the way in which NJ acts on distance matrices in general, and we do not concern ourselves with the way in which the distance measure d is computed from data. Although the examples that motivate the problem arise from consideration of trees of populations, with distances between populations computed as averages across many loci in the genome, the taxa in the distance matrix can represent any taxonomic entities for which it is of interest to compute a tree.

2.1 Exact Calculation

We now predict the probability that a random tree—with three distinct taxa chosen at random to fulfill the roles of taxa t_1, t_2, and t_n—satisfies (or violates) the three properties. Note that although t_n is in the role of an admixed taxon and t_1 and t_2 are in the roles of source taxa, no admixture occurs. We consider random trees to be selected from the uniform distribution on labeled unrooted topologies with n taxa. There exist $(2n-5)!! = (2n-5)(2n-3) \times \cdots \times 5 \times 3 \times 1$ such topologies (Felsenstein 2004), each with $2n-3$ edges.

2.1.1 Property 1: Antecedence of Clustering

The three taxa of interest are t_1, t_2, and t_n. Three choices exist for the agglomeration that occurs first: $((C_{t_1}, C_{t_2}), C_{t_n})$, $((C_{t_1}, C_{t_n}), C_{t_2})$, and $((C_{t_2}, C_{t_n}), C_{t_1})$, where C_{t_i}

represents a clade containing taxon t_i. The first of these cases, in which taxa t_1 and t_2 are agglomerated first, produces a violation of Property 1. We would, therefore, predict a probability of violation of antecedence of clustering equal to 1/3.

2.1.2 Property 2: Intermediacy of Distances

The three distances on the tree are $d_{t_1t_2}$, $d_{t_1t_n}$, and $d_{t_2t_n}$. By symmetry, with random choices of the three values, all six possible arrangements of the values from smallest to largest are equiprobable. For Property 2 to be satisfied, $d_{t_1t_2} > d_{t_1t_n}$ and $d_{t_1t_2} > d_{t_2t_n}$ are required. Assuming that d has probability 0 that two distances are equal, each of the six orderings has probability 1/6. Four of the six, $d_{t_1,t_2} < d_{t_1,t_n} < d_{t_2,t_n}$, $d_{t_1,t_2} < d_{t_2,t_n} < d_{t_1,t_n}$, $d_{t_1,t_n} < d_{t_1,t_2} < d_{t_2,t_n}$, and $d_{t_2,t_n} < d_{t_1,t_2} < d_{t_1,t_n}$, fail to satisfy one or both of the conditions. Thus, the predicted probability for Property 2 to be false is 2/3.

2.1.3 Property 3: Intermediacy of Path Lengths

For Property 3 to hold, the inequalities $b_{t_1t_2} \geq b_{t_1,t_n}$ and $b_{t_1,t_2} \geq b_{t_2,t_n}$ must simultaneously hold. We compute the probability that they hold for randomly chosen unrooted labeled topologies. As the output of NJ is an unrooted bifurcating tree, we compute the probability that a random unrooted labeled bifurcating tree, with three fixed taxa selected at random, labeled t_1, t_2, and t_n, satisfies the pair of inequalities.

Let X be the unique node that is an intersection of the paths $t_1 - t_2$, $t_1 - t_n$, and $t_2 - t_n$ (Fig. 2). X is the unique node assigning t_1, t_2, and t_n to different subtrees. Then

$$b_{t_1t_2} \geq b_{t_1t_n} \iff b_{t_1X} + b_{t_2X} \geq b_{t_1X} + b_{t_nX} \iff b_{t_2X} \geq b_{t_nX}$$
$$b_{t_1t_2} \geq b_{t_2t_n} \iff b_{t_1X} + b_{t_2X} \geq b_{t_2X} + b_{t_nX} \iff b_{t_1X} \geq b_{t_nX}.$$

We can assign a 4-vector $(n, b_{t_1X}, b_{t_2X}, b_{t_nX})$ to each of the $(2n-5)!!$ trees in the set of unrooted binary trees with n leaves. Each of the entries, b_{t_1X}, b_{t_2X}, and b_{t_nX}, ranges from 1 to $n-2$, the maximal distance possible from a leaf to an internal node of an unrooted bifurcating tree with n taxa. Note that multiple trees can share the same vector of four variables.

A taxon can be added to a tree characterized by $(n, b_{t_1X}, b_{t_2X}, b_{t_nX})$ in one of four ways:

(i) The new taxon, taxon $n+1$, is added to an edge in path $t_1 - X$. The new tree then has the 4-vector $(n+1, b_{t_1X}+1, b_{t_2X}, b_{t_nX})$. There are b_{t_1X} ways to place taxon $n+1$ on the path $t_1 - X$, each corresponding to a distinct edge that is subdivided by the placement of the new taxon.

(ii) The new taxon is added to an edge in path $t_2 - X$. The new tree has the 4-vector $(n+1, b_{t_1X}, b_{t_2X}+1, b_{t_nX})$. There are b_{t_2X} ways to place taxon $n+1$ on the path $t_2 - X$.

(iii) The new taxon is added to an edge in path $t_n - X$. The new tree has the 4-vector $(n+1, b_{t_1X}, b_{t_2X}, b_{t_nX}+1)$. There are b_{t_nX} ways to place taxon $n+1$ on the path $t_n - X$.

(iv) The new taxon is added to an edge that is not in any of paths $t_1 - X$, $t_2 - X$ or $t_n - X$. The new tree has the 4-vector $(n+1, b_{t_1X}, b_{t_2X}, b_{t_nX})$. Because the tree has $2n-3$ edges in total, the number of ways to place taxon $n+1$ is $(2n-3) - (b_{t_1X} + b_{t_2X} + b_{t_nX})$.

Let $a_{n,i,j,k}$ be the number of trees having the same 4-vector (n, i, j, k). Because the total number of unrooted labeled binary trees with n taxa is $(2n-5)!!$,

$$\sum_{i=1}^{n-2}\sum_{j=1}^{n-2}\sum_{k=1}^{n-2} a_{n,i,j,k} = (2n-5)!!.$$

The probability for a random tree with n taxa to have 4-vector (n, i, j, k) is then $P_{n,i,j,k} \equiv a_{n,i,j,k}/(2n-5)!!$. By considering the ways to add taxon $n+1$ to an n-taxon tree described in (i)-(iv), we can produce a recursion:

$$\begin{aligned} a_{n+1,i,j,k} = {} & (i-1)a_{n,i-1,j,k} + (j-1)a_{n,i,j-1,k} + (k-1)a_{n,i,j,k-1} \\ & + (2n-3-i-j-k)a_{n,i,j,k}. \end{aligned} \tag{1}$$

The unique unrooted labeled bifurcating tree for $n = 3$ gives the initial condition for the recursion, or $a_{3,1,1,1} = 1$, with $a_{3,i,j,k} = 0$ for all (i, j, k) with $(i, j, k) \neq (1, 1, 1)$.

Dividing Eq. 1 by the total number of trees with $n+1$ taxa, $[2(n+1)-5]!!$, we have

$$\begin{aligned} \frac{1}{(2n-3)!!}a_{n+1,i,j,k} = {} & \frac{1}{2n-3}\left[\frac{i-1}{(2n-5)!!}a_{n,i-1,j,k} + \frac{j-1}{(2n-5)!!}a_{n,i,j-1,k}\right. \\ & \left. + \frac{k-1}{(2n-5)!!}a_{n,i,j,k-1} + \frac{2n-3-i-j-k}{(2n-5)!!}a_{n,i,j,k}\right], \end{aligned}$$

and

$$\begin{aligned} P_{n+1,i,j,k} = {} & \left(\frac{i-1}{2n-3}\right)P_{n,i-1,j,k} + \left(\frac{j-1}{2n-3}\right)P_{n,i,j-1,k} + \left(\frac{k-1}{2n-3}\right)P_{n,i,j,k-1} \\ & + \left(1 - \frac{i+j+k}{2n-3}\right)P_{n,i,j,k}, \end{aligned} \tag{2}$$

with initial conditions $P_{3,1,1,1} = 1$ and $P_{3,i,j,k} = 0$ for all (i, j, k) with $(i, j, k) \neq (1, 1, 1)$.

For a tree having a 4-vector (n, i, j, k) to satisfy Property 3, we must have $b_{t_2X} \geq b_{t_nX}$ and $b_{t_1X} \geq b_{t_nX}$. This condition is equivalent to $j \geq k$ and $i \geq k$. To evaluate the probability that Property 3 holds, we must sum over all possible $\{i, j, k\}$ that also satisfy $j \geq k$ and $i \geq k$. This sum is

$$P(\text{Property 3 holds}) = \sum_{k=1}^{n-2}\sum_{j=k}^{n-2}\sum_{i=k}^{n-2} P_{n,i,j,k}. \tag{3}$$

Table 1 The exact probability that Property 3 fails to hold in a random unrooted labeled binary tree with n taxa

n	Number of Trees Satisfying Property 3	Number of Labeled Unrooted Trees	P(Property 3 Fails)
3	1	1	0
4	2	3	0.333333
5	8	15	0.466667
6	54	105	0.485714
7	468	945	0.504762
8	4950	10,395	0.523810
9	62,640	135,135	0.536464
10	920,430	2,027,025	0.545921
11	15,373,260	34,459,425	0.553874
12	287,746,830	654,729,075	0.560510
13	5,965,860,600	13,749,310,575	0.566097
14	135,691,860,150	316,234,143,225	0.570913
15	3,359,026,786,500	7,905,853,580,625	0.575122
16	89,901,262,801,350	213,458,046,676,875	0.578834

The number of unrooted labeled binary trees is $(2n-5)!!$. The probabilities are computed using Eq. 3

The exact probabilities that Property 3 holds for small values of n can be obtained from Eq. 3 by recursively evaluating Eq. 2. The probability it is violated is one minus the expression in Eq. 3 (Table 1). We can observe from Table 1 that the probability of a violation increases from 0 for $n = 3$ to 0.579 for $n = 16$.

2.2 Simulation

Having examined the exact probabilities for the three properties to fail in a random unrooted binary tree, we now evaluate by simulation the frequency of occurrence of these properties in trees obtained by application of NJ to random distance matrices without admixture.

For n taxa, we represent a symmetric $n \times n$ distance matrix by $\mathbf{D}^{(n)} = [x_{ij}]$, where x_{ij} is the pairwise distance between taxa i and j in a set of n taxa $\{1, 2, \ldots, n\}$. Because the number of independent entries in $\mathbf{D}^{(n)}$ is $\binom{n}{2}$, we can represent $\mathbf{D}^{(n)}$ as an $\binom{n}{2}$-dimensional vector that gives the lower triangle of $\mathbf{D}^{(n)}$:

$$\mathbf{x}^{(n)} = (x_{21}, x_{31}, x_{32}, \ldots, x_{n1}, x_{n2}, \ldots, x_{n,n-1})^{\mathsf{T}} \in \mathbb{R}^{\binom{n}{2}}, \tag{4}$$

ordering the vector by the first index i and then by the second index j, with $i > j$.

Our null model considers n species with no admixture and an $n \times n$ distance matrix $\mathbf{D}^{(n)}$. The entries in $\mathbf{x}^{(n)}$, the lower triangle of the matrix, are chosen independently at random, each from a uniform-[0, 1] distribution. The distance matrix is symmetric,

Table 2 Percentage of violations of the three properties among 99,000 inferred NJ trees using random input distance matrices without admixture

Distances are sampled independently as uniform-[0, 1] random variables. In each matrix, the roles of source taxa are played by t_1 and t_2 and the admixed taxon is t_n

n	Property violated						
	1	2	3	1 & 2	1 & 3	2 & 3	1 & 2 & 3
4	33.3	66.9	33.3	29.9	33.3	29.9	29.9
5	33.1	66.7	46.5	28.4	33.1	40.1	28.4
6	33.2	66.5	44.1	28.0	29.7	38.3	26.3
7	33.2	66.4	44.9	28.1	29.2	39.3	26.2
8	33.2	66.6	46.9	28.1	29.6	40.9	26.4
9	33.1	66.7	48.6	27.9	29.6	42.4	26.5
10	33.6	66.8	49.6	28.3	30.0	43.1	26.8
11	33.3	66.5	50.1	28.2	29.8	43.4	26.7
12	33.5	66.6	50.2	28.3	29.8	43.5	26.8
13	33.2	66.8	50.3	28.1	29.4	43.7	26.5
14	33.3	66.7	50.8	28.3	29.6	44.0	26.7
15	33.5	66.9	51.3	28.6	29.8	44.3	26.9
16	33.4	66.7	51.4	28.5	29.9	44.4	27.0

and we fill in the upper triangle by symmetry, assigning values of 0 along the diagonal. All off-diagonal entries are assumed to be strictly positive.

For each number of taxa n, we generated 99,000 $n \times n$ random symmetric matrices $\mathbf{D}^{(n)}$. This total number of matrices was chosen to match the number that will be used in simulation models with admixture in Sect. 3, where 1000 copies of random admixed distance matrices will be generated for each of the 99 admixture values in $\{0.01, 0.02, \ldots, 0.99\}$.

For each distance vector $\mathbf{x}^{(n)}$, NJ outputs a specific tree topology. Because $(2n-5)!!$ labeled unrooted binary tree topologies exist for a given n, the space of distance matrices $[0, 1]^{\binom{n}{2}}$ can be divided into $(2n-5)!!$ subsets $D^{(n)}_{\mathcal{T}_i}$ with $i = 1, \ldots, (2n-5)!!$ and $\bigcup_{i=1}^{(2n-5)!!} D^{(n)}_{\mathcal{T}_i} = [0, 1]^{\binom{n}{2}}$, each corresponding to an inferred NJ tree topology $\mathcal{T}_i$ (Davidson et al. 2017). Thus, NJ was applied to each simulated $\mathbf{x}^{(n)}$ to identify its associated topology.

The numbers of violations of Properties 1, 2, and 3 and the numbers of simultaneous violations of two or all three properties among 99,000 matrices for $n = 4, 5, \ldots, 16$ are shown in Table 2.

2.2.1 Property 1

In the exact calculation in Sect. 2.1, the expected number of violations for Property 1 is $99{,}000 \times 1/3 = 33{,}000$, independent of the number of taxa. The simulation depicted in Table 2 shows that the number of violations from the null model simulation is near 1/3 of the total number of simulations for all choices of n, in close agreement with the expected number of Property 1 violations.

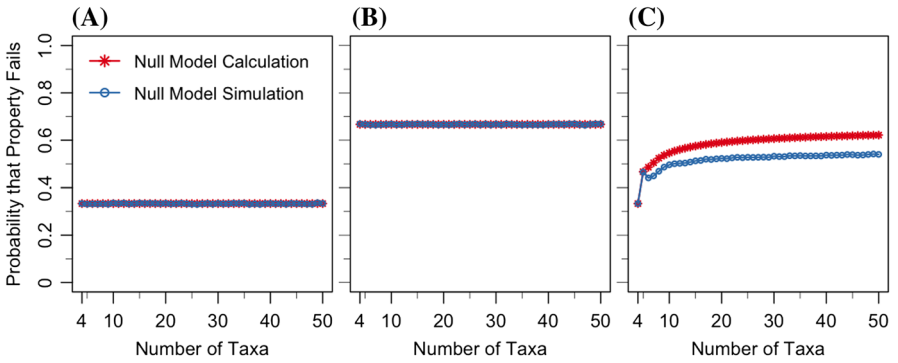

Fig. 3 Simulated probabilities that Properties 1, 2, and 3 are violated under a null model in which distance matrices are generated randomly without admixture. **a** Property 1. **b** Property 2. **c** Property 3. The exact probability for a random labeled unrooted binary tree to violate the properties and the simulated probability for a neighbor-joining tree inferred from a random input distance matrix to violate the properties are both shown. For the calculation for Property 3 for $n = 4$ to $n = 16$, the values plotted are taken from Table 1. Simulation values are based on 99,000 matrices for each n; the values for $n = 4$ to $n = 16$ appear in Table 2

2.2.2 Property 2

In the null model, the exact probability that Property 2 fails is 2/3, independent of the number of taxa n. Based on the exact calculation in Sect. 2.1, the expected number of violations of Property 2 is $99{,}000 \times 2/3 = 66{,}000$. Our null model simulation result in Table 2 agrees well with the predicted number from the exact calculation. It shows no systematic dependence of the number of Property 2 violations on n.

2.2.3 Property 3

As Table 1 shows, the null probability of violations for a random tree with a random assignment of source and admixed taxa is 1/3 when $n = 4$, and it increases with n. In agreement with the exact calculation, the number of violations in the NJ simulation under the null model is near 1/3 of the total number of simulations, and it also increases with n (Table 2). As shown in Fig. 3, however, the simulated results differ slightly from the values from the recursion relation in Eq. 3, except when $n = 4$ and 5.

The discrepancy between the simulated and exact values arises because for $n \geq 6$, the sampling of the distance space used in the simulation is not uniform across the discrete tree space used in the exact calculation. In our exact calculation in Sect. 2.1, a random tree topology is considered from a uniform distribution of $(2n - 5)!!$ labeled unrooted binary tree topologies. In our null model simulation, the input distance matrices are instead sampled from a uniform distribution on the distance matrix space $[0, 1]^{\binom{n}{2}}$.

This uniform sampling from the distance matrix space does not translate into uniform sampling from the tree space. The probability that a randomly chosen distance matrix results in a given tree topology $\mathcal{T}_i$ is proportional to the volume of the sub-

set $D_{\mathcal{T}_i}^{(n)}$ of $[0,1]^{\binom{n}{2}}$, and different tree topologies $\mathcal{T}_i$ can produce $D_{\mathcal{T}_i}^{(n)}$ with different volume (Eickmeyer et al. 2008; Eickmeyer and Yoshida 2008).

The cases of $n = 4$ and 5 have only one unlabeled unrooted tree topology, and thus, each labeled unrooted tree is equally likely to be produced: all $D_{\mathcal{T}_i}^{(n)}$ have the same volume in $[0,1]^{\binom{n}{2}}$. In these cases, the uniform random sampling of a distance matrix is equivalent to the uniform random sampling of a labeled unrooted tree topology. For $n = 4$, of the three possible labelings, one violates Property 3, so that the expected number of violations for $n = 4$ is $99{,}000 \times 1/3 = 33{,}000$. Of the 15 labelings with $n = 5$, 7 violate Property 3, and the expected number of Property 3 violations for $n = 5$ is $99{,}000 \times 7/15 = 46{,}200$. Our simulated results for $n = 4$ and 5 accord with the expected values from the exact calculation (Table 2).

For $n \geq 6$, more than one unlabeled unrooted tree topology exists, and the difference in volume in $[0,1]^{\binom{n}{2}}$ for distinct $\mathcal{T}_i$ results in nonuniform sampling of the labeled unrooted tree topology. This difference accounts for the difference between the probability from an NJ simulation using random matrices under the null model and the exact probability from random trees.

3 Model with Admixture

The simulation without admixture in Sect. 2.2 closely accords with the exact computation under the null model in Sect. 2.1, which did not include the effect of admixture. To the extent that the simulations and exact computations differ, we can attribute the difference to differences in the assumptions of the simulation and the calculation. Having explored the null model with no admixture, we now examine the three properties on distance matrices with admixture.

As before, let t_1 and t_2 be the source taxa, and let t_n be the taxon admixed with sources t_1 and t_2. The remaining $n - 3$ taxa $(t_3, t_4, \ldots, t_{n-1})$ have no systematic relationships with one another. To model admixture in distance matrices, we adopt the linear mixture model of Kopelman et al. (2013), in which the distance between t_n and any t_i, $i = 1, 2, \ldots, n-1$, is

$$x_{ni} = \alpha x_{1i} + (1-\alpha) x_{2i}. \tag{5}$$

Here, α is an admixture fraction for population 1, representing the contribution of population 1 to the admixed population. Using Eq. 5, the distance matrix including the admixed population is

$$\mathbf{D}^{(n)} = \begin{bmatrix} 0 & x_{21} & \cdots & x_{(n-1)1} & (1-\alpha)x_{21} \\ x_{21} & 0 & \cdots & x_{(n-1)2} & \alpha x_{21} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{(n-1)1} & x_{(n-1)2} & \cdots & 0 & \alpha x_{(n-1)1} + (1-\alpha)x_{(n-1)2} \\ (1-\alpha)x_{21} & \alpha x_{21} & \cdots & \alpha x_{(n-1)1} + (1-\alpha)x_{(n-1)2} & 0 \end{bmatrix}. \tag{6}$$

We denote the $(n-1)\times(n-1)$ source distance matrix from which the $n\times n$ admixed distance matrix is created by $\mathbf{S}^{(n-1)}$, with $S_{ij}^{(n-1)} = D_{ij}^{(n)}$ for $i, j = 1, 2, \ldots, n-1$. We assume all off-diagonal entries are strictly positive and that the matrix $\mathbf{S}^{(n-1)}$ is symmetric.

Kopelman et al. showed that, for arbitrarily many taxa, Properties 2 and 3 hold when the admixed distance matrix $\mathbf{D}^{(n)}$ in Eq. 6 is additive. A distance matrix with its last row and column generated by Eq. 5, however, is not additive in general, and in fact, it is "rarely" additive. To illustrate what this statement means, for the $n = 4$ case, where the source taxa are labeled 1 and 2, and the admixed taxon is taxon 4, consider the distance matrix $\mathbf{D}^{(4)}$:

$$\mathbf{D}^{(4)} = \begin{bmatrix} 0 & x_{12} & x_{13} & x_{14} \\ x_{21} & 0 & x_{23} & x_{24} \\ x_{31} & x_{32} & 0 & x_{34} \\ x_{41} & x_{42} & x_{43} & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & x_{21} & x_{31} & (1-\alpha)x_{21} \\ x_{21} & 0 & x_{32} & \alpha x_{21} \\ x_{31} & x_{32} & 0 & \alpha x_{31} + (1-\alpha)x_{32} \\ (1-\alpha)x_{21} & \alpha x_{21} & \alpha x_{31} + (1-\alpha)x_{32} & 0 \end{bmatrix},$$

with $(x_{21}, x_{31}, x_{32}) \in \mathbb{R}_+^3$.

The necessary and sufficient condition on (x_{21}, x_{31}, x_{32}) for $\mathbf{D}^{(4)}$ to be additive—or, in other words, to represent a tree metric—is that a four-point condition (Buneman 1974; Steel 2016) is satisfied for all taxon quartets (i, j, k, ℓ):

$$x_{ij} + x_{k\ell} \leq \max\{x_{ik} + x_{j\ell}, x_{i\ell} + x_{jk}\}. \tag{7}$$

It suffices to consider this condition for $(i, j, k, \ell) = (1, 2, 3, 4)$, $(1, 3, 2, 4)$, and $(1, 4, 2, 3)$, as other cases are redundant. Aggregating inequalities from each of these assignments of taxa 1, 2, 3, and 4 to (i, j, k, ℓ), for each assignment, the largest two of $x_{21} + x_{43}$, $x_{31} + x_{42}$, and $x_{41} + x_{32}$ among the three values must be equal.

First, consider the case of $x_{21} + x_{43} = x_{31} + x_{42}$. Because $x_{43} = \alpha x_{31} + (1-\alpha)x_{32}$ and $x_{42} = \alpha x_{21}$, we can rewrite the equality as $x_{21} + \alpha x_{31} + (1-\alpha)x_{32} = x_{31} + \alpha x_{21}$, which reduces to $x_{21} + x_{32} = x_{31}$. Similarly, in a second case, in which $x_{21} + x_{43} = x_{41} + x_{32}$, substituting x_{43} with $\alpha x_{31} + (1-\alpha)x_{32}$ and x_{41} with $(1-\alpha)x_{21}$, we obtain $x_{21} + x_{31} = x_{32}$. In the third case, $x_{31} + x_{42} = x_{41} + x_{32}$, the accompanying inequalities $x_{21} + x_{43} \leq x_{31} + x_{42}$ and $x_{21} + x_{43} \leq x_{41} + x_{32}$ give rise to $x_{21} \leq 0$ when substituting $x_{41} = (1-\alpha)x_{21}$, $x_{42} = \alpha x_{21}$, and $x_{43} = \alpha x_{31} + (1-\alpha)x_{32}$, contradicting the assumption that all off-diagonal entries of matrix $\mathbf{D}^{(n)}$ are strictly positive. We conclude that $\mathbf{D}^{(4)}$ is additive if and only if $x_{21} + x_{32} = x_{31}$ or $x_{21} + x_{31} = x_{32}$. These equations each force the four taxa to be collinear (Kopelman et al. 2013). Note that for the general n-taxon case of additive $\mathbf{D}^{(n)}$, the characterization of Kopelman et al. (2013) has an error ("Appendix B").

In this section, without imposing the additivity constraint, we explore the behavior of NJ using input distance matrices of the form of Eq. 6. We investigate two scenarios

involving the admixture: first, when the original $(n-1) \times (n-1)$ distance matrix is random, and second, when the original $n-1$ populations follow treelike evolution (i.e., the source distance matrix $\mathbf{S}^{(n-1)}$ is additive). For each case, we vary the number of taxa and the admixture fraction.

3.1 Random Source Matrix with Admixture

As a direct extension of the null model simulation, we generated random source distance matrices with $n-1$ populations, and we then introduced an admixed taxon t_n. In this model, the $n-1$ non-admixed source populations have random and independent distances, and distances involving the nth admixed taxon have a linear relationship with the distances involving taxa t_1 and t_2.

For a given number of taxa n, we generated an $(n-1) \times (n-1)$ random symmetric source distance matrix $\mathbf{S}^{(n-1)}$, which was not necessarily additive, by sampling the entries in the lower triangle from the uniform distribution on $[0, 1]$ as we did in Sect. 2.2. We set t_1 and t_2 as the source taxa, and the admixed taxon t_n was created as a linear admixture of the source taxa t_1 and t_2. The distances between the admixed taxon t_n and the other taxa were computed using the relation in Eq. 5. The resulting $n \times n$ admixed distance matrix $\mathbf{D}^{(n)}$ was used as an input to the NJ algorithm. The admixture fraction α was varied from 0.01 to 0.99 at an increment of 0.01. For each number of taxa and each admixture fraction, we generated 1000 random matrices $\mathbf{S}^{(n-1)}$, producing 99,000 total $\mathbf{D}^{(n)}$ matrices for a given n. On each inferred tree, we examined the three properties.

3.1.1 $n = 4$

Table 3 shows the numbers of violations of the three properties in the sets of 99,000 matrices for each n. For the case of $n = 4$, no violations are observed, in accord with the mathematical result of Kopelman et al. (2013), that, for a 4-taxon scenario, all three properties hold even without the additivity constraint, as long as an input distance matrix follows the form of Eq. 6.

3.1.2 $n \geq 5$

Property 1: As illustrated in Table 3 and Fig. 4a, the probability for Property 1 to be violated depends on the number of taxa. Initially, the number of violations for Property 1 increases rapidly. Among the values $n = 4, 5, \ldots, 50$, it reaches its maximum at $n = 13$, where the number of violations is 2314 (2.3%). Once the number of taxa passes $n = 13$, the number of violations steadily decreases and approaches a plateau of values of 750 to 900 for $n > 30$ (0.75–0.9%).

In the null model in Sect. 2, the probability of violations for Property 1 stays constant at 1/3 across all n. When admixture is introduced, however, Property 1 has a substantially higher chance of being satisfied. Even at its maximum of 2.3% at $n = 13$, the probability that Property 1 fails is much smaller than in the null model.

Table 3 Percentage of violations of the three properties among 99,000 inferred NJ trees using random source distance matrices with admixture, where the taxon t_n is a linear admixture of taxa t_1 and t_2

n	Property Violated						
	1	2	3	1 & 2	1 & 3	2 & 3	1 & 2 & 3
4	0	0	0	0	0	0	0
5	0.1	3.4	0.1	< 0.05	0.1	< 0.05	< 0.05
6	0.4	7.3	0.1	0.1	0.1	< 0.05	< 0.05
7	0.8	10.6	0.2	0.3	0.2	0.2	0.1
8	1.3	13.4	0.4	0.6	0.4	0.3	0.3
9	1.5	15.6	0.6	0.7	0.5	0.6	0.4
10	1.9	17.3	0.9	1.0	0.6	0.8	0.6
11	2.0	19.2	1.1	1.1	0.7	1.0	0.6
12	2.2	20.9	1.3	1.3	0.8	1.1	0.7
13	2.3	22.1	1.5	1.4	0.9	1.3	0.8
14	2.3	23.1	1.6	1.4	0.9	1.4	0.8
15	2.3	24.7	1.8	1.4	0.8	1.5	0.8
16	2.2	25.6	1.8	1.4	0.8	1.6	0.7

The distances between the $n-1$ source taxa are independently chosen from the uniform distribution on the interval [0, 1], and the distances involving the admixed taxon t_n are computed using Eq. 5. The numbers shown are aggregated from 99 sets of 1000 simulations, each set with a distinct admixture fraction in $\{0.01, 0.02, \ldots, 0.99\}$

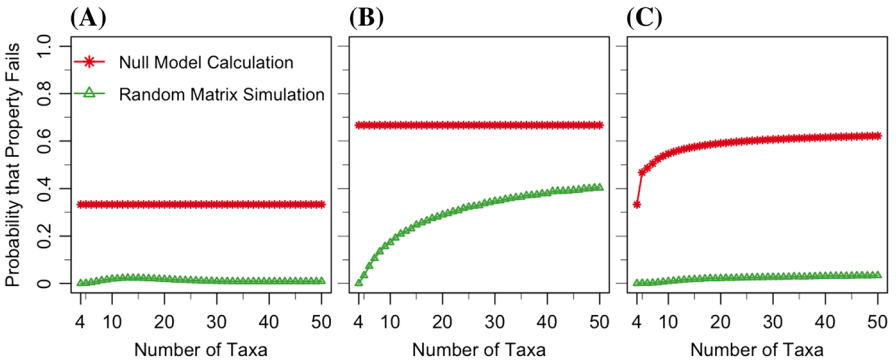

Fig. 4 Simulated probabilities that Properties 1, 2, and 3 are violated under a model in which $(n-1)\times(n-1)$ source distance matrices $\mathbf{S}^{(n-1)}$ are generated randomly and an admixed nth taxon is added using Eq. 5. **a** Property 1. **b** Property 2. **c** Property 3. The exact probability for a random labeled unrooted binary tree to violate the properties and the simulated probability for an NJ tree inferred from a random distance matrix with admixture to violate the properties are both shown. For the null model, the values are copied from Fig. 3. For the random matrix model, simulation values are based on 99,000 matrices for each n; the values for $n = 4$ to $n = 16$ appear in Table 3

Property 2: Our simulation finds a rapid increase in the number of violations for Property 2 as n increases for small $n < 10$ (Fig. 4b). The rate of increase slows when n becomes large.

The dependence of Property 2 on the number of taxa n differs from the null model simulation, in which the number of violations is constant across all n. From $n = 4$ to 6, the number of violations for Property 2 is an order of magnitude smaller than in the null model. The deviation from the null model for Property 2 reduces gradually as n increases. At $n = 29$, the number of violations reaches 50% of the value from the null model. The effect of admixture in reducing violations compared to the null model is not as drastic for Property 2 as for Property 1.

Property 3: As shown in Table 3, when $n = 5$, the number of Property 3 violations is the same as that of Property 1, 107 among 99,000 simulations. Interestingly, all 107 violations of Properties 1 and 3 occur on the same set of distance matrices. The number of matrices $\mathbf{D}^{(5)}$ violating both properties is also 107.

The number of violations of Property 3 at $n = 6$ is 109, close to the result at $n = 5$. Once $n > 6$, the number of violations for Property 3 increases steadily from 224 (0.2%) for $n = 7$ to 1788 (1.8%) for $n = 16$. Compared to the null model, the fraction of input distance matrices violating Property 3 is considerably smaller. As was seen for Property 1, Property 3 is more sensitive to the introduction of admixture than is Property 2.

3.1.3 $n = 5$

The $n = 5$ case is of particular interest as $n = 5$ is the smallest number of taxa giving violations of the properties, and it has only one unrooted unlabeled bifurcating tree topology. Therefore, to understand the effects of model parameters in more detail, for $n = 5$, we performed simulations with a larger number of replicates, 100,000, for each admixture fraction in $\{0.01, 0.02, \ldots, 0.99\}$.

For each $(n-1) \times (n-1)$ input source distance matrix $\mathbf{S}^{(n-1)}$, we estimated the $(n-1)$-taxon source tree $\mathcal{T}_S^{(n-1)}$ using NJ. The corresponding admixed NJ tree $\mathcal{T}_D^{(n)}$ was inferred from the associated $\mathbf{D}^{(n)}$. Based on their topologies, the source NJ trees $\mathcal{T}_S^{(n-1)}$ were classified into three possible unrooted labeled binary 4-taxon topologies, and the admixed NJ trees $\mathcal{T}_D^{(n)}$ were assigned to one of 15 possible unrooted labeled binary 5-taxon topologies. We then characterized the inferred topologies in relation to Properties 1 and 3, as these properties closely depend on tree topology.

Table 4 aggregates the topology classifications of $\mathcal{T}_S^{(4)}$ and $\mathcal{T}_D^{(5)}$ from 99 sets of 100,000 matrices, each with a distinct admixture fraction, totaling 9,900,000 matrices. The three unrooted labeled binary 4-taxon source topologies are equally likely to be produced. This result accords with the $n = 4$ case in our null model simulation in Sect. 2.2.

Topology of $\mathcal{T}_D^{(5)}$: Unlike the source NJ trees, the admixed NJ trees $\mathcal{T}_D^{(5)}$ display an uneven distribution of topologies (Table 4B). For notational simplicity, we use t_S to indicate either of the source taxa, t_1 and t_2. We indicate the admixed taxon t_5 by t_A, and we use t_O to indicate either of the other taxa, t_3 and t_4. We also define $(B_1, B_2) \equiv (\max\{b_{t_1t_5}, b_{t_2t_5}\}, \min\{b_{t_1t_5}, b_{t_2t_5}\})$.

The most frequent topologies have shape $((t_S, t_A), t_S, (t_O, t_O))$. Two topologies have this form, with 18.86 and 18.87% occurrences. They possess topological attributes

Table 4 Topology classification of inferred NJ trees at $n = 5$ using random source distance matrices with admixture

Topology $\mathcal{T}_S^{(4)}$	Number of observations	Frequency (%)	
(A)			
$((t_1, t_2), (t_3, t_4))$	3,298,318	33.32	
$((t_1, t_3), (t_2, t_4))$	3,302,590	33.36	
$((t_1, t_4), (t_2, t_3))$	3,299,092	33.32	
Topology $\mathcal{T}_D^{(5)}$	Topology category	Number of observations	Frequency (%)
(B)			
$((t_4, t_5), t_1, (t_2, t_3))$	$((t_O, t_A), t_S, (t_S, t_O))$	3117	0.03
$((t_3, t_5), t_1, (t_2, t_4))$	$((t_O, t_A), t_S, (t_S, t_O))$	3084	0.03
$((t_2, t_5), t_1, (t_3, t_4))$	$((t_S, t_A), t_S, (t_O, t_O))$	1,867,130	18.86
$((t_4, t_5), t_2, (t_1, t_3))$	$((t_O, t_A), t_S, (t_S, t_O))$	3075	0.03
$((t_3, t_5), t_2, (t_1, t_4))$	$((t_O, t_A), t_S, (t_S, t_O))$	3121	0.03
$((t_1, t_5), t_2, (t_3, t_4))$	$((t_S, t_A), t_S, (t_O, t_O))$	1,868,166	18.87
$((t_4, t_5), t_3, (t_1, t_2))$	$((t_O, t_A), t_O, (t_S, t_S))$	0	0
$((t_2, t_5), t_3, (t_1, t_4))$	$((t_S, t_A), t_O, (t_S, t_O))$	936,540	9.46
$((t_1, t_5), t_3, (t_2, t_4))$	$((t_S, t_A), t_O, (t_S, t_O))$	936,777	9.46
$((t_3, t_5), t_4, (t_1, t_2))$	$((t_O, t_A), t_O, (t_S, t_S))$	0	0
$((t_2, t_5), t_4, (t_1, t_3))$	$((t_S, t_A), t_O, (t_S, t_O))$	936,127	9.46
$((t_1, t_5), t_4, (t_2, t_3))$	$((t_S, t_A), t_O, (t_S, t_O))$	934,402	9.44
$((t_3, t_4), t_5, (t_1, t_2))$	$((t_O, t_O), t_A, (t_S, t_S))$	0	0
$((t_2, t_4), t_5, (t_1, t_3))$	$((t_S, t_O), t_A, (t_S, t_O))$	1,204,777	12.17
$((t_2, t_3), t_5, (t_1, t_4))$	$((t_S, t_O), t_A, (t_S, t_O))$	1,203,684	12.16

(A) NJ-inferred source tree ($\mathcal{T}_S^{(4)}$). (B) NJ-inferred tree with admixture ($\mathcal{T}_D^{(5)}$). Here, t_1 and t_2 are source taxa and t_5 is the admixed taxon. The topologies of trees appear in Newick format. The numbers shown are aggregated from 99 sets of 100,000 simulations, each set with a distinct admixture fraction in $\{0.01, 0.02, \ldots, 0.99\}$

commonly observed in NJ trees containing an admixed population. Specifically, the admixed taxon branches off the path connecting the two source taxa. The admixed taxon and one of the source taxa form a cherry, satisfying Property 1. The numbers of edges separating the source taxa and the admixed taxon are minimal: $(B_1, B_2) = (3, 2)$. These trees have $b_{t_1 t_2} = 3 \geq B_1$, so Property 3 holds true as well.

Two topologies, represented by $((t_S, t_O), t_A, (t_S, t_O))$, are the next most commonly inferred labeled topologies, with frequencies 12.16 and 12.17%. In these trees, the admixed taxon is incident to an internal edge joining two cherries. Each cherry contains one of the two source taxa, so that the topological distance from the admixed taxon to either of the two source taxa is three, $(B_1, B_2) = (3, 3)$, whereas $b_{t_1 t_2} = 4 \geq B_1$. The topologies therefore satisfy Property 3. In the first step of the NJ agglomeration, either cherry is formed. In the second step of the NJ algorithm, by adopting the tie-breaking rule from Sect. 2, the admixed taxon clusters with one of the cherries containing a source taxon. Thus, Property 1 is satisfied.

The next most frequent topologies have shape $((t_S, t_A), t_O, (t_S, t_O))$. This category has four topologies, each occurring with frequency $\sim$9%. Two source taxa belong to separate cherries, and the admixed taxon shares one of the cherries with a source taxon. This structure gives $(B_1, B_2) = (4, 2)$ and $b_{t_1 t_2} = 4$, satisfying Property 3. If (t_S, t_A) is the first cherry of the NJ algorithm, then Property 1 holds. If (t_S, t_O) is the first cherry, then by our convention in the event of a tie, (t_S, t_A) becomes the second cluster in constructing the final tree shape $((t_S, t_A), t_O, (t_S, t_O))$, and thus, Property 1 holds under this condition as well.

All topologies with high frequencies ($>9\%$) satisfy Properties 1 and 3. Some topologies for which the admixed taxon does not appear on the path connecting two source taxa, however, do occur. Four topologies of the form $((t_O, t_A), t_S, (t_S, t_O))$ appear at frequency 0.03%. With $(B_1, B_2) = (4, 3)$ and $b_{t_1 t_2} = 3$, one source taxon is maximally separated from the admixed taxon, and the topologies fail to satisfy Property 3.

Three topologies that violate Property 3 are never observed. Two of these topologies have shape $((t_O, t_A), t_O, (t_S, t_S))$, and the third has the form $((t_O, t_O), t_A, (t_S, t_S))$; all contain the two source taxa as a cherry. Both source taxa in topologies $((t_O, t_A), t_O, (t_S, t_S))$ are maximally separated from the admixed taxon: $(B_1, B_2) = (4, 4)$. The source taxa in a topology $((t_O, t_O), t_A, (t_S, t_S))$ have $(B_1, B_2) = (3, 3)$. Because the source taxa form a cherry and $b_{t_1 t_2} = 2$, all three topologies violate Property 3.

The topologies that are never observed also fail Property 1, as none of the possible clustering sequences producing $((t_O, t_A), t_O, (t_S, t_S))$ or $((t_O, t_O), t_A, (t_S, t_S))$ involves a clustering of clades containing a source taxon and the admixed taxon prior to clustering of the two clades containing the source taxa. This observation leads to the following propositions, the proofs of which appear in "Appendix C".

Proposition 1 *Suppose an input distance matrix to the NJ algorithm has the form in Eq. 6 with $n = 5$, t_1 and t_2 being source taxa, and t_5 being an admixed taxon. Then the following pairs do not cluster in the first step of the NJ algorithm: (t_1, t_2), (t_3, t_5), and (t_4, t_5).*

Proposition 2 *Under the same conditions on the input distance matrix as in Proposition 1, the two source populations t_1 and t_2 cannot form a cherry in a final NJ-estimated tree.*

Our simulation finds that an NJ tree that violates Property 1 has one of four topologies: $((t_4, t_5), t_1, (t_2, t_3))$, $((t_3, t_5), t_1, (t_2, t_4))$, $((t_4, t_5), t_2, (t_1, t_3))$, and $((t_3, t_5), t_2, (t_1, t_4))$. We claim that these four are in fact the only possible topologies that violate Property 1. In the first step of NJ, of $\binom{5}{2} = 10$ possible pairs, Proposition 1 shows that (t_1, t_2), (t_3, t_5), and (t_4, t_5) cannot agglomerate. If (t_1, t_5) or (t_2, t_5) is the first cherry, then Property 1 is satisfied. We are left with five potential pairs—(t_1, t_3), (t_1, t_4), (t_2, t_3), (t_2, t_4), and (t_3, t_4)—as the possible first cherry of an NJ tree whose construction might violate Property 1.

Suppose that NJ clusters a pair (t_3, t_4) first. In the second step, NJ clusters two taxa from $\{t_1, t_2, t_5, t_{34}\}$, where t_{34} is a node representing the (t_3, t_4) cluster. As shown in Eq. 9, a tie always occurs in the choice of the next pair to agglomerate when four

nodes remain. Applying our convention in case of a tie described in Sect. 2, the possible second cherries are (t_1, t_5), (t_2, t_5), (t_1, t_2), and (t_5, t_{34}). The first two choices satisfy Property 1, and we can exclude them. The last two choices result in (t_1, t_2) being a cherry, violating Proposition 2. Therefore, when (t_3, t_4) is the first cherry, no violations in Property 1 occur.

Now consider the remaining four potential pairs for the first cherry. Each pair contains one taxon from $\{t_1, t_2\}$ and one taxon from $\{t_3, t_4\}$. It suffices to examine (t_1, t_3) because the roles of source taxa t_1 and t_2 are interchangeable, as are the roles of t_3 and t_4. If (t_1, t_3) is the first cherry, then the remaining nodes are $\{t_2, t_4, t_5, t_{13}\}$, where t_{13} represents the (t_1, t_3) cluster. Again applying our tie-breaking rule, the possible second cherries are (t_2, t_5), (t_5, t_{13}), (t_2, t_{13}), and (t_4, t_5). The first two choices satisfy Property 1. The last two pairs have a tie at the 4-taxon stage of the NJ algorithm, and either choice violates Property 1. Therefore, the resulting topology, $((t_4, t_5), t_2, (t_1, t_3))$, arising from a choice of (t_2, t_{13}) or (t_4, t_5) as the second cherry, is the only topology that violates Property 1 when the first cherry is (t_1, t_3). By exchanging t_1 and t_2 or t_3 and t_4 or both, we can deduce Corollary 1.

Corollary 1 *When* $\mathbf{D}^{(5)}$ *is used as an input,* $((t_4, t_5), t_1, (t_2, t_3))$, $((t_3, t_5), t_1, (t_2, t_4))$, $((t_4, t_5), t_2, (t_1, t_3))$, *and* $((t_3, t_5), t_2, (t_1, t_4))$ *are the only topologies possible when the NJ algorithm violates Property 1.*

A similar analysis can be carried out for Property 3. Of 15 unrooted labeled binary tree topologies for $n = 5$, 7 fail to satisfy $b_{t_1t_2} \geq b_{t_1t_5}$ or $b_{t_1t_2} \geq b_{t_2t_5}$ (shaded boxes in Figures S1–S3). However, not all of them are attainable by NJ with the admixed distance matrix in Eq. 6 as an input. This observation can be explained by direct application of Proposition 2. Excluding three topologies containing (t_1, t_2) as a cherry from the 7 topologies violating Property 3, we reach the following Corollary.

Corollary 2 *When* $\mathbf{D}^{(5)}$ *is used as an input,* $((t_4, t_5), t_1, (t_2, t_3))$, $((t_3, t_5), t_1, (t_2, t_4))$, $((t_4, t_5), t_2, (t_1, t_3))$, *and* $((t_3, t_5), t_2, (t_1, t_4))$ *are the only topologies possible when the NJ algorithm violates Property 3.*

From Corollaries 1 and 2, we can deduce that input admixed distance matrices $\mathbf{D}^{(5)}$ violating Property 1 also fail Property 3 and vice versa, as any time one of the four topologies in the corollaries is produced, it violates Property 3. By Proposition 1, it must be produced in an order that violates Property 1. Our simulation result from 9,900,000 runs confirms that Properties 1 and 3 coincide when $n = 5$.

Topology of $\mathcal{T}_S^{(4)}$: Because an admixed distance matrix $\mathbf{D}^{(n)}$ is created from a source distance matrix $\mathbf{S}^{(n-1)}$ by a linear transformation, for a given admixture fraction α, the independent variables in $\mathbf{D}^{(n)}$ remain the same as the ones in $\mathbf{S}^{(n-1)}$. For this reason, the underlying structure for the $n - 1$ non-admixed populations affects the topology of an NJ tree with admixture.

Table 4A shows that $\mathcal{T}_S^{(4)}$ has a uniform distribution across all three unrooted labeled binary 4-taxon tree topologies. This distribution changes, however, when we condition on the properties being attained. We choose input admixed distance matrices $\mathbf{D}^{(5)}$ whose inferred NJ trees $\mathcal{T}_D^{(5)}$ fail Property 3 and construct their source NJ trees

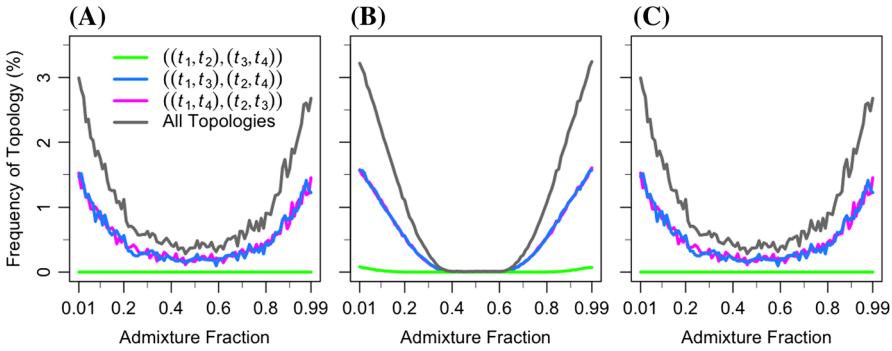

Fig. 5 Frequencies of unrooted labeled binary tree topologies for $\mathcal{T}_S^{(4)}$ inferred from random source distance matrices $\mathbf{S}^{(4)}$, considering all $\mathcal{T}_S^{(4)}$ whose corresponding admixed distance matrices $\mathbf{D}^{(5)}$ violate one of the three properties. **a** Property 1. **b** Property 2. **c** Property 3. For each admixture fraction, a source distance matrix $\mathbf{S}^{(4)}$ giving rise to a violation of each property is analyzed, and the frequencies are computed conditioning on the total number of violations of each property. Each source NJ tree inferred from $\mathbf{S}^{(4)}$ is assigned to one of three possible unrooted labeled binary topologies with four taxa. The "All Topologies" line sums the values in the other three lines, and thus represents each property's dependence on the admixture fraction without regard to the topology of $\mathcal{T}_S^{(4)}$. In each panel, values appear for 99 admixture fractions, summing to 100%. The simulations shown, 100,000 at each value of α, are the same as those considered in Table 4

from $\mathbf{S}^{(4)}$. Because Properties 1 and 3 coincide for $n = 5$, we consider Property 3 only.

Of 12,397 NJ trees of $\mathbf{S}^{(4)}$ resulting in Property 3 violations, 50.32% have topology $((t_1, t_4), (t_2, t_3))$, 49.68% have topology $((t_1, t_3), (t_2, t_4))$, and none has topology $((t_1, t_2), (t_3, t_4))$. When $\mathcal{T}_S^{(4)}$ has topology $((t_1, t_4), (t_2, t_3))$, its admixed distance matrix results in either $((t_4, t_5), t_1, (t_2, t_3))$ or $((t_3, t_5), t_2, (t_1, t_4))$ with approximately equal frequency. Similarly, when $\mathcal{T}_S^{(4)}$ has topology $((t_1, t_3), (t_2, t_4))$, its admixed distance matrix results in either $((t_3, t_5), t_1, (t_2, t_4))$ or $((t_4, t_5), t_2, (t_1, t_3))$ with approximately equal frequency. When $\mathcal{T}_S^{(4)}$ has topology $((t_1, t_2), (t_3, t_4))$, the resulting admixed NJ tree $\mathcal{T}_D^{(5)}$ always satisfies Property 3 (Fig. 5c). This observation leads to the following proposition, the proof of which appears in "Appendix C".

Proposition 3 *If $\mathcal{T}_S^{(4)}$, inferred from a source distance matrix $\mathbf{S}^{(4)}$, has topology $((t_1, t_2), (t_3, t_4))$, then its corresponding admixed NJ tree $\mathcal{T}_D^{(5)}$ always satisfies Properties 1 and 3 and always contains either (t_1, t_5) or (t_2, t_5) as a cherry.*

Admixture Fraction Dependence: Having performed a detailed analysis on topological characteristics of source and admixed NJ trees violating Properties 1 and 3, we now examine the effect of the admixture fraction α on the properties. As can be seen in Fig. 5, the number of inferred admixed NJ trees violating Properties 1, 2, and 3 is minimal near $\alpha = 0.5$, and it increases as α deviates from 0.5. The frequency of violations is symmetric around $\alpha = 0.5$; the behavior of the source taxon t_1 at admixture fraction α is equivalent to that of the source taxon t_2 at admixture fraction $1-\alpha$. Because the roles of the two source taxa t_1 and t_2 are interchangeable, it suffices

to consider $\alpha \leq 0.5$. We examine three cases: $\alpha = 0.01$ (Figure S1), $\alpha = 0.25$ (Figure S2), and $\alpha = 0.5$ (Figure S3).

We infer the 4-taxon source NJ tree $\mathcal{T}_S^{(4)}$ from a random source distance matrix $\mathbf{S}^{(4)}$. Based on its topology, each source NJ tree is characterized into one of three possible 4-taxon unrooted labeled binary trees, which we denote $\{T_1^{(4)}, T_2^{(4)}, T_3^{(4)}\}$. For a given admixture fraction α, an admixed distance matrix $\mathbf{D}^{(5)}$ is constructed, and its admixed NJ tree $\mathcal{T}_D^{(5)}$ is estimated. We categorize each admixed NJ tree into one of 15 possible 5-taxon unrooted labeled binary trees, represented by $\{T_1^{(5)}, T_2^{(5)}, \ldots, T_{15}^{(5)}\}$. In each graph in Figures S1–S3, we represent each topology with a node and connect a $(T_i^{(4)}, T_j^{(5)})$ pair with a directed edge from a 4-taxon topology to a 5-taxon topology when a source and admixed NJ pair having such topologies appear in our simulation. The edge weight is proportional to the frequency with which a connected pair occurs. For $n = 5$, we have shown that Properties 1 and 3 coincide, so we focus on Property 3 only.

When $\alpha = 0.01$, the proportion of ancestry from population 1 in the admixed population is small and ancestry from population 2 is large. As illustrated in Figure S1, most estimated admixed NJ trees satisfying Properties 1 and 3 have (t_2, t_5) as a cherry. Owing to the small admixture fraction, t_2 and t_5 are the most closely related pair, and NJ picks (t_2, t_5) as a cluster with high probability. Of the four feasible topologies $\mathcal{T}_D^{(5)}$ violating Property 3, the ones in which t_2 and t_5 are maximally separated have the highest frequency among violations: $((t_3, t_5), t_1, (t_2, t_4))$ with 53.30% and $((t_4, t_5), t_1, (t_2, t_3))$ with 45.12%.

When $\alpha = 0.25$, population 2 still contributes a majority of the ancestry to population 5, but less so than for $\alpha = 0.01$. Among $(\mathcal{T}_S^{(4)}, \mathcal{T}_D^{(5)})$ topology pairs satisfying Property 3 (Figure S2A), three pairs with the highest frequencies place the admixed taxon t_5 on the t_2 leaf of $\mathcal{T}_S^{(4)}$ and thereby contain (t_2, t_5) as a cherry in $\mathcal{T}_D^{(5)}$. The frequencies of those pairs decrease from ∼26% to ∼17–23%, and other $(\mathcal{T}_S^{(4)}, \mathcal{T}_D^{(5)})$ pairs increase in frequency. As in the $\alpha = 0.01$ case, among $\mathcal{T}_D^{(5)}$ topologies violating Property 3, $((t_3, t_5), t_1, (t_2, t_4))$ and $((t_4, t_5), t_1, (t_2, t_3))$ occur most often, with frequencies 53.93% and 37.08%, respectively. From $\mathcal{T}_D^{(5)}$ topologies accessible by NJ, the other two topologies violating Property 3, $((t_3, t_5), t_2, (t_1, t_4))$ and $((t_4, t_5), t_2, (t_1, t_4))$, appear more often, 5.62% and 3.37%, compared to 1.06% and 0.53% when $\alpha = 0.01$.

When two sources contribute equally to the admixed taxon ($\alpha = 0.5$), all four feasible topologies $\mathcal{T}_D^{(5)}$ violating Property 3 are approximately equally likely, as shown in Figure S3: $((t_4, t_5), t_1, (t_2, t_3))$ with 24.14%, $((t_3, t_5), t_1, (t_2, t_4))$ with 27.59%, $((t_4, t_5), t_2, (t_1, t_3))$ with 25.86%, and $((t_3, t_5), t_2, (t_1, t_4))$ with 22.41%.

3.2 Random Additive Source Matrix with Admixture

In the previous section, we investigated the properties of inferred NJ trees when the underlying source distance matrices $\mathbf{S}^{(n-1)}$, from which the nth admixed taxon was created, were arbitrary. Here, we focus on a model in which the underlying $n - 1$ populations evolve independently in a treelike manner, and the admixture occurs

between two randomly sampled populations. In the absence of the admixed taxon, NJ recovers the true tree given additive $\mathbf{S}^{(n-1)}$ (Atteson 1999; Bryant 2005; Steel 2016). Inclusion of admixed taxon t_n in phylogeny reconstruction, however, can disrupt the NJ inference for the underlying $n-1$ populations as well as relationships between the source taxa and the admixed taxon on the inferred NJ tree.

Kopelman et al. proved that when the admixed distance matrix $\mathbf{D}^{(n)}$ is additive, Properties 2 and 3 must hold, irrespective of the number of taxa n. Under the only slightly weaker condition of assuming an additive source distance matrix $\mathbf{S}^{(n-1)}$, however, we find that they do not necessarily hold. In this section, we examine the three properties of the inferred NJ tree in the case that an admixed taxon is constructed from an additive source distance matrix.

To simulate an evolutionary tree with $n-1$ populations, we employed a Yule pure birth process (Yule 1925; Aldous 2001; Nee 2006) with birth rate $\lambda = 1$ and scaled the total tree height to equal 1. From each simulated tree, we computed a source distance matrix $\mathbf{S}^{(n-1)}$ by summing edge lengths along the shortest path between each pair of taxa on the tree. We then added an admixed taxon t_n and constructed an admixed distance matrix $\mathbf{D}^{(n)}$ from the additive $\mathbf{S}^{(n-1)}$ using Eq. 5. For each admixture fraction α in $\{0.01, 0.02, \ldots, 0.99\}$, we generated 1000 random admixed distance matrices $\mathbf{D}^{(n)}$ formed from random additive source distance matrices $\mathbf{S}^{(n-1)}$, resulting in 99,000 total matrices for a given number of taxa n. We varied n from 4 to 50. Note that the scaling of the source tree height—multiplying $\mathbf{S}^{(n-1)}$ by a constant—only scales the NJ branch lengths and does not affect the order of agglomeration, as the admixture step (Eq. 5), the NJ Q-criterion ("Appendix A"), and the inference of edge lengths all utilize linear transformations of the source distance matrix $\mathbf{S}^{(n-1)}$.

The simulated numbers of violations for Properties 1, 2, and 3 from $n = 4$ to 16 appear in Table 5. No violation is observed for Properties 1 and 3 across all numbers of taxa used in our simulation. Property 2, however, exhibits violations, the number of which depends both on the number of taxa n and on the admixture fraction α. When $n = 4$, Property 2 shows no violations. This result is consistent with the mathematical result of Kopelman et al. (2013) that all three properties hold when $n = 4$, irrrespective of the additivity of the input distance matrix $\mathbf{D}^{(n)}$. As n increases, the probability that Property 2 fails increases as well. The number of simulations violating Property 2 rapidly grows with a small number of taxa, from 0 to 20% for $n = 4$ to 10. The rate of increase slows gradually as n becomes large (Fig. 6b).

The probability that Property 2 holds also depends on the admixture fraction. For all numbers of taxa in our simulation, the number of violations is minimal when the admixture fraction is 0.5. In particular, when $n = 5$, Property 2 shows no violations in the range from $\alpha = 0.37$ to 0.63. As α deviates further from 0.5, the number of violations increases (Fig. 7), reaching maxima at the extreme values $\alpha = 0.01$ and 0.99. As discussed in Sect. 3.1, the dependence of the number of violations on the admixture fraction is symmetric with respect to $\alpha = 0.5$, as the roles of the source taxa, t_1 and t_2, are interchangeable.

To gain a better understanding of the dependence of Property 2 on α, we analyzed $\Delta_1 \equiv d_{t_1 t_n} - d_{t_1 t_2}$ and $\Delta_2 \equiv d_{t_2 t_n} - d_{t_1 t_2}$ for all NJ-inferred trees from input admixed distance matrices. Property 2 requires that $d_{t_1 t_n} < d_{t_1 t_2}$ and $d_{t_2 t_n} < d_{t_1 t_2}$, so NJ trees violating Property 2 have either $\Delta_1 \geq 0$ or $\Delta_2 \geq 0$. We find no occurrences in which

Table 5 Percentage of violations of the three properties among 99,000 inferred NJ trees using random additive source distance matrices with admixture, where the source distance matrix $\mathbf{S}^{(n-1)}$ is additive, and the taxon t_n is a linear admixture of taxa t_1 and t_2

n	Property violated						
	1	2	3	1 & 2	1 & 3	2 & 3	1 & 2 & 3
4	0	0	0	0	0	0	0
5	0	5.4	0	0	0	0	0
6	0	10.7	0	0	0	0	0
7	0	14.0	0	0	0	0	0
8	0	16.4	0	0	0	0	0
9	0	18.3	0	0	0	0	0
10	0	19.8	0	0	0	0	0
11	0	21.3	0	0	0	0	0
12	0	22.6	0	0	0	0	0
13	0	23.8	0	0	0	0	0
14	0	24.9	0	0	0	0	0
15	0	25.6	0	0	0	0	0
16	0	26.6	0	0	0	0	0

The distances between the $n-1$ source taxa correspond to pairwise distances between tips in a randomly generated tree, and the distances involving the admixed taxon t_n are computed using Eq. 5. The numbers shown are aggregated from 99 sets of 1000 simulations, each set with a distinct admixture fraction in $\{0.01, 0.02, \ldots, 0.99\}$

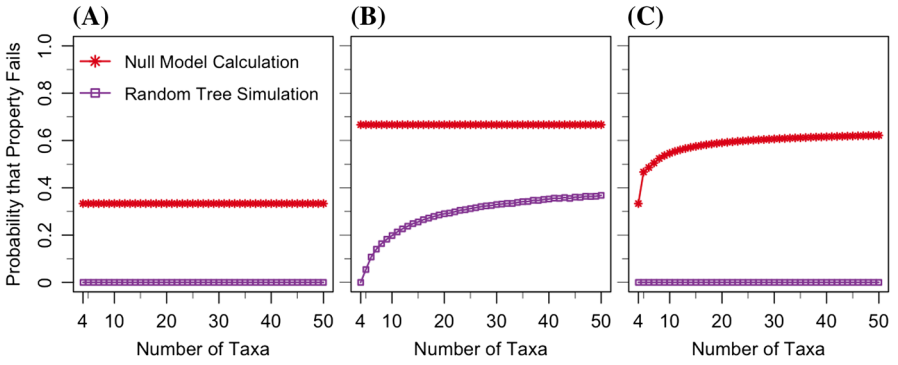

Fig. 6 Simulated probabilities that Properties 1, 2, and 3 are violated under a model in which $(n-1)\times(n-1)$ source distance matrices $\mathbf{S}^{(n-1)}$ are additive, and an admixed nth taxon is added using Eq. 5. **a** Property 1. **b** Property 2. **c** Property 3. The exact probability for a random labeled unrooted binary tree to violate the properties is also shown; for the null model, the values are copied from Fig. 3. For the random tree model, simulation values are based on 99,000 matrices for each n; the values for $n = 4$ to $n = 16$ appear in Table 5

$\mathbf{D}^{(n)}$ results in both $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$, so that if $\mathbf{D}^{(n)}$ generated from an additive $\mathbf{S}^{(n-1)}$ violates Property 2, then only one of $d_{t_1 t_n}$ and $d_{t_2 t_n}$ is observed to be greater than or equal to $d_{t_1 t_2}$. For this reason, we can assign $\Delta = \max(\Delta_1, \Delta_2)$ for each NJ tree violating Property 2, and the quantity Δ can be interpreted as a degree of violation of

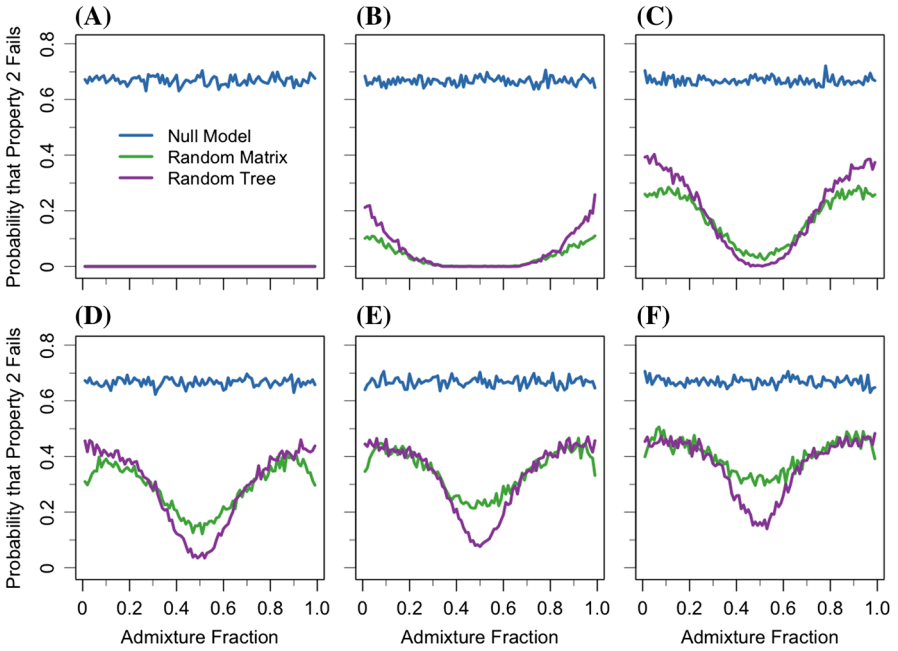

Fig. 7 Dependence of the probability that Property 2 is violated on the admixture fraction α under a model in which an admixed distance matrix $\mathbf{D}^{(n)}$ is constructed from an additive $(n-1) \times (n-1)$ source distance matrix $\mathbf{S}^{(n-1)}$. **a** $n = 4$. **b** $n = 5$. **c** $n = 10$. **d** $n = 20$. **e** $n = 30$. **f** $n = 50$. Simulated values from the null model without admixture and a random source matrix model are also shown. The simulations shown are the same ones considered for the null model in Fig. 3, the random matrix model in Fig. 4, and the random tree model in Fig. 6. Because the null model has no parameter α, each of 99 values of α is randomly associated with 1000 simulations

Property 2. Whether $\Delta = \Delta_1$ or Δ_2 depends on α. For lower numbers of taxa, $n < 10$, all NJ trees have $\Delta = \Delta_1$ when $\alpha < 0.5$ and $\Delta = \Delta_2$ when $\alpha > 0.5$. As n increases, it is possible to observe either $\Delta = \Delta_1$ or $\Delta = \Delta_2$ for $\alpha \approx 0.5$ (Fig. 8).

We have seen in Fig. 7 that the number of Property 2 violations is maximal when the admixture fraction is skewed toward 0 or 1. For those admixture fractions close to the boundary, however, Δ is small (Fig. 8). This result can be explained as follows. For small admixture fractions ($\alpha \ll 1$), most of the genetic contribution to t_n comes from source population t_2. Based on Eq. 5, the distance between t_n and t_2 is small, and distances from all the other taxa t_i to t_n ($x_{t_i t_n}$) are close to distances $x_{t_i t_2}$. Therefore, t_n and t_2 likely form a cherry, and the admixed taxon t_n behaves like the source taxon t_2 on the NJ-inferred tree $\mathcal{T}_D^{(n)}$. If we denote a node representing the most recent common ancestor of t_2 and t_n as C, then $d_{t_2 C} \approx d_{t_n C}$ and $\Delta_1 = d_{t_1 t_n} - d_{t_1 t_2} = (d_{t_1 C} + d_{t_n C}) - (d_{t_1 C} + d_{t_2 C}) = d_{t_n C} - d_{t_2 C} \ll 1$. The same explanation applies to $\Delta_2 \ll 1$ at large admixture fractions ($1 - \alpha \ll 1$).

As α deviates further from the boundary values, Property 2 violations become less frequent, but the degree of the violation Δ is larger than when α is near 0 or 1. When $n = 5$, the maximal Δ_1 occurs at $\alpha = 0.18$ and the maximal Δ_2 at $\alpha = 0.84$. As the number of taxa increases, the admixture fraction giving the maximal degree of violation shifts

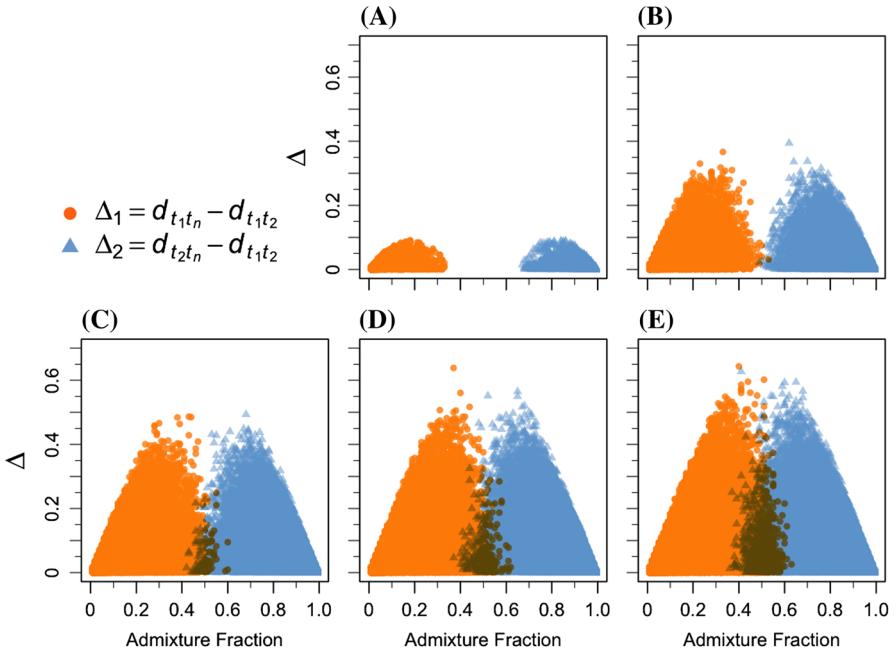

Fig. 8 The Δ measurement of the magnitude of Property 2 violations and its dependence on admixture fractions α under the random additive source matrix model with admixture. **a** $n = 5$. **b** $n = 10$. **c** $n = 20$. **d** $n = 30$. **e** $n = 50$. For each n, among 99,000 simulations, 1000 for each admixture fraction α in $\{0.01, 0.02, \ldots, 0.99\}$, NJ trees violating Property 2 are plotted. For each NJ tree violating Property 2, the degree of violation Δ is computed. Because no NJ tree violates both $d_{t_1t_2} > d_{t_1t_n}$ and $d_{t_1t_2} > d_{t_2t_n}$, each NJ tree violating Property 2 is assigned to either $\Delta = \Delta_1 = d_{t_1t_n} - d_{t_1t_2}$ (orange) or $\Delta = \Delta_2 = d_{t_2t_n} - d_{t_1t_2}$ (blue). The simulations shown are the same ones considered for the random tree model in Fig. 7

toward $\alpha = 0.5$. These results suggest that when the admixed population is formed from similar proportions of two source populations, the probability that Property 2 is violated is smaller. If a violation does occur, however, then the chance is higher that its magnitude is higher compared to the case in which either one of two populations is a major source for the admixture.

A comparison of Property 2 under all three models appears in Fig. 7. When $n = 4$, no difference is observed between the random source matrix model and the additive source matrix model because all properties are satisfied irrespective of the source matrix used. Both models involving admixture show fewer violations compared to the null model without admixture. For $n = 5$ to 20, compared to the random source matrix model (Section 3.1), Property 2 fails more frequently when the source distance matrix $\mathbf{S}^{(n-1)}$ is additive. For $n > 20$, the random source matrix model displays more violations, and the difference between the two models increases with the number of taxa. Across the number of taxa used, the additive source distance matrix model consistently shows a stronger dependence on the admixture fraction.

3.3 Perturbed Random Additive Source Matrix with Admixture

The previous section examined source population distance matrices that represented the outcome of an ideal treelike evolution. Next, to model deviations from additivity that occur in distance matrices, owing to the frequently imperfect representation by distance matrices of an underlying treelike evolutionary descent, we simulated scenarios that consider "noise" added to an underlying additive metric.

For a given number of taxa n, we followed the procedure of Sect. 3.2 to generate additive source distance matrices $\mathbf{S}^{(n-1)}$ from simulated random trees. For each entry d_{ij} in $\mathbf{S}^{(n-1)}$, we added a random noise term drawn from a truncated normal distribution whose underlying untruncated normal random variable had mean 0 and variance σ^2. To keep the perturbed distances between taxa nonnegative, the distribution from which the noise term was drawn was truncated below at $-d_{ij}$. From the perturbed distance matrix $\tilde{\mathbf{S}}^{(n-1)}$, the nth taxon was created as an admixture of source taxa t_1 and t_2, and the distances between the admixed taxon t_n and the other taxa were computed using Eq. 5. The resulting $n \times n$ perturbed admixed distance matrix $\tilde{\mathbf{D}}^{(n)}$ was used as an input to the NJ algorithm.

For each n, we varied the perturbation of matrices $\tilde{\mathbf{S}}^{(n-1)}$ from additivity by choosing values of σ from 0.00 to 1.00 with increment 0.01. We also varied the admixture fraction α from 0.01 to 0.99 with increment 0.01. For each (σ, α) pair, we simulated 1000 distance matrices, giving 9,999,000 in total.

Figure 9 displays the frequency with which Properties 1, 2, and 3 fail to hold for $n = 4, 5, 8, 12$, and 16 as functions of the perturbation σ and the admixture fraction α. For $\sigma = 0$, no noise is added to the additive $\mathbf{S}^{(n-1)}$, and the failure percentages are expected to match those of the random tree model with admixture (Sect. 3.2). Indeed, in accord with Sect. 3.2, no violations are observed for Properties 1 and 3 across all values of n and α, and the number of violations of Property 2 shows a similar dependence on α to that seen in Sect. 3.2. When the noise is small ($\sigma < 0.1$), its effect is negligible, and all three properties behave similarly to those in the random tree model with admixture. In particular, the fraction of violations of Properties 1 and 3 is at most 0.1%, with no systematic dependence on n and α.

As σ increases, more noise is added to the additive source distance matrix, and thus, $\tilde{\mathbf{S}}^{(n-1)}$ deviates further from additivity. With sufficiently large σ, the noise contributes more to a distance entry than does the underlying additive matrix, and results are similar to those of the random matrix model with admixture (Sect. 3.1). For $n = 5$, the maximum failure rates for Properties 1 and 3 are 0.8% and 0.8%, respectively, both observed at $(\sigma, \alpha) = (0.97, 0.02)$; for $n = 16$, they are 3.8%, at $(\sigma, \alpha) = (0.96, 0.45)$, and 3.4%, at $(\sigma, \alpha) = (0.90, 0.16)$, respectively. These values accord with the corresponding values from the random matrix model with admixture: 0.8% at $\alpha = 0.97$ and 0.8% at $\alpha = 0.97$ for $n = 5$, and 3.9% at $\alpha = 0.49$ and 2.8% at $\alpha = 0.12$ for $n = 16$. As was seen in previous sections, the behavior of all three properties is symmetric around $\alpha = 0.5$, as the roles of taxa t_1 and t_2 are interchangeable.

In general, the numbers of violations for all three properties increase with n. For a given n, the total numbers of violations of Properties 1 and 3 summed over all admixture fraction values increase with σ. For a given set of parameter values (n, σ, α),

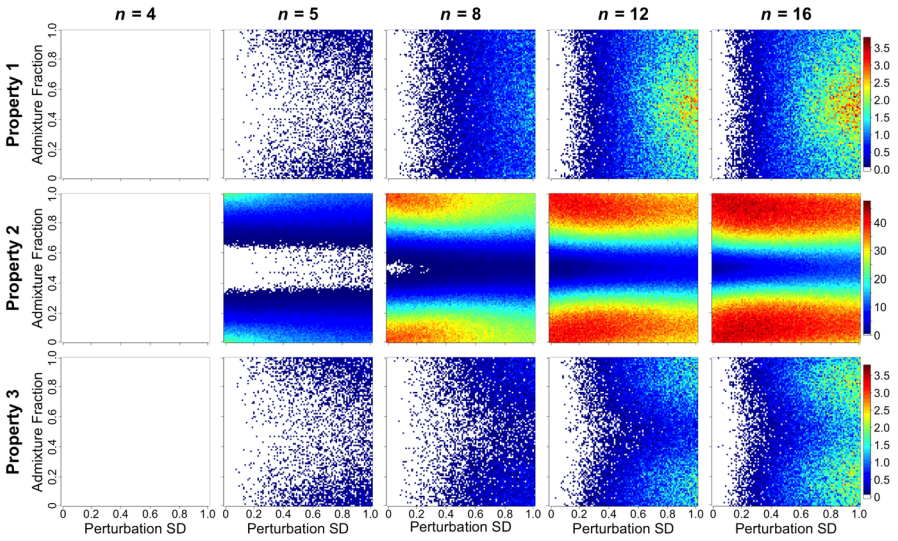

Fig. 9 Percentage of simulated matrices in which Properties 1, 2, and 3 are violated under a model in which a truncated normal noise term is added to each entry in additive source distance matrices $\mathbf{S}^{(n-1)}$, and admixed nth taxon is created using Eq. 5. The standard deviation σ of the noise term (prior to truncation) was varied from 0.00 to 1.00 at increment 0.01, and the admixture fraction (α) was varied from 0.01 to 0.99 at increment 0.01. Each point in the plot represents a percentage among 1000 simulated matrices for a pair (σ, α) and is colored according to the scale at right

the failure rate for Property 2 substantially exceeds those of Properties 1 and 3, and the number of Property 2 violations depends more on n and α than on the noise σ. We can compare the dependence on σ of the number of violations for a given n across properties by summing the total number of violations over all α and normalizing the difference between the maximal and minimal numbers of violations by the maximum. As σ varies from 0 to 1, the number of Property 2 violations changes by 38.2% when $n = 5$ and 9.9% when $n = 16$. By comparison, Properties 1 and 3 both change by 100%.

4 Discussion

In this paper, we have analyzed the effect of an admixed taxon on neighbor-joining inference of evolutionary trees. We have focused on three specific characteristics of the inferred trees that concern an admixed taxon and its two source taxa: antecedence of clustering, intermediacy of distances, and intermediacy of path lengths. Compared to a null model containing no admixture, the three properties were satisfied more frequently when pairwise distances involving an admixed taxon were included in the distance matrix (Eq. 6). The fact that violations were sometimes observed, however, shows that the conjectures of Kopelman et al. (2013) are not true in general, failing on some subsets of distance matrices.

In the case of additive source distance matrices $\mathbf{S}^{(n-1)}$, from which admixed distance matrices $\mathbf{D}^{(n)}$ were constructed, we observed no violations of Properties 1 and 3.

Additive source distance matrices also generated fewer violations of Property 2 than did either random source distance matrices without the additivity constraint on $\mathbf{S}^{(n-1)}$ or additive source distance matrices that were modified by perturbation terms. These results are sensible in light of the result of Kopelman et al. (2013) that no violations of Properties 2 and 3 occur when $\mathbf{D}^{(n)}$, the distance matrix including the admixed taxon, is additive—a result that pertains to a more stringent setting than those considered in our simulations.

For the 5-taxon case including an admixed taxon, we found that certain tree topologies are not accessible as a result of neighbor-joining. In a null model, all 15 labeled topologies of a 5-taxon unrooted labeled binary tree can be obtained. When we introduce the linear admixture relationship between the admixed taxon and the source taxa, however, only 12 and 8 of 15 topologies are accessible when the 4-taxon source distance matrices are drawn from random matrices and random additive trees, respectively. Note that although the mathematical results for the five-taxon model in Propositions 1–3 and Corollaries 1 and 2 have been motivated by observations from the simulations in the random source matrix model with admixture, they hold for all source distance matrices with four taxa.

4.1 Rogue Taxon Effect

The admixed taxon can be considered as a "rogue taxon" (Sanderson and Shaffer 2002; Thomson and Shaffer 2010; Cueto and Matsen 2011; Westover et al. 2013). The presence of an admixed population in the distance matrix perturbs the topology and branch lengths of the reconstructed tree compared to the case of exclusively non-admixed populations. To examine the admixed taxon's rogue effect, we restrict an NJ tree $\mathcal{T}_D^{(n)}$ with an admixed taxon to the $n-1$ non-admixed taxa, removing leaf t_n from $\mathcal{T}_D^{(n)}$ to construct a pruned tree $\widetilde{\mathcal{T}}_D^{(n-1)}$. We can compare the topology of $\widetilde{\mathcal{T}}_D^{(n-1)}$ to the topology $\mathcal{T}_S^{(n-1)}$ of the source NJ tree constructed from the source distance matrix $\mathbf{S}^{(n-1)}$.

In the null model in which distances involving the admixed taxon are chosen randomly, many pairs $(\widetilde{\mathcal{T}}_D^{(n-1)}, \mathcal{T}_S^{(n-1)})$ have different topologies (Fig. 10). When the distances to the admixed taxon are chosen by the linear mixture model, however, fewer disruptions of evolutionary relationships among the non-admixed $n-1$ taxa are observed. Finally, compared to the random source matrix model with admixture, in which the distances between the $n-1$ taxa are chosen randomly, enforcing the additivity constraint to the source distance matrix $\mathbf{S}^{(n-1)}$ greatly reduces the frequency of the rogue taxon effect.

For all three models, the rogue taxon effect increases in frequency as the number of taxa increases (Fig. 10a). In the null model, in which the additional taxon is chosen arbitrarily in a matrix with random distances, 18.0% of $\widetilde{\mathcal{T}}_D^{(n-1)}$ trees do not contain the original source tree when $n = 5$, growing to 99.9% for $n = 50$. In the random source matrix model, the corresponding value at $n = 50$ is 92.7%, and in the additive source matrix model, it is a comparatively low value of 37.4%.

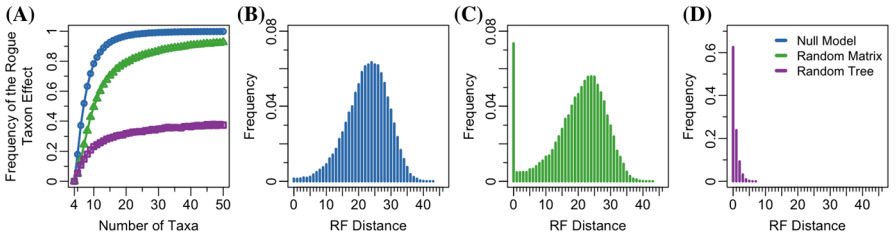

Fig. 10 Frequency and magnitude of the rogue taxon effect. **a** Simulated probabilities that the addition of an admixed taxon disturbs evolutionary relationships between $n-1$ non-admixed taxa in an inferred NJ tree. **b**, **c**, **d** Distribution across simulated trees with $n = 50$ of the Robinson–Foulds distance between the NJ tree of the 49 non-admixed taxa with and without inclusion of the admixed taxon. The maximal Robinson–Foulds distance is 46. An NJ tree is classified as having experienced the rogue taxon effect if the relationships in an n-taxon tree of the $n-1$ non-admixed taxa differ in topology from the tree constructed from the $n-1$ taxa without the admixed taxon included (nonzero Robinson–Foulds distance). **b** Null model with no admixture (Sect. 2.2). **c** Admixed distance matrix $\mathbf{D}^{(n)}$ constructed from random source matrix $\mathbf{S}^{(n-1)}$ (Sect. 3.1). **d** Admixed distance matrix $\mathbf{D}^{(n)}$ constructed from additive source matrix $\mathbf{S}^{(n-1)}$ (Sect. 3.2). The simulations shown are the same ones considered for the null model in Fig. 3, the random matrix model in Fig. 4, and the random tree model in Figure 6

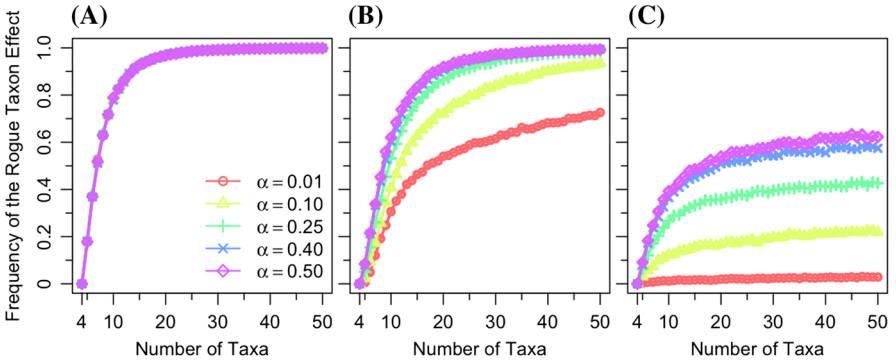

Fig. 11 Influence of the admixture fraction on the rogue taxon effect, the probability that the pruned tree topology $\widetilde{\mathcal{T}}_D^{(n-1)}$ differs from the source tree topology $\mathcal{T}_S^{(n-1)}$. For each admixture fraction in $\{0.01, 0.1, 0.25, 0.4, 0.5\}$ and each number of taxa in $\{4, 5, \ldots, 50\}$, 10,000 input admixed distance matrices $\mathbf{D}^{(n)}$ are considered for each of three different models. **a** Null model (Sect. 2.2). **b** Random source matrix model with admixture (Sect. 3.1). **c** Random additive source matrix model with admixture (Sect. 3.2)

The frequency of the rogue taxon effect for an admixed taxon depends on the admixture fraction and the number of taxa (Fig. 11). For a given n, considering $\alpha \leq 0.5$, the rogue taxon effect is rarest when α is small and most frequent when $\alpha = 0.5$. This result is sensible because when the admixture fraction from source taxon t_1 is small, the admixed taxon is closer to source taxon t_2, so the probability is high that the admixed taxon forms a cherry with t_2 without disturbing the topology of other taxa in the original tree. When $\alpha = 0.5$ and contributions from the two source taxa to the admixed taxon are equal, NJ inference is unstable because the admixed taxon can join a clade containing either source taxon, potentially producing greater influence on other steps in the inference. Thus, the rogue taxon effect occurs more often.

To measure the extent to which the rogue taxon effect alters the topology of source NJ trees, we quantify the topological difference between $\widetilde{\mathcal{T}}_D^{(n-1)}$ and $\mathcal{T}_S^{(n-1)}$ using the Robinson–Foulds (RF) distance (Robinson and Foulds 1981; Steel 2016), which counts bipartitions of $\widetilde{\mathcal{T}}_D^{(n-1)}$ that are absent in $\mathcal{T}_S^{(n-1)}$, and has maximal value three fewer than the number of taxa in the tree, or $n-4$. In the null model, the mean observed RF distance is 0.18 when $n = 5$ (maximum 1), increasing to 22.65 with $n = 50$ (maximum 43, Fig. 10b). In the random source matrix model, the mean RF distance at $n = 50$ is 19.83 (maximum 43, Fig. 10c); the additive source matrix model has smaller RF distance at $n = 50$, reaching mean only 0.57 and maximum 7 (Fig. 10d).

The rogue taxon effect with the admixed taxon under NJ parallels a corresponding analysis that investigated the effect of a rogue taxon on a tree inferred by balanced minimum evolution (BME) (Cueto and Matsen 2011). Because the NJ and BME methods are closely related (Gascuel and Steel 2006), we expect the behavior of the rogue taxon on NJ trees to be similar to the effect of a rogue taxon on BME trees. First, the high frequency and magnitude of the rogue taxon effect in the null model accord with Cueto and Matsen (2011), in which a frequent and severe rogue taxon effect was observed in simulations using treelike distance matrices augmented by distances to a rogue taxon chosen at random from a specified distribution.

Second, they obtained mathematical results that in BME trees, a rogue taxon can have significant, but not arbitrary effects, so that some topologies cannot be the inferred BME tree when the rogue taxon is present. In the simplest case, when $n = 5$, if the source BME tree is $((t_1, t_2), (t_3, t_4))$ and t_5 is added, then the BME tree cannot be $((t_2, t_4), t_5, (t_1, t_3))$ or $((t_2, t_3), t_5, (t_1, t_4))$ (Theorem 3.5 in Cueto and Matsen (2011)). Our NJ simulations with $n = 5$ in Fig. 10a, when sorted by the topologies of the source tree with and without the admixed taxon, suggest that the same rule might apply in the random source matrix and additive source matrix models, which do not produce instances of these cases (notice also that in Figures S1–S3, no arrows connect the associated trees in the random source matrix model). In the null model, however, the NJ simulations do infer the two topologies forbidden by BME, each with a nonzero but small frequency ($\sim$0.01%). This result can be attributed to the fact that NJ does not always infer the same topology as BME (Eickmeyer et al. 2008).

4.2 Tree Models

We note that our various analyses make use of a variety of probability distributions on trees, so that differences in results observed across Sects. 2.1, 2.2, 3.1, 3.2, and 3.3 could be partially attributable to differences in the probability model. Our exact computations (Sect. 2.1) use a uniform distribution on the set of labeled unrooted topologies with n taxa, a model under which we can obtain the exact probability that Property 3 holds by use of a recursion. As it is not straightforward to draw random distance matrices so that when NJ is applied, the simulated matrices produce a uniform distribution on labeled unrooted topologies, our null model simulations (Sect. 2.2) and the admixture simulations that rely on random matrices (Sect. 3.1) instead consider draws taken uniformly at random from a space $[0, 1]^{\binom{n}{2}}$ of distance matrices for n taxa. The null

model simulations only approximately reflect the exact computations, as they do not simulate from the same model under which the exact computations were performed.

Further, in another difference from the null model, our simulations with random additive source matrices (Sect. 3.2) and the associated simulations with perturbed additive source matrices (Sect. 3.3) begin from trees simulated by a Yule birth process. Rather than a uniform distribution, such trees follow the Yule–Harding distribution on the space of labeled rooted topologies, a distribution we chose because it is more suitable as a generative evolutionary model than is the uniform model (Blum and François 2006; Steel 2016).

The difference in tree models across our analyses likely contributes to some of the difference seen across analyses in the frequency of violations of the three properties. Notably, however, the models have similar trends of increasing numbers of violations of the properties when the number of taxa is increased. In addition, the model with random additive source matrices has the weakest rogue taxon effect, substantially smaller than in the null model and in the random source matrix model (Fig. 10). This observation accords with the theoretical result that a distance matrix that is fully additive when the admixed taxon is included does not have a rogue taxon effect—trivially, because NJ recovers the input tree from an additive distance matrix—and the random additive source matrix model is the closest of the three models to this idealized rogue-free scenario.

To further study the effect of the model choice on the results, it would be useful to extend the theoretical calculations of Sect. 2.1 to consider the Yule–Harding model, which does not produce the convenient recursive Eq. 1; to analyze simulations analogous to those of Sect. 3.2 under the less naturally generative uniform model on labeled tree shapes; and to devise a method for simulating random distance matrices that produce a uniform distribution on the set of labeled unrooted topologies after NJ is applied.

5 Conclusions

This work expands the understanding of the effect of admixture on neighbor-joining trees. Past empirical NJ trees with admixture and the conjectures of Kopelman et al. (2013) have suggested that observations of antecedence of clustering, intermediacy of distances, and intermediacy of path lengths might be taken as evidence of admixture. We have found that the three properties are not necessarily observed in our simulations with admixture models; it is therefore possible that counterexamples in data might also be found. However, the properties do occur more often in a random additive source matrix model with admixture between a pair of taxa than in null models without admixture; it is thus useful to continue to treat them as admixture related and to consider admixture as a potential explanation when the properties are observed.

As noted by Kopelman et al. (2013), a limitation of our approach is that common distance functions in population genetics need not be linear in the admixture coefficient α when applied to admixed populations (Boca and Rosenberg 2011), so that Eq. 5 will not necessarily apply to specific distance functions. In future work, it will be informative to consider a form of the admixture model in which the linear combination

is applied to the allele frequencies that underlie genetic distance computations and specific distance functions are then computed, rather than treating the distance function itself as linear.

Acknowledgements This work has been supported by National Institutes of Health grant R01 GM117590, by National Science Foundation grant BCS-1515127, and by a 2014 Rita Levi Montalcini grant from the Ministero dell'Istruzione, dell'Università e della Ricerca.

Appendix

A The Q-Criterion

For each agglomerative step of the NJ algorithm, the algorithm evaluates the *Q-criterion* (Bryant 2005; Gascuel and Steel 2006) and picks a pair of taxa i and j with the minimal q value, $q_{ij} = (n-2)x_{ij} - \sum_{k=1}^{n} x_{ik} - \sum_{k=1}^{n} x_{kj}$. Here, $i \neq j$ and i, j range from 1 to n. The Q-criterion gives a linear transformation of the original input distance matrix $\mathbf{D}^{(n)}$.

Capitalizing on the linearity to make use of matrix algebra (Eickmeyer and Yoshida 2008), we write $\mathbf{Q} = \mathbf{A}^{(n)}\mathbf{x}^{(n)}$. In the matrix $\mathbf{A}^{(n)}$, a and b represent taxon pairs, ranging from 1 to $\binom{n}{2}$, and r, ℓ, s, and t represent taxa.

$$A_{ab}^{(n)} = A_{r\ell,st}^{(n)} = \begin{cases} n-4 & \text{if } a = b \\ -1 & \text{if } a \neq b \text{ and } \{r, \ell\} \cap \{s, t\} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

For example, for $n = 4$, annotating rows and columns of $\mathbf{A}^{(4)}$ by the order of the entries in $\mathbf{x}^{(4)}$,

$$\mathbf{A}^{(4)} = \begin{array}{c} \\ 21 \\ 31 \\ 32 \\ 41 \\ 42 \\ 43 \end{array} \begin{array}{c} \begin{array}{cccccc} 21 & 31 & 32 & 41 & 42 & 43 \end{array} \\ \begin{bmatrix} 0 & -1 & -1 & -1 & -1 & 0 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ 0 & -1 & -1 & -1 & -1 & 0 \end{bmatrix} \end{array}, \quad \mathbf{x}^{(4)} = \begin{bmatrix} x_{21} \\ x_{31} \\ x_{32} \\ x_{41} \\ x_{42} \\ x_{43} \end{bmatrix}. \tag{9}$$

The process of cherry picking continues until only three nodes are left, at which point the three remaining taxon groups are joined to a shared node. NJ produces a particular tree topology given an input distance matrix $\mathbf{D}^{(n)}$ if and only if the original distances x_{ij} satisfy an associated set of linear inequalities defined by the Q-criterion.

B Correction to Section 4.1 of Kopelman et al. (2013)

In the case of an additive n-taxon matrix $\mathbf{D}^{(n)}$ with taxon t_n admixed between source taxa t_1 and t_2, Section 4.1 of Kopelman et al. (2013) demonstrates that in the NJ tree of all n taxa, (1) taxa t_1, t_2, and t_n must be collinear, with t_n between t_1 and t_2. Moreover,

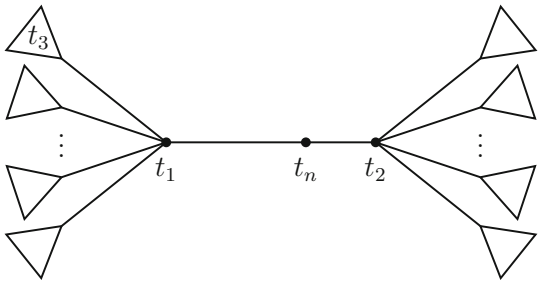

Fig. 12 The structure required for a tree corresponding to an additive n-taxon admixed distance matrix whose distances satisfy Eq. 5. Taxa t_1 and t_2 are source taxa, and t_n is the admixed taxon. Triangles represent subtrees, one of which contains taxon t_3 either at an internal node or as a leaf node. Admixed taxon t_n must lie on the path connecting the source taxa, with no other node placed on edges $t_1 - t_n$ and $t_2 - t_n$. The remaining taxa $t_3, t_4, \ldots, t_{n-1}$ can be in any subtrees connected to either t_1 or t_2 by edges external to the path $t_1 - t_2$

they showed that (2) each taxon t_i for $i = 3, 4, \ldots, n-1$ must be collinear with t_1, t_2, and t_n, with taxon t_i exterior to the path from t_1 to t_n to t_2. Finally, they claimed that (3) the NJ tree structure can be characterized by a line from t_1 to t_n to t_2, with t_1 and t_2 each placed at a multifurcating node to one of which each taxon $t_3, t_4, \ldots, t_{n-1}$ must be connected by a single edge (depicted in their Fig. 3b).

We comment here that although points (1) and (2) are correct, the claim (3) does not follow from (1) and (2) as assumed by Kopelman et al. (2013). It is possible to place the various taxa t_i for $i = 3, 4, \ldots, n-1$ in relation to the line segment $t_1 - t_n - t_2$ in a manner in which each t_i is collinear with and exterior to the segment, but the nodes for t_1 and t_2 are not necessarily multifurcating, and the t_i are not necessarily connected to one of those nodes by only a single edge. It follows from (1) and (2) merely that exterior to the segment $t_1 - t_n - t_2$ are two possibly but not necessarily multifurcating trees, one rooted at t_1 and the other rooted at t_2, and that each taxon t_i for $i = 3, 4, \ldots, n-1$ is placed in one of those trees (Fig. 12).

C Proofs for the 5-Taxon Case

In this appendix, we provide the proofs of Propositions 1, 2, and 3 stated in Sect. 3.1 pertaining to distance matrices with admixture $\mathbf{D}^{(5)}$ constructed from random source matrices $\mathbf{S}^{(4)}$.

From the linear mixture model in Eq. 5, the distance matrix for $n = 5$ taxa, including one admixed taxon t_5 and two source taxa t_1 and t_2, is:

$$\mathbf{D}^{(5)} = \begin{bmatrix} 0 & x_{21} & x_{31} & x_{41} & (1-\alpha)x_{21} \\ x_{21} & 0 & x_{32} & x_{42} & \alpha x_{21} \\ x_{31} & x_{32} & 0 & x_{43} & \alpha x_{31} + (1-\alpha)x_{32} \\ x_{41} & x_{42} & x_{43} & 0 & \alpha x_{41} + (1-\alpha)x_{42} \\ (1-\alpha)x_{21} & \alpha x_{21} & \alpha x_{31} + (1-\alpha)x_{32} & \alpha x_{41} + (1-\alpha)x_{42} & 0 \end{bmatrix},$$

with positive non-diagonal entries ($x_{ij} > 0$ for all i and j with $i \neq j$).

Using the matrix representation in Eq. 8, the Q-criterion in the first step in NJ with 5 taxa can be written as $\mathbf{Q} = \mathbf{A}^{(5)}\mathbf{x}^{(5)}$, where $\mathbf{A}^{(5)} \in \mathbb{R}^{10\times 10}$, $\mathbf{x}^{(5)} \in \mathbb{R}^{10\times 1}$. We have

$$\mathbf{A}^{(5)} = \begin{array}{c} \\ 21 \\ 31 \\ 32 \\ 41 \\ 42 \\ 43 \\ 51 \\ 52 \\ 53 \\ 54 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 21 & 31 & 32 & 41 & 42 & 43 & 51 & 52 & 53 & 54 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & -1 & -1 & 0 & -1 & -1 & 0 & 0 \\ -1 & 1 & -1 & -1 & 0 & -1 & -1 & 0 & -1 & 0 \\ -1 & -1 & 1 & 0 & -1 & -1 & 0 & -1 & -1 & 0 \\ -1 & -1 & 0 & 1 & -1 & -1 & -1 & 0 & 0 & -1 \\ -1 & 0 & -1 & -1 & 1 & -1 & 0 & -1 & 0 & -1 \\ 0 & -1 & -1 & -1 & -1 & 1 & 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & -1 & 0 & 0 & 1 & -1 & -1 & -1 \\ -1 & 0 & -1 & 0 & -1 & 0 & -1 & 1 & -1 & -1 \\ 0 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix} \end{array}$$

$$\equiv \begin{bmatrix} \mathbf{a}_{\mathbf{1}}^{(4)} \\ \mathbf{a}_{\mathbf{2}}^{(4)} \\ \mathbf{a}_{\mathbf{3}}^{(4)} \\ \mathbf{a}_{\mathbf{4}}^{(4)} \\ \mathbf{a}_{\mathbf{5}}^{(4)} \\ \mathbf{a}_{\mathbf{6}}^{(4)} \\ \mathbf{a}_{\mathbf{7}}^{(4)} \\ \mathbf{a}_{\mathbf{8}}^{(4)} \\ \mathbf{a}_{\mathbf{9}}^{(4)} \\ \mathbf{a}_{\mathbf{10}}^{(4)} \end{bmatrix} \tag{10}$$

$$\mathbf{x}^{(5)} = \begin{bmatrix} x_{21} \\ x_{31} \\ x_{32} \\ x_{41} \\ x_{42} \\ x_{43} \\ x_{51} \\ x_{52} \\ x_{53} \\ x_{54} \end{bmatrix} = \begin{bmatrix} x_{21} \\ x_{31} \\ x_{32} \\ x_{41} \\ x_{42} \\ x_{43} \\ (1-\alpha)x_{21} \\ \alpha x_{21} \\ \alpha x_{31} + (1-\alpha)x_{32} \\ \alpha x_{41} + (1-\alpha)x_{42} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1-\alpha & 0 & 0 & 0 & 0 & 0 \\ \alpha & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha & 1-\alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha & 1-\alpha & 0 \end{bmatrix} \begin{bmatrix} x_{21} \\ x_{31} \\ x_{32} \\ x_{41} \\ x_{42} \\ x_{43} \end{bmatrix} \equiv \mathbf{M}\mathbf{s}^{(4)}, \tag{11}$$

denoting by $\mathbf{M}$ the matrix in Eq. 11, with $\mathbf{s}^{(4)} = [x_{21}, x_{31}, x_{32}, x_{41}, x_{42}, x_{43}]^\intercal$.

Only six independent variables appear in the source distance vector $\mathbf{s}^{(4)}$, and we can rewrite the Q-criterion as

$$\begin{aligned} \mathbf{Q}^{(5)} &= [q_{21}, q_{31}, q_{32}, q_{41}, q_{42}, q_{43}, q_{51}, q_{52}, q_{53}, q_{54}]^\intercal \\ &= \mathbf{A}^{(5)}\mathbf{x}^{(5)} = \mathbf{A}^{(5)}\mathbf{M}\mathbf{s}^{(4)} = \tilde{\mathbf{A}}^{(5)}\mathbf{s}^{(4)}, \end{aligned} \tag{12}$$

where

$$\tilde{\mathbf{A}}^{(5)} \equiv \mathbf{A}^{(5)}\mathbf{M}$$

$$= \begin{array}{c} 21 \\ 31 \\ 32 \\ 41 \\ 42 \\ 43 \\ 51 \\ 52 \\ 53 \\ 54 \end{array} \begin{bmatrix} 0 & -1 & -1 & -1 & -1 & 0 \\ -2+\alpha & 1-\alpha & -2+\alpha & -1 & 0 & -1 \\ -1-\alpha & -1-\alpha & \alpha & 0 & -1 & -1 \\ -2+\alpha & -1 & 0 & 1-\alpha & -2+\alpha & -1 \\ -1-\alpha & 0 & -1 & -1-\alpha & \alpha & -1 \\ 0 & -1-\alpha & -2+\alpha & -1-\alpha & -2+\alpha & 1 \\ -2\alpha & -1-\alpha & -1+\alpha & -1-\alpha & -1+\alpha & 0 \\ -2+2\alpha & -\alpha & -2+\alpha & -\alpha & -2+\alpha & 0 \\ -1 & -1+\alpha & -\alpha & -\alpha & -1+\alpha & -1 \\ -1 & -\alpha & -1+\alpha & -1+\alpha & -\alpha & -1 \end{bmatrix} \equiv \begin{bmatrix} \tilde{\mathbf{a}}_1^{(5)} \\ \tilde{\mathbf{a}}_2^{(5)} \\ \tilde{\mathbf{a}}_3^{(5)} \\ \tilde{\mathbf{a}}_4^{(5)} \\ \tilde{\mathbf{a}}_5^{(5)} \\ \tilde{\mathbf{a}}_6^{(5)} \\ \tilde{\mathbf{a}}_7^{(5)} \\ \tilde{\mathbf{a}}_8^{(5)} \\ \tilde{\mathbf{a}}_9^{(5)} \\ \tilde{\mathbf{a}}_{10}^{(5)} \end{bmatrix}, \tag{13}$$

and $\tilde{\mathbf{a}}_{\boldsymbol{i}}^{(5)}$ $(i = 1, \ldots, 10)$ is the ith row vector in $\tilde{\mathbf{A}}^{(5)}$. The pair that is agglomerated together is the pair corresponding to the row with the minimal value in $\mathbf{Q}^{(5)}$.

Proof of Proposition 1 We must show that (t_1, t_2), (t_3, t_5), and (t_4, t_5) cannot be the minimum for the Q-criterion in the first step of NJ. We begin with (t_1, t_2). By contradiction, suppose that (t_1, t_2) has the minimum Q-criterion. This means that $q_{21} = \tilde{\mathbf{a}}_1^{(5)} \cdot \mathbf{s}^{(4)}$ is less than or equal to $q_{\ell m} = \tilde{\mathbf{a}}_{\boldsymbol{i}}^{(5)} \cdot \mathbf{s}^{(4)}$ for all $i = 2, 3, \ldots, 10$. In other words,

$$(\tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_{\boldsymbol{i}}^{(5)}) \cdot \mathbf{s}^{(4)} \leq 0. \tag{14}$$

This linear system of 9 inequalities in the 6 entries of $\mathbf{s}^{(4)}$ can be represented as a matrix $\mathbf{B}^{(5)}\mathbf{s}^{(4)} \leq \mathbf{0}$, where $\mathbf{0} = [0, \ldots, 0]^\intercal \in \mathbb{R}^{9\times 1}$ and the ith row of the 9×6 matrix $\mathbf{B}^{(5)}$ is $\tilde{\mathbf{a}}_1 - \tilde{\mathbf{a}}_{\boldsymbol{i+1}}$:

$$\mathbf{B}^{(5)} \equiv \begin{bmatrix} \tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_2^{(5)} \\ \tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_3^{(5)} \\ \tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_4^{(5)} \\ \tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_5^{(5)} \\ \tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_6^{(5)} \\ \tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_7^{(5)} \\ \tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_8^{(5)} \\ \tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_9^{(5)} \\ \tilde{\mathbf{a}}_1^{(5)} - \tilde{\mathbf{a}}_{10}^{(5)} \end{bmatrix} = \begin{array}{c} \\ 2-\alpha \\ 1+\alpha \\ 2-\alpha \\ 1+\alpha \\ 0 \\ 2\alpha \\ 2-2\alpha \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{cccccc} 21 & 31 & 32 & 41 & 42 & 43 \end{array} \\ \begin{bmatrix} -2+\alpha & 1-\alpha & 0 & -1 & 1 \\ \alpha & -1-\alpha & -1 & 0 & 1 \\ 0 & -1 & -2+\alpha & 1-\alpha & 1 \\ -1 & 0 & \alpha & -1-\alpha & 1 \\ \alpha & 1-\alpha & \alpha & 1-\alpha & -1 \\ \alpha & -\alpha & \alpha & -\alpha & 0 \\ -1+\alpha & 1-\alpha & -1+\alpha & 1-\alpha & 0 \\ -\alpha & -1+\alpha & -1+\alpha & -\alpha & 1 \\ -1+\alpha & -\alpha & -\alpha & -1+\alpha & 1 \end{bmatrix} \end{array}.$$

We use Fourier–Motzkin elimination (FME, Dantzig and Eaves 1973; Schrijver 1986; Ziegler 1995) to prove that this system of linear inequalities $\mathbf{B}^{(5)}\mathbf{s}^{(4)} \leq \mathbf{0}$ with the constraint $\mathbf{s}^{(4)} > \mathbf{0}$ has no solution, thereby showing that (t_1, t_2) cannot have the lowest Q-criterion and hence cannot be the first pair to agglomerate in the NJ algorithm.

In FME, we sequentially eliminate variables from the system of linear inequalities, at each step transforming the system into a new system with an equivalent solution space. Eventually, we eliminate all variables and reach a set of trivially satisfied inequalities only involving constants, so that the system has solutions. Alternatively, we reach an inequality with no solution, and the system therefore has no solution.

Briefly, consider a system of inequalities $\mathbf{Bx} \leq \mathbf{c}$, where $\mathbf{B} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^{n\times 1}$, and $\mathbf{c} \in \mathbb{R}^{m\times 1}$. Each row of $\mathbf{B}$ represents an inequality, and the solution space for the system is $\mathcal{P} = \{\mathbf{x} \mid \mathbf{Bx} \leq \mathbf{c}\} \subset \mathbb{R}^n$. FME eliminates the variable x_k by performing row operations on the augmented matrix $(\mathbf{B} \mid \mathbf{c})$. In Step 1, we rearrange the inequalities in $\mathbf{Bx} \leq \mathbf{c}$ into three sets, based on the signs of the kth column: $I_+ = \{i \in \{1, 2, \ldots, m\} \mid b_{ik} > 0\}$, $I_- = \{i \in \{1, 2, \ldots, m\} \mid b_{ik} < 0\}$, and $I_0 = \{i \in \{1, 2, \ldots, m\} \mid b_{ik} = 0\}$.

In Step 2, for each $i \in I_+ \cup I_-$, we scale each row $(b_{i1}, \ldots, b_{in} \mid c_i)$ by $|b_{ik}|$ so that elements in column k of $(\mathbf{B} \mid \mathbf{c})$ are either 1, -1, or 0. In Step 3, we then define a new set of inequalities:

$$\frac{1}{b_{rk}}\sum_{i=1}^{n} b_{ri}x_i + \frac{1}{|b_{sk}|}\sum_{i=1}^{n} b_{si}x_i \leq \frac{1}{b_{rk}}c_r + \frac{1}{|b_{sk}|}c_s \qquad (r, s) \in I_+ \times I_-$$

$$\sum_{i=1}^{n} b_{\ell i}x_i \leq c_\ell \qquad \ell \in I_0.$$

Because the coefficient of x_k is normalized to $+1$ for row r and -1 for row s, the combined new set of inequalities does not have x_k. In other words, column k has zeroes as entries. We now have a total of $|I_+||I_-| + |I_0|$ new inequalities in $x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n$. The system of inequalities has solutions if and only if FME can successively eliminate all the variables without generating a contradiction. Otherwise, if the system is inconsistent at one point during the elimination, then it has no solutions.

We first seek to eliminate the third variable, x_{32}, represented in the third column:

$$\mathbf{B}^{(5)} = \begin{array}{c} \begin{array}{cccccc} 21 & 31 & 32 & 41 & 42 & 43 \end{array} \\ \left[\begin{array}{cccccc} 2-\alpha & -2+\alpha & 1-\alpha & 0 & -1 & 1 \\ 1+\alpha & \alpha & -1-\alpha & -1 & 0 & 1 \\ 2-\alpha & 0 & -1 & -2+\alpha & 1-\alpha & 1 \\ 1+\alpha & -1 & 0 & \alpha & -1-\alpha & 1 \\ 0 & \alpha & 1-\alpha & \alpha & 1-\alpha & -1 \\ 2\alpha & \alpha & -\alpha & \alpha & -\alpha & 0 \\ 2-2\alpha & -1+\alpha & 1-\alpha & -1+\alpha & 1-\alpha & 0 \\ 1 & -\alpha & -1+\alpha & -1+\alpha & -\alpha & 1 \\ 1 & -1+\alpha & -\alpha & -\alpha & -1+\alpha & 1 \end{array}\right] \end{array}.$$

Note that $0 < \alpha < 1$ and $0 < 1 - \alpha < 1$. The rearrangement by the sign of x_{32} gives:

$$\mathbf{B}^{(5)}_{\text{step1}} = \begin{array}{c} \begin{array}{cccccc} 21 & 31 & 32 & 41 & 42 & 43 \end{array} \\ \begin{bmatrix} 2-\alpha & -2+\alpha & 1-\alpha & 0 & -1 & 1 \\ 0 & \alpha & 1-\alpha & \alpha & 1-\alpha & -1 \\ 2-2\alpha & -1+\alpha & 1-\alpha & -1+\alpha & 1-\alpha & 0 \\ 1+\alpha & \alpha & -1-\alpha & -1 & 0 & 1 \\ 2-\alpha & 0 & -1 & -2+\alpha & 1-\alpha & 1 \\ 2\alpha & \alpha & -\alpha & \alpha & -\alpha & 0 \\ 1 & -\alpha & -1+\alpha & -1+\alpha & -\alpha & 1 \\ 1 & -1+\alpha & -\alpha & -\alpha & -1+\alpha & 1 \\ 1+\alpha & -1 & 0 & \alpha & -1-\alpha & 1 \end{bmatrix} \end{array}.$$

Once the rows have been rearranged, we normalize each row with a nonzero entry for x_{32} by the absolute value of the entry for x_{32}:

$$\mathbf{B}^{(5)}_{\text{step2}} = \begin{array}{c} \begin{array}{cccccc} 21 & 31 & 32 & 41 & 42 & 43 \end{array} \\ \begin{bmatrix} \frac{2-\alpha}{1-\alpha} & -\frac{2-\alpha}{1-\alpha} & 1 & 0 & -\frac{1}{1-\alpha} & \frac{1}{1-\alpha} \\ 0 & \frac{\alpha}{1-\alpha} & 1 & \frac{\alpha}{1-\alpha} & 1 & -\frac{1}{1-\alpha} \\ 2 & -1 & 1 & -1 & 1 & 0 \\ 1 & \frac{\alpha}{1+\alpha} & -1 & -\frac{1}{1+\alpha} & 0 & \frac{1}{1+\alpha} \\ 2-\alpha & 0 & -1 & -2+\alpha & 1-\alpha & 1 \\ 2 & 1 & -1 & 1 & -1 & 0 \\ \frac{1}{1-\alpha} & -\frac{\alpha}{1-\alpha} & -1 & -1 & -\frac{\alpha}{1-\alpha} & \frac{1}{1-\alpha} \\ \frac{1}{\alpha} & -\frac{1-\alpha}{\alpha} & -1 & -1 & -\frac{1-\alpha}{\alpha} & \frac{1}{\alpha} \\ 1+\alpha & -1 & 0 & \alpha & -1-\alpha & 1 \end{bmatrix} \end{array}.$$

Considering linear combinations of rows in I_+ and I_-, we then have a new system of inequalities:

$$
\mathbf{B}^{(5)}_{\text{step3}} = \begin{array}{c} \begin{array}{cccccc} 21 & 31 & 32 & 41 & 42 & 43 \end{array} \\ \left[\begin{array}{cccccc}
\frac{-3+2\alpha}{-1+\alpha} & \frac{2}{-1+\alpha^2} & 0 & -\frac{1}{1+\alpha} & \frac{1}{-1+\alpha} & -\frac{2}{-1+\alpha^2} \\
-\frac{(-2+\alpha)^2}{-1+\alpha} & \frac{2-\alpha}{-1+\alpha} & 0 & -2+\alpha & -\frac{(-2+\alpha)\alpha}{-1+\alpha} & 1+\frac{1}{1-\alpha} \\
\frac{-4+3\alpha}{-1+\alpha} & \frac{1}{-1+\alpha} & 0 & 1 & \frac{2-\alpha}{-1+\alpha} & \frac{1}{1-\alpha} \\
\frac{-3+\alpha}{-1+\alpha} & \frac{2}{-1+\alpha} & 0 & -1 & \frac{1+\alpha}{-1+\alpha} & -\frac{2}{-1+\alpha} \\
\frac{-2+\alpha}{-1+\alpha}+\frac{1}{\alpha} & \frac{1}{(-1+\alpha)\alpha} & 0 & -1 & \frac{1-\alpha+\alpha^2}{(-1+\alpha)\alpha} & \frac{1}{\alpha-\alpha^2} \\
1 & -\frac{2\alpha}{-1+\alpha^2} & 0 & \frac{\alpha}{1-\alpha}-\frac{1}{1+\alpha} & 1 & \frac{2\alpha}{-1+\alpha^2} \\
2-\alpha & \frac{\alpha}{1-\alpha} & 0 & -2+\alpha+\frac{\alpha}{1-\alpha} & 2-\alpha & \frac{\alpha}{-1+\alpha} \\
2 & \frac{1}{1-\alpha} & 0 & \frac{1}{1-\alpha} & 0 & \frac{1}{-1+\alpha} \\
\frac{1}{1-\alpha} & 0 & 0 & \frac{1-2\alpha}{-1+\alpha} & \frac{-1+2\alpha}{-1+\alpha} & 0 \\
\frac{1}{\alpha} & \frac{1-2\alpha}{(-1+\alpha)\alpha} & 0 & \frac{1-2\alpha}{-1+\alpha} & 2-\frac{1}{\alpha} & \frac{1}{-1+\alpha}+\frac{1}{\alpha} \\
3 & -\frac{1}{1+\alpha} & 0 & -\frac{2+\alpha}{1+\alpha} & 1 & \frac{1}{1+\alpha} \\
4-\alpha & -1 & 0 & -3+\alpha & 2-\alpha & 1 \\
4 & 0 & 0 & 0 & 0 & 0 \\
\frac{-3+2\alpha}{-1+\alpha} & \frac{1}{-1+\alpha} & 0 & -2 & \frac{-1+2\alpha}{-1+\alpha} & \frac{1}{1-\alpha} \\
2+\frac{1}{\alpha} & -\frac{1}{\alpha} & 0 & -2 & 2-\frac{1}{\alpha} & \frac{1}{\alpha} \\
1+\alpha & -1 & 0 & \alpha & -1-\alpha & 1
\end{array}\right] \end{array}.
$$

The highlighted 13th row implies that if $\mathbf{B}^{(5)}\mathbf{s}^{(4)} \leq \mathbf{0}$, then $\mathbf{B}^{(5)}_{\text{step3}}\mathbf{s}^{(4)} \leq \mathbf{0}$. We would then have $4x_{21} \leq 0$. However, this inequality contradicts our assumption that all off-diagonal entries of distance matrices are positive, by which we must have $x_{21} > 0$.

Similarly, we can show that pairs (t_3, t_5) and (t_4, t_5) cannot have the minimum Q-value and thus do not form cherries in the first step of NJ. For (t_3, t_5), we start with a system of inequalities $(\tilde{\mathbf{a}}^{(5)}_9 - \tilde{\mathbf{a}}^{(5)}_i) \cdot \mathbf{s}^{(4)} \leq 0\,(i \neq 9)$ and eliminate variable x_{41} to reach the contradiction $x_{21} \leq 0$. For (t_4, t_5), we start with a system of inequalities $(\tilde{\mathbf{a}}^{(5)}_{10} - \tilde{\mathbf{a}}^{(5)}_i) \cdot \mathbf{s}^{(4)} \leq 0\,(i \neq 10)$ and eliminate variable x_{32} to reach the contradiction $x_{21} \leq 0$. □

Note that the fact that (t_1, t_2), (t_3, t_5), and (t_4, t_5) cannot form cherries in the first step does not imply that they cannot form cherries in the final NJ tree, as they can agglomerate in subsequent iterations of the NJ algorithm. In Proposition 2, however, we show that the final NJ tree of $\mathbf{D}^{(5)}$ cannot have (t_1, t_2) as a cherry at all.

Proof of Proposition 2 In Sect. 3.1, our simulations found that the three unrooted labeled binary tree topologies containing the source taxa t_1 and t_2 as a cherry are unattainable by NJ inference when distance matrices of the form of Eq. 6 are used. Such topologies are $((t_4, t_5), t_3, (t_1, t_2))$, $((t_3, t_5), t_4, (t_1, t_2))$, and $((t_3, t_4), t_5, (t_1, t_2))$.

Because (t_1, t_2), (t_3, t_5) and (t_4, t_5) cannot form a cherry in the first step of cherry picking (Proposition 1), we have so far shown that topologies $((t_4, t_5), t_3, (t_1, t_2))$ and $((t_3, t_5), t_4, (t_1, t_2))$ cannot be the final NJ tree. To show that (t_1, t_2) is not a cherry in the NJ tree and therefore to prove the proposition, it remains to show that $((t_3, t_4), t_5, (t_1, t_2))$, the last remaining topology containing (t_1, t_2) as a cherry, is not accessible.

Consider a case in which pair (t_3, t_4) agglomerates in the first iteration and we are left with four unclustered nodes, $\{t_1, t_2, t_5, t_6\}$. Here, t_6 represents a clade containing (t_3, t_4). Distances between the new node t_6 and the remaining taxa are:

$$x_{6i} = \frac{1}{2}(x_{3i} + x_{4i} - x_{43}),\ i \in \{1, 2, 5\}.$$

These distances can be written in matrix form as $\tilde{\mathbf{x}} = \mathbf{R}\mathbf{x}^{(5)}$, where

$$\tilde{\mathbf{x}} = \begin{bmatrix} x_{21} \\ x_{51} \\ x_{52} \\ x_{61} \\ x_{62} \\ x_{65} \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Using the matrix representation in Eq. 8, the Q-criterion for the four remaining nodes is

$$\begin{aligned} \mathbf{Q}^{(4)} &= [q_{21}, q_{51}, q_{52}, q_{61}, q_{62}, q_{65}]^{\mathsf{T}} \\ &= \mathbf{A}^{(4)}\tilde{\mathbf{x}} = \mathbf{A}^{(4)}\mathbf{R}\mathbf{x}^{(5)} = \mathbf{A}^{(4)}\mathbf{R}\mathbf{M}\mathbf{s}^{(4)} = \tilde{\mathbf{A}}^{(4)}\mathbf{s}^{(4)}, \end{aligned}$$

where

$$\mathbf{A}^{(4)} = \begin{array}{c} \\ 21 \\ 51 \\ 52 \\ 61 \\ 62 \\ 65 \end{array} \begin{array}{c} \begin{array}{cccccc} 21 & 51 & 52 & 61 & 62 & 65 \end{array} \\ \begin{bmatrix} 0 & -1 & -1 & -1 & -1 & 0 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ 0 & -1 & -1 & -1 & -1 & 0 \end{bmatrix} \end{array}$$

$$
\begin{aligned}
\tilde{\mathbf{A}}^{(4)} &\equiv \mathbf{A}^{(4)}\mathbf{RM} \\
&= \begin{array}{c} \\ 21 \\ 51 \\ 52 \\ 61 \\ 62 \\ 65 \end{array}
\begin{bmatrix}
-1 & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 1 \\
-(1+\alpha) & -\frac{1}{2}(1+\alpha) & -\frac{1}{2}(1-\alpha) & -\frac{1}{2}(1+\alpha) & -\frac{1}{2}(1-\alpha) & 1 \\
-(2-\alpha) & -\frac{1}{2}\alpha & -\frac{1}{2}(2-\alpha) & -\frac{1}{2}\alpha & -\frac{1}{2}(2-\alpha) & 1 \\
-(2-\alpha) & -\frac{1}{2}\alpha & -\frac{1}{2}(2-\alpha) & -\frac{1}{2}\alpha & -\frac{1}{2}(2-\alpha) & 1 \\
-(1+\alpha) & -\frac{1}{2}(1+\alpha) & -\frac{1}{2}(1-\alpha) & -\frac{1}{2}(1+\alpha) & -\frac{1}{2}(1-\alpha) & 1 \\
-1 & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 1
\end{bmatrix} \\
&= \begin{bmatrix}
\tilde{\mathbf{a}}_1^{(4)} \\ \tilde{\mathbf{a}}_2^{(4)} \\ \tilde{\mathbf{a}}_3^{(4)} \\ \tilde{\mathbf{a}}_4^{(4)} \\ \tilde{\mathbf{a}}_5^{(4)} \\ \tilde{\mathbf{a}}_6^{(4)}
\end{bmatrix}.
\end{aligned}
$$

As before, $\tilde{\mathbf{a}}_i^{(4)}$ $(i = 1, \ldots, 6)$ is defined as the ith row vector in $\tilde{\mathbf{A}}^{(4)}$. Because $\tilde{\mathbf{a}}_i^{(4)} = \tilde{\mathbf{a}}_{7-i}^{(4)}$ $(i = 1, 2, 3)$, pairs $\{(t_2, t_1), (t_6, t_5)\}$ have the same Q-criterion, as do the pairs $\{(t_5, t_1), (t_6, t_2)\}$ and $\{(t_5, t_2), (t_6, t_1)\}$. To prove that t_1 and t_2 cannot form a cherry, it suffices to show that the Q-criterion for (t_1, t_2), $q_{21} = \tilde{\mathbf{a}}_{\mathbf{1}}^{(4)} \cdot \mathbf{s}^{(4)}$, is not minimum, so that either $\tilde{\mathbf{a}}_2^{(4)} \cdot \mathbf{s}^{(4)} < \tilde{\mathbf{a}}_{\mathbf{1}}^{(4)} \cdot \mathbf{s}^{(4)}$ or $\tilde{\mathbf{a}}_3^{(4)} \cdot \mathbf{s}^{(4)} < \tilde{\mathbf{a}}_{\mathbf{1}}^{(4)} \cdot \mathbf{s}^{(4)}$ must hold. Equivalently, one of the following pair of inequalities must hold:

$$
\begin{aligned}
(\tilde{\mathbf{a}}_2^{(4)} - \tilde{\mathbf{a}}_{\mathbf{1}}^{(4)}) \cdot \mathbf{s}^{(4)} &< 0 \\
(\tilde{\mathbf{a}}_3^{(4)} - \tilde{\mathbf{a}}_{\mathbf{1}}^{(4)}) \cdot \mathbf{s}^{(4)} &< 0.
\end{aligned} \tag{15}
$$

By the assumption that (t_3, t_4) is the cherry from the first iteration of the NJ algorithm, the Q-criterion for (t_3, t_4), $q_{43} = \tilde{\mathbf{a}}_{\mathbf{6}}^{(5)} \cdot \mathbf{s}^{(4)}$, from the first iteration of the NJ algorithm must be minimum:

$$
(\tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_{\boldsymbol{i}}^{(5)}) \cdot \mathbf{s}^{(4)} \leq 0, \tag{16}
$$

for all i from 1 to 10 $(i \neq 6)$ and $\tilde{\mathbf{a}}_{\boldsymbol{i}}^{(5)}$ as defined in Eq. 13.

To show that Eq. 15 follows from Eq. 16, suppose to the contrary that (t_1, t_2) has the minimal Q-criterion in the second iteration of the NJ algorithm, so that both of the following hold:

$$
\begin{aligned}
(\tilde{\mathbf{a}}_1^{(4)} - \tilde{\mathbf{a}}_{\mathbf{2}}^{(4)}) \cdot \mathbf{s}^{(4)} &\leq 0 \\
(\tilde{\mathbf{a}}_1^{(4)} - \tilde{\mathbf{a}}_{\mathbf{3}}^{(4)}) \cdot \mathbf{s}^{(4)} &\leq 0.
\end{aligned} \tag{17}
$$

We then have a system of 11 linear inequalities that can be represented in matrix form, $\mathbf{B}^{(4)}\mathbf{s}^{(4)} \leq \mathbf{0}$, where

$$\mathbf{B}^{(4)} \equiv \begin{bmatrix} \tilde{\mathbf{a}}_1^{(4)} - \tilde{\mathbf{a}}_2^{(4)} \\ \tilde{\mathbf{a}}_1^{(4)} - \tilde{\mathbf{a}}_3^{(4)} \\ \tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_1^{(5)} \\ \tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_2^{(5)} \\ \tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_3^{(5)} \\ \tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_4^{(5)} \\ \tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_5^{(5)} \\ \tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_7^{(5)} \\ \tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_8^{(5)} \\ \tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_9^{(5)} \\ \tilde{\mathbf{a}}_6^{(5)} - \tilde{\mathbf{a}}_{10}^{(5)} \end{bmatrix}$$

$$= \begin{array}{c} \begin{matrix} 21 & 31 & 32 & 41 & 42 & 43 \end{matrix} \\ \begin{bmatrix} \alpha & \frac{1}{2}\alpha & -\frac{1}{2}\alpha & \frac{1}{2}\alpha & -\frac{1}{2}\alpha & 0 \\ 1-\alpha & -\frac{1}{2}(1-\alpha) & \frac{1}{2}(1-\alpha) & -\frac{1}{2}(1-\alpha) & \frac{1}{2}(1-\alpha) & 0 \\ 0 & -\alpha & -1+\alpha & -\alpha & -1+\alpha & 1 \\ 2-\alpha & -2 & 0 & -\alpha & -2+\alpha & 2 \\ 1+\alpha & 0 & -2 & -1-\alpha & -1+\alpha & 2 \\ 2-\alpha & -\alpha & -2+\alpha & -2 & 0 & 2 \\ 1+\alpha & -1-\alpha & -1+\alpha & 0 & -2 & 2 \\ 2\alpha & 0 & -1 & 0 & -1 & 1 \\ 2-2\alpha & -1 & 0 & -1 & 0 & 1 \\ 1 & -2\alpha & -2+2\alpha & -1 & -1 & 2 \\ 1 & -1 & -1 & -2\alpha & -2+2\alpha & 2 \end{bmatrix} \end{array}.$$

We apply FME to $\mathbf{B}^{(4)}$, noting again that $0 < \alpha < 1$. After eliminating variable x_{31}, we reach the inequality $x_{21} \leq 0$, contradicting the assumption that $x_{21} > 0$. This proves that the pair (t_1, t_2) cannot have the minimal Q-criterion in the second NJ clustering step when (t_3, t_4) forms a cherry in the first NJ clustering step, completing the proof of Proposition 2. □

Proof of Proposition 3 We must show that if a source distance matrix generates topology $((t_1, t_2), (t_3, t_4))$, then its corresponding admixed neighbor-joining tree satisfies Properties 1 and 3 and contains (t_1, t_5) or (t_2, t_5) as a cherry.

The Q-criterion for four taxa $\{t_1, t_2, t_3, t_4\}$ is $\mathbf{Q}^{(4)} = \mathbf{A}^{(4)}\mathbf{s}^{(4)}$,

$$\mathbf{A}^{(4)} = \begin{array}{c} \begin{matrix} & 21 & 31 & 32 & 41 & 42 & 43 \end{matrix} \\ \begin{matrix} 21 \\ 31 \\ 32 \\ 41 \\ 42 \\ 43 \end{matrix} \begin{bmatrix} 0 & -1 & -1 & -1 & -1 & 0 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ 0 & -1 & -1 & -1 & -1 & 0 \end{bmatrix} \end{array} \equiv \begin{bmatrix} \mathbf{a}_1^{(4)} \\ \mathbf{a}_2^{(4)} \\ \mathbf{a}_3^{(4)} \\ \mathbf{a}_4^{(4)} \\ \mathbf{a}_5^{(4)} \\ \mathbf{a}_6^{(4)} \end{bmatrix}, \quad \mathbf{s}^{(4)} = \begin{bmatrix} x_{21} \\ x_{31} \\ x_{32} \\ x_{41} \\ x_{42} \\ x_{43} \end{bmatrix}.$$

Because $\mathbf{s}^{(4)} \in D^{(4)}_{((t_1,t_2),(t_3,t_4))}$, the Q-criterion $q_{21} = \mathbf{a}_\mathbf{1}^{(4)} \cdot \mathbf{s}^{(4)}$ for (t_1, t_2), or equivalently for (t_3, t_4), $q_{43} = \mathbf{a}_\mathbf{6}^{(4)} \cdot \mathbf{s}^{(4)}$, is minimal. That is, for $i = 2, 3$,

$$(\mathbf{a}_\mathbf{1}^{(4)} - \mathbf{a}_\mathbf{2}^{(4)}) \cdot \mathbf{s}^{(4)} \leq 0 \tag{18}$$

$$(\mathbf{a}_\mathbf{1}^{(4)} - \mathbf{a}_\mathbf{3}^{(4)}) \cdot \mathbf{s}^{(4)} \leq 0. \tag{19}$$

Consider an admixed distance matrix $\mathbf{D}^{(5)}$ constructed from source distance matrix $\mathbf{S}^{(4)}$. The Q-criterion for pairs in $\mathbf{D}^{(5)}$ follows Eq. 12. Proposition 1 excludes three pairs from being the first cherry for any input admixed distance matrix, so we are left with 7 potential pairs for the first cherry: (t_1, t_3), (t_2, t_3), (t_1, t_4), (t_2, t_4), (t_3, t_4), (t_1, t_5), and (t_2, t_5). We claim that the first four pairs, (t_1, t_3), (t_2, t_3), (t_1, t_4), and (t_2, t_4), cannot cluster in the first iteration of the NJ algorithm when $\mathbf{s}^{(4)} \in D^{(4)}_{((t_1,t_2),(t_3,t_4))}$.

Suppose for contradiction that (t_1, t_3) is the first pair to agglomerate in constructing $\mathcal{T}_D^{(5)}$, so that the Q-criterion for (t_1, t_3), $q_{31} = \tilde{\mathbf{a}}_2^{(5)} \cdot \mathbf{s}^{(4)}$, is minimal when $\mathbf{s}^{(4)} \in D^{(4)}_{((t_1,t_2),(t_3,t_4))}$. Using Eq. 12 and 13, for all i from 1 to 10 ($i \neq 2$), the following must hold:

$$(\tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_i^{(5)}) \cdot \mathbf{s}^{(4)} \leq 0.$$

Considering this inequality along with Eqs. 18 and 19, we have a linear system of 11 inequalities in the x_{ij}, $\mathbf{B}^{(5)}\mathbf{s}^{(4)} \leq \mathbf{0}$, where

$$\mathbf{B}^{(5)} \equiv \begin{bmatrix} \mathbf{a}_\mathbf{1}^{(4)} - \mathbf{a}_\mathbf{2}^{(4)} \\ \mathbf{a}_\mathbf{1}^{(4)} - \mathbf{a}_\mathbf{3}^{(4)} \\ \tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_1^{(5)} \\ \tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_3^{(5)} \\ \tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_4^{(5)} \\ \tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_5^{(5)} \\ \tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_6^{(5)} \\ \tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_7^{(5)} \\ \tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_8^{(5)} \\ \tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_9^{(5)} \\ \tilde{\mathbf{a}}_2^{(5)} - \tilde{\mathbf{a}}_{10}^{(5)} \end{bmatrix} = \begin{array}{c} \begin{array}{cccccc} 21 & 31 & 32 & 41 & 42 & 43 \end{array} \\ \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 \\ 1 & 0 & -1 & -1 & 0 & 1 \\ -2+\alpha & 2-\alpha & -1+\alpha & 0 & 1 & -1 \\ -1+2\alpha & 2 & -2 & -1 & 1 & 0 \\ 0 & 2-\alpha & -2+\alpha & -2+\alpha & 2-\alpha & 0 \\ -1+2\alpha & 1-\alpha & -1+\alpha & \alpha & -\alpha & 0 \\ -2+\alpha & 2 & 0 & \alpha & 2-\alpha & -2 \\ -2+3\alpha & 2 & -1 & \alpha & 1-\alpha & -1 \\ -\alpha & 1 & 0 & -1+\alpha & 2-\alpha & -1 \\ -1+\alpha & 2-2\alpha & -2+2\alpha & -1+\alpha & 1-\alpha & 0 \\ -1+\alpha & 1 & -1 & -\alpha & \alpha & 0 \end{bmatrix} \end{array}.$$

Application of the FME procedure to $\mathbf{B}^{(5)}$ to eliminate x_{31} and x_{41} results in $x_{21} \leq 0$, a contradiction. Therefore, (t_1, t_3) cannot be the first cherry of $\mathcal{T}_D^{(5)}$.

It follows by symmetry that (t_2, t_3), (t_1, t_4) and (t_2, t_4) cannot cluster in the first step of the NJ algorithm when an inferred source NJ tree of $\mathbf{S}^{(4)}$ has topology $((t_1, t_2), (t_3, t_4))$. In the 4-taxon source distance matrix $\mathbf{S}^{(4)}$ with t_1 and t_2 being source populations from which the admixed taxon t_5 is created, the roles of t_1 and t_2 are interchangeable, as are the roles of t_3 and t_4.

We have so far proven that (t_3, t_4), (t_1, t_5), and (t_2, t_5) are the only possible clusters in the first step of the construction of $\mathcal{T}_D^{(5)}$ when $\mathbf{s}^{(4)} \in D_{((t_1,t_2),(t_3,t_4))}$. If a pair (t_3, t_4) clusters first, then four nodes, $\{t_1, t_2, t_5, t_{34}\}$, are left to join in the second iteration.

Because Proposition 2 says (t_1, t_2) cannot be a cherry in any $\mathcal{T}_D^{(5)}$, the only possible NJ tree topologies are $((t_1, t_5), (t_2, t_{34}))$ and $((t_2, t_5), (t_1, t_{34}))$. The node t_{34} represents the cluster (t_3, t_4), so the final topologies are $((t_1, t_5), t_2, (t_3, t_4))$ and $((t_2, t_5), t_1, (t_3, t_4))$.

If a pair (t_1, t_5) clusters first, then four nodes, $\{t_2, t_3, t_4, t_{15}\}$, are left to join in the second iteration. The possible NJ tree topologies are $((t_{15}, t_2), (t_3, t_4))$, $((t_{15}, t_3), (t_2, t_4))$ and $((t_{15}, t_4), (t_2, t_3))$. Because the node t_{15} represents the cluster (t_1, t_5), the final topologies are $((t_1, t_5), t_2, (t_3, t_4))$, $((t_1, t_5), t_3, (t_2, t_4))$ and $((t_1, t_5), t_4, (t_2, t_3))$. By the same argument, possible final topologies when a pair (t_2, t_5) is the first cherry are $((t_2, t_5), t_1, (t_3, t_4))$, $((t_2, t_5), t_3, (t_1, t_4))$, and $((t_2, t_5), t_4, (t_1, t_3))$.

All 6 topologies of $\mathcal{T}_D^{(5)}$ from three possible choices for the first cluster satisfy Property 3, and the procedures for their construction comply with Property 1. Also, they contain a cherry involving one of the source taxa and the admixed taxon. □

References

Aldous DJ (2001) Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statist Sci 16:23–34

Atteson K (1999) The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25:251–278

Blum MGB, François O (2006) Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst Biol 55:685–691

Boca SM, Rosenberg NA (2011) Mathematical properties of F_{st} between admixed populations and their parental source populations. Theor Popul Biol 80:208–216

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Nat Acad Sci USA 88:839–843

Bryant D (2005) On the uniqueness of the selection criterion in neighbor-joining. J Classif 22:3–15

Buneman P (1974) A note on the metric properties of trees. J Combin Theory Ser B 17:48–50

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Cueto MA, Matsen FA (2011) Polyhedral geometry of phylogenetic rogue taxa. Bull Math Biol 73:1202–1226

Dantzig GB, Eaves BC (1973) Fourier-Motzkin elimination and its dual. J Comb Theory Ser A 14:288–297

Davidson R, Rusinko J, Vernon Z, Xi J (2017) Modeling the distribution of distance data in Euclidean space. In: Harrington HA, Omar M, Wright M (eds) Algebraic and geometric methods in discrete mathematics. American Mathematical Society, Providence, pp 117–136

Eickmeyer K, Huggins P, Pachter L, Yoshida R (2008) On the optimality of the neighbor-joining algorithm. Algorithms Mol Biol 3:5

Eickmeyer K, Yoshida R (2008) The geometry of the neighbor-joining algorithm for small trees. In: Horimoto K, Regensburger G, Rosenkranz M, Yoshida H (eds) Algebraic Biology: AB 2008. Lecture Notes in Computer Science, vol 5147. Springer, Berlin, pp 81–95

Felsenstein J (1984) Distance methods for inferring phylogenies: a justification. Evolution 38:16–24

Felsenstein J (2004) Inferring phylogenies. Sinauer, Sunderland

Gascuel O, Steel M (2006) Neighbor-joining revealed. Mol Biol Evol 23:1997–2000

Kopelman NM, Stone L, Gascuel O, Rosenberg NA (2013) The behavior of admixed populations in neighbor-joining inference of population trees. Pacific Symp Biocomput 18:273–284

Mountain JL, Cavalli-Sforza LL (1994) Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc Nat Acad Sci USA 91:6515–6519

Nee S (2006) Birth-death models in macroevolution. Annu Rev Ecol Evol Syst 37:1–17

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147

Ruiz-Linares A, Minch E, Meyer D, Cavalli-Sforza LL (1995) Analysis of classical and DNA markers for reconstructing human population history. In: Brenner S, Hanihara K (eds) The origin and past of modern humans as viewed from DNA. World Scientific, Singapore, pp 123–148

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sanderson MJ, Shaffer HB (2002) Troubleshooting molecular phylogenetic analyses. Annu Rev Ecol Syst 33:49–72

Schrijver A (1986) Theory of linear and integer programming. Wiley, Chichester

Steel M (2016) Phylogeny: discrete and random processes in evolution, Society for Industrial and Applied Mathematics, Philadelphia

Studier JA, Keppler KJ (1988) A note on the neighbor-joining algorithm of Saitou and Nei. Mol Biol Evol 5:729–731

Thomson RC, Shaffer HB (2010) Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Syst Biol 59:42–58

Westover KM, Rusinko JP, Hoin J, Neal M (2013) Rogue taxa phenomenon: a biological companion to simulation analysis. Mol Phylogenet Evol 69:1–3

Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S., Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213:21–87

Ziegler GM (1995) Lectures on Polytopes. Springer-Verlag, New York, NY