THE BEHAVIOR OF ADMIXED POPULATIONS IN NEIGHBOR-JOINING INFERENCE OF POPULATION TREES

NAAMA M. KOPELMAN and LEWI STONE

Porter School of Environmental Studies, Department of Zoology, Tel Aviv University, Ramat Aviv, Israel

OLIVIER GASCUEL

Méthodes et Algorithmes pour la Bioinformatique, LIRMM-CNRS, Montpellier, France

NOAH A. ROSENBERG*

Department of Biology, Stanford University, Stanford, California, USA
**E-mail: noahr@stanford.edu*

Neighbor-joining is one of the most widely used methods for constructing evolutionary trees. This approach from phylogenetics is often employed in population genetics, where distance matrices obtained from allele frequencies are used to produce a representation of population relationships in the form of a tree. In phylogenetics, the utility of neighbor-joining derives partly from a result that for a class of distance matrices including those that are *additive* or tree-like—generated by summing weights over the edges connecting pairs of taxa in a tree to obtain pairwise distances—application of neighbor-joining recovers exactly the underlying tree. For populations within a species, however, migration and admixture can produce distance matrices that reflect more complex processes than those obtained from the bifurcating trees typical in the multispecies context. Admixed populations—populations descended from recent mixture of groups that have long been separated—have been observed to be located centrally in inferred neighbor-joining trees, with short external branches incident to the path connecting their source populations. Here, using a simple model, we explore mathematically the behavior of an admixed population under neighbor-joining. We show that with an additive distance matrix, a population admixed among two source populations necessarily lies on the path between the sources. Relaxing the additivity requirement, we examine the smallest nontrivial case—four populations, one of which is admixed between two of the other three—showing that the two source populations never merge with each other before one of them merges with the admixed population. Furthermore, the distance on the constructed tree between the admixed population and either source population is always smaller than the distance between the source populations, and the external branch for the admixed population is always incident to the path connecting the sources. We define three properties that hold for four taxa and that we hypothesize are satisfied under more general conditions: *antecedence of clustering*, *intermediacy of distances*, and *intermediacy of path lengths*. Our findings can inform interpretations of neighbor-joining trees with admixed groups, and they provide an explanation for patterns observed in trees of human populations.

Keywords: admixture; neighbor-joining; phylogenetics; population genetics

1. Introduction

Distance matrix methods in phylogenetics construct trees of taxa using algorithms applied to matrices that tabulate pairwise evolutionary distances between the taxa.[1,2] Among these methods, neighbor-joining[3,4] is one of the most popular.[5–7] One of its key features is its consistency: if the distance matrix is *additive*, such that a tree of taxa exists that generates the distances in the matrix, then neighbor-joining recovers this exact tree.[5,8,9] Further, neighbor-

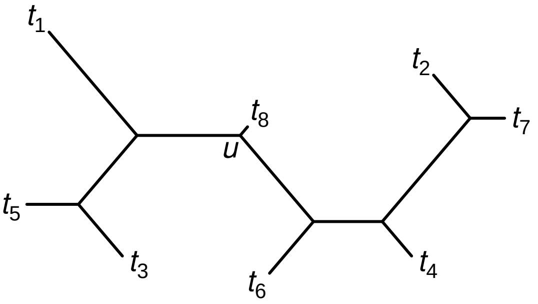

Fig. 1. Properties observed for admixed taxa in neighbor-joining trees. Taxon t_8 represents an admixture of source populations t_1 and t_2. The admixed taxon appears on a short external branch incident to the path connecting the source populations. Denoting distances on the tree by $\hat{d}$ and topological path lengths that count edges separating pairs of taxa by $\hat{b}$, the tree illustrates the properties of *intermediacy of distances* ($\hat{d}_{t_1,t_8} < \hat{d}_{t_1,t_2}$ and $\hat{d}_{t_2,t_8} < \hat{d}_{t_1,t_2}$, or equivalently, $\hat{d}_{u,t_8} < \hat{d}_{u,t_1}$ and $\hat{d}_{u,t_8} < \hat{d}_{u,t_2}$, where u is the unique node that places t_1, t_2, and t_8 in different subtrees), and *intermediacy of path lengths* ($\hat{b}_{t_1,t_8} \leq \hat{b}_{t_1,t_2}$ and $\hat{b}_{t_2,t_8} \leq \hat{b}_{t_1,t_2}$).

joining is robust in that theoretical and simulation-based studies have found it to infer sensible trees under a broad range of mathematical and biological conditions.[7,9–13]

As trees have long been used in population genetics to describe relationships among populations,[14,15] the neighbor-joining algorithm has been applied extensively as a population clustering tool, using distance matrices calculated from population-level allele frequencies. In humans, neighbor-joining trees have been and continue to be a regular feature of studies of population relationships.[16–20] In population-genetic studies, because migration and admixture sometimes generate evolutionary histories that cannot easily be described by a bifurcating tree of populations, a neighbor-joining tree is treated as a type of population clustering diagram rather than a precise representation of the evolutionary history of the populations.

When neighbor-joining has been used with admixed populations—populations recently descended from two or more source groups that have long been separated—particular characteristics of the inferred trees have often been observed (Fig. 1). For example, one simulation study based on human data identified a reduction in the external branch length leading to an admixed population as the strength of gene flow with other populations was increased.[21] It has also been suggested on the basis of observed human population trees that a short external branch for a population on a constructed neighbor-joining tree can imply recent admixture of the population, and that admixed populations often appear in the "middle" of a neighbor-joining tree, on branches incident to paths connecting possible source populations.[21–25] This pattern is evident in Fig. 2, in which admixed Mestizo populations from Latin America lie on branches incident to the path connecting Native American and European populations. Here, we seek to understand these results on the behavior of admixed populations in the application of the neighbor-joining algorithm. We therefore apply neighbor-joining to populations that satisfy a simple admixture model, first considering the case in which the distance matrix is additive. Next, for the case of $n = 4$ taxa, we use a mechanistic mathematical investigation to examine three specific properties of neighbor-joining trees involving an admixed population.

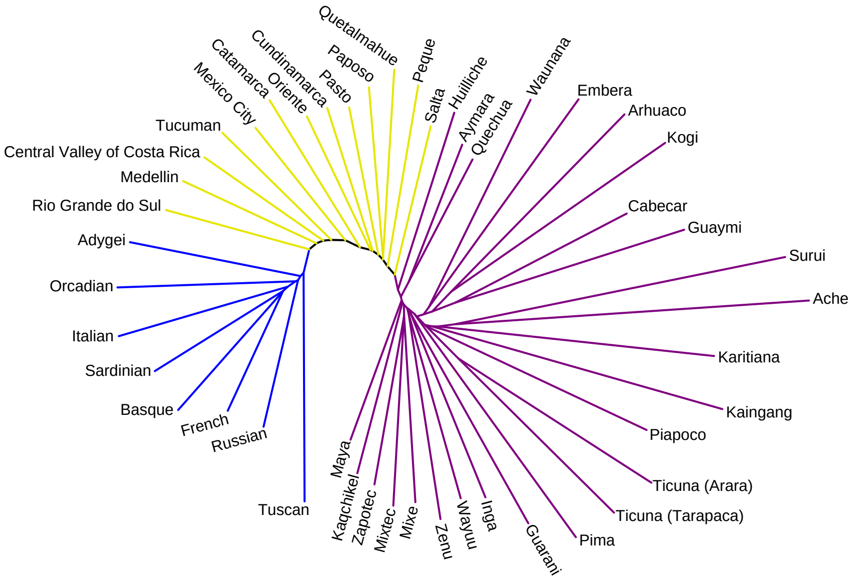

Fig. 2. Neighbor-joining tree of admixed Mestizo populations together with Native American and European populations that represent ancestral source regions for the admixed populations. The tree, obtained using `Neighbor` and `Drawtree` in the `Phylip` package,[26] uses data on 678 microsatellite loci in 13 Mestizo, 26 Native American, and 8 European populations.[27–29] Allele frequencies were computed from 872 individuals—249 Mestizo, 463 Native American, and 160 European—and distances were computed with `Microsat`[30] using one minus the proportion of shared alleles.[31] Mestizo, Native American, and European branches appear in yellow, purple, and blue, respectively. Mestizos lie in the "middle" of the tree, connecting to the path that links the Native Americans and Europeans. External branches for Mestizo populations are shorter on average (0.102) than for Native American (0.146) and European populations (0.109); Mestizo populations have 9 of the 15 shortest external branches.

2. The neighbor-joining algorithm

We briefly review the neighbor-joining algorithm.[3,4] Consider a set of n taxa, together with a distance function d computed for each pair of taxa, such that the distance between taxa i and j is denoted d_{ij}. The algorithm takes as input the distance matrix D containing entries d_{ij}, with i and j ranging from 1 to n, and it outputs a bifurcating unrooted tree. D is symmetric $(d_{ij} = d_{ji})$, with zeroes on the diagonals $(d_{ii} = 0)$ and nonnegative real entries $(d_{ij} \geq 0)$.

As in other agglomerative algorithms that construct bifurcating trees,[2,32] at each of a series of steps, the two nearest taxa according to a selection criterion are connected to a new interior node, becoming "neighbors" on the constructed tree. Branch lengths from the new node to the nodes it agglomerates, as well as the distances to all remaining nodes, are then calculated, and a new distance matrix is obtained. This procedure is repeated iteratively until the last three nodes remain, and these three nodes are then connected to a final interior node. Because the last three nodes are always joined, the number of taxa must exceed three for neighbor-joining to have a nontrivial decision at the first step.

At each step, the key decision is the choice of the pair of taxa that are agglomerated. Neighbor-joining uses an $n \times n$ matrix Q, containing entries q_{ij} for pairs of taxa (i, j):

$$q_{ij} = (n-2)d_{ij} - \sum_{k=1}^{n} d_{ik} - \sum_{k=1}^{n} d_{jk}. \tag{1}$$

The two taxa that are agglomerated are those with the minimal value of q_{ij} (choosing randomly in case of ties). If taxa i and j are agglomerated, then their distances to the new node u become

$$d_{iu} = \frac{1}{2}d_{ij} + \frac{1}{2(n-2)}\left(\sum_{k=1}^{n} d_{ik} - \sum_{k=1}^{n} d_{jk}\right) \tag{2}$$

$$d_{ju} = \frac{1}{2}d_{ij} + \frac{1}{2(n-2)}\left(\sum_{k=1}^{n} d_{jk} - \sum_{k=1}^{n} d_{ik}\right). \tag{3}$$

The distances of all remaining nodes k to node u are computed as

$$d_{ku} = (d_{ik} + d_{jk} - d_{ij})/2. \tag{4}$$

The next agglomeration then proceeds from an $(n-1) \times (n-1)$ distance matrix that replaces distances involving nodes i and j with those involving the single node u.

3. An admixture scenario

We examine a scenario in which one of the taxa is admixed among two of the others. This taxon can be viewed as having been formed from its two source taxa, such that individual members of the taxon have ancestors in both source groups. We label the taxa $t_1, t_2, \ldots, t_n$. Without loss of generality, let taxon t_n be the admixed group, and suppose that it is an admixture of taxa t_1 and t_2. The relationships among the remaining $n-3$ taxa $(t_3, t_4, \ldots, t_{n-1})$ and between these taxa and t_1, t_2, and t_n are not specified; we do not consider any additional admixture relationships that might exist among these taxa. We assume $n \geq 4$, so that at least one taxon is considered in addition to t_1, t_2, and t_n.

In a standard statistical model of admixture used in population genetics, allele frequencies in an admixed taxon are given by linear combinations of the allele frequencies of the source taxa.[33–37] We denote by λ the proportion of the ancestry of taxon t_n arising from t_1 and by $1-\lambda$ the corresponding proportion arising from t_2, where $0 < \lambda < 1$. For any allelic type, if p_{t_i} denotes the frequency of the specified allele in taxon t_i, then

$$p_{t_n} = \lambda p_{t_1} + (1-\lambda)p_{t_2}. \tag{5}$$

It follows that if for each of the taxa in a pair, a distance function d is linear in each component of the allele frequency vector at a locus, then the distances between the admixed taxon and other taxa are obtained as linear combinations of corresponding distances involving taxa t_1 and t_2. Therefore, for $1 \leq i \leq n-1$,

$$d_{t_n,t_i} = \lambda d_{t_1,t_i} + (1-\lambda)d_{t_2,t_i}. \tag{6}$$

Eq. 6 continues to hold if for a series of loci, the distance function d is linear for each taxon in each component of the allele frequency vector *at each locus*, as would occur if the distance between a pair of taxa at a set of loci were computed as the mean of locus-wise distances that were each linear in the components of the allele frequency vector at the specified locus.

We assume that the distance function supplied to neighbor-joining satisfies eq. 6, and that it is symmetric, nonnegative, and zero if and only if it is computed between a taxon and itself; we otherwise do not concern ourselves with the form of the function. While typical population-genetic distance functions often involve nonlinear relationships with allele frequencies and do not necessarily follow eq. 6—consider the nonlinear graphs in Figure 3 of Boca & Rosenberg,[38] which illustrate that for the F_{ST} measure and an admixed population t_n whose frequencies are linear combinations of those of populations t_1 and t_2, $F_{ST}(t_n, t_1) \neq \lambda F_{ST}(t_1, t_1) + (1 - \lambda) F_{ST}(t_2, t_1)$—eq. 6 is a natural extension of the ubiquitous eq. 5 from allele frequencies to distance functions. For F_{ST}, it can be shown from eqs. 1 and 7 of Boca & Rosenberg[38] that for small λ, $F_{ST}(t_n, t_1) \approx \lambda F_{ST}(t_1, t_1) + (1 - \lambda) F_{ST}(t_2, t_1)$. Thus, we view eq. 6 as a reasonable first approximation for examining properties of neighbor-joining in an admixture scenario.

4. The neighbor-joining algorithm in an admixture scenario

Our goal is to construct a distance matrix according to the admixture rule in eq. 6, mechanistically apply neighbor-joining to the matrix, and characterize the properties of the inference process and the resulting inferred tree. We examine two settings. In the first, arbitrarily many taxa are considered, and their distances produce an additive distance matrix (and therefore satisfy a *tree metric*[39]). In the second, a general matrix is investigated, with distances that do not necessarily follow a tree metric, but the matrix includes only four taxa.

4.1. *The additive case for n taxa*

We first assume that the distance matrix is additive. In this case, by the consistency property of the neighbor-joining algorithm,[5,8,9] distances between taxa on the constructed neighbor-joining tree exactly equal those of the input matrix. Denote by $\hat{d}$ the distance function computed for pairs of nodes in the inferred neighbor-joining tree, such that for taxa t_i and t_j, $\hat{d}_{t_i,t_j}$ is the sum of the lengths of the branches on the path connecting t_i and t_j. Recalling that d represents distance in the input distance matrix, if the matrix is additive, then for all (t_i, t_j),

$$\hat{d}_{t_i,t_j} = d_{t_i,t_j}. \tag{7}$$

Because $d_{t_1,t_n} = (1 - \lambda) d_{t_1,t_2}$ and $d_{t_2,t_n} = \lambda d_{t_1,t_2}$ by eq. 6,

$$\hat{d}_{t_1,t_n} = (1 - \lambda) \hat{d}_{t_1,t_2} \tag{8}$$

$$\hat{d}_{t_2,t_n} = \lambda \hat{d}_{t_1,t_2}. \tag{9}$$

It follows that $\hat{d}_{t_1,t_n} + \hat{d}_{t_2,t_n} = \hat{d}_{t_1,t_2}$, from which we can infer that taxa t_1, t_2, and t_n are collinear in the inferred neighbor-joining tree, with t_n in the interior of the path from t_1 to t_2.

We can obtain an even stronger result. Consider a case with at least four taxa: t_1, t_2, t_n, and, without loss of generality, t_3 (Fig. 3A). In the inferred neighbor-joining tree, a path of length c connects taxon t_3 to some point P on the path from t_1 to t_2 (including the endpoints). Without loss of generality, we can assume that P lies on the path from t_1 to t_n (including the endpoints). We denote the distances $\hat{d}_{t_1,P}$ and $\hat{d}_{P,t_n}$ by nonnegative values y and z, respectively. We denote $\hat{d}_{t_1,t_2} = d_{t_1,t_2} = x$, for some nonnegative x.

By eq. 7, $\hat{d}_{t_1,t_n} = y + z = d_{t_1,t_n} = (1 - \lambda)x$. By eqs. 6 and 7,

$$d_{t_3,t_n} = \lambda d_{t_3,t_1} + (1 - \lambda) d_{t_3,t_2} \tag{10}$$

$$\hat{d}_{t_3,t_n} = \lambda \hat{d}_{t_3,t_1} + (1 - \lambda) \hat{d}_{t_3,t_2}. \tag{11}$$

In other words, $c + z = \lambda(c + y) + (1 - \lambda)(c + x - y)$. Together with the relationship $y + z = (1 - \lambda)x$ and the assumption that $\lambda > 0$, eq. 11 implies that $y = 0$. It then follows that taxon t_3 lies on

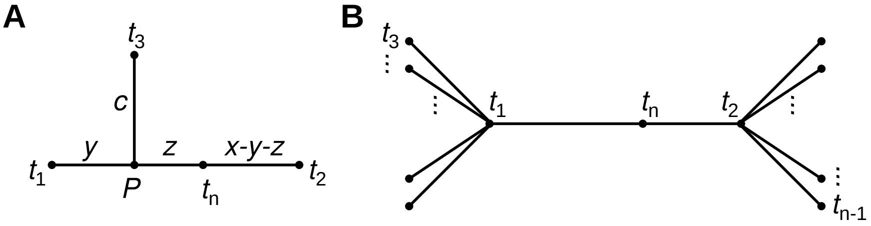

Fig. 3. The case of an additive distance matrix for n taxa. (A) Illustration of distances in the model. In the text it is shown that $y = 0$. (B) The structure required for a tree. Taxa t_1 and t_2 lie at multifurcating nodes, with t_n on the line connecting them. The leaves connected to the multifurcations have labels $t_3, t_4, \ldots, t_{n-1}$.

a line with taxa t_1, t_2, and t_n. Further, taxon t_3 lies on the side of taxon t_1 opposite to taxa t_2 and t_n; otherwise, by eq. 11, we would have $(1-\lambda)x - c = \lambda c + (1-\lambda)(x-c)$, which requires $c = 0$. In turn, $c = 0$ implies $\hat{d}_{t_3,t_1} = 0$, and hence, $d_{t_3,t_1} = 0$, contradicting the assumption that all pairs of taxa are separated by positive distances in the distance matrix.

We have therefore shown that for an additive tree with taxon t_n admixed between t_1 and t_2, any additional taxon beyond t_1, t_2, and t_n must be collinear with t_1, t_2, and t_n, and must lie exterior to the path connecting t_1 and t_2. Thus, each additional taxon $t_3, t_4, \ldots, t_{n-1}$ is connected to t_1 or t_2 by an external branch. The admixture model together with the assumption of an additive distance matrix imposes such a strong restriction on the set of allowed distance matrices that it forces all taxa onto a highly constrained tree (Fig. 3B). When we consider the placement of each taxon $t_3, t_4, \ldots, t_{n-1}$, we find that this tree has two multifurcating nodes separated by a line that joins taxa t_1 and t_2, with t_n as the only intervening taxon.

The additive case can assist in explaining phenomena observed empirically with admixed populations in the application of neighbor-joining:[21–25] in the additive case, t_n has external branch length 0, a result compatible with the short external branches detected for admixed taxa. Further, t_n lies on the path connecting t_1 and t_2, compatible with the observation that admixed taxa lie in the "middle" of inferred neighbor-joining trees, with external branches incident to the paths connecting their source taxa. We can thus see that the empirical Fig. 2 resembles Fig. 3B, as the short internal branches among Native Americans and Europeans give rise to a shape with near multifurcations on each side of the admixed Mestizo groups.

4.2. *The case of $n = 4$ taxa, not necessarily additive*

The additive case is restrictive and atypical of the population-genetic context, in which migration and admixture generate non-tree-like evolution. We can then consider the more general setting of arbitrary genetic distance matrices with positive entries, examining the smallest nontrivial case, with $n = 4$ taxa. In this case, the admixed taxon is t_4, with source taxa t_1 and t_2. We set the distances among taxa t_1, t_2, and t_3 to be $d_{t_1,t_2} = x$, $d_{t_1,t_3} = y$ and $d_{t_2,t_3} = z$, for some positive x, y, and z. Employing eq. 6, the distance matrix D has the form:

$$D = \begin{pmatrix} 0 & x & y & (1-\lambda)x \\ x & 0 & z & \lambda x \\ y & z & 0 & \lambda y + (1-\lambda)z \\ (1-\lambda)x & \lambda x & \lambda y + (1-\lambda)z & 0 \end{pmatrix}. \tag{12}$$

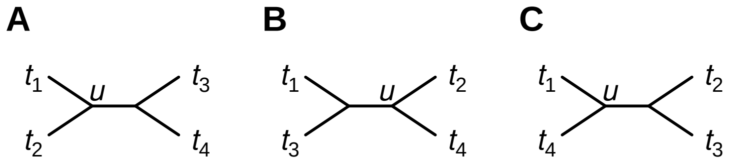

Fig. 4. The three possible topologies for $n = 4$ taxa. Node u is the unique node that places t_1, t_2, and t_4 in different subtrees.

Using eq. 1 to calculate the matrix Q used in deciding which taxa will agglomerate, we obtain

$$Q = \begin{pmatrix} 0 & q_1 & q_2 & q_3 \\ q_1 & 0 & q_3 & q_2 \\ q_2 & q_3 & 0 & q_1 \\ q_3 & q_2 & q_1 & 0 \end{pmatrix}, \tag{13}$$

where

$$q_1 = -(x + y + z) \tag{14}$$
$$q_2 = -(2 - \lambda)x - \lambda y - (2 - \lambda)z \tag{15}$$
$$q_3 = -(1 + \lambda)x - (1 + \lambda)y - (1 - \lambda)z. \tag{16}$$

Examining the relationships among q_1, q_2, and q_3, we have that

$$q_1 < q_2 \quad \Leftrightarrow \quad x + z < y \tag{17}$$
$$q_1 < q_3 \quad \Leftrightarrow \quad x + y < z \tag{18}$$
$$q_2 < q_3 \quad \Leftrightarrow \quad y < (1 - 2\lambda)x + z. \tag{19}$$

As in the work of Eickmeyer & Yoshida,[40] we partition the four-dimensional space of possible values of (λ, q_1, q_2, q_3) according to the tree topologies produced by neighbor-joining.

Three tree topologies are possible with the four taxa (Fig. 4). In Fig. 4A, taxon t_4 is separated by three edges from taxa t_1 and t_2, which themselves are separated by only two edges. In Fig. 4B, t_4 is separated by two edges from t_2 and by three edges from t_1; t_1 and t_2 are separated by three edges. Taxa t_1 and t_2 are also separated by three edges in Fig. 4C, but t_4 is instead separated by two edges from t_1 and by three edges from t_2.

Seven possibilities exist for the smallest entry of Q: (1) q_1, (2) q_2, (3) q_3, (4) q_1 and q_2 (tied), (5) q_1 and q_3 (tied), (6) q_2 and q_3 (tied), and (7) q_1, q_2, and q_3 (all tied). Each choice leads to a particular outcome among the three tree topologies in Fig. 4, with two or more topologies being possible outcomes in cases that involve ties. For each value among q_1, q_2, and q_3, two pairs of taxa produce the same value in the matrix Q. It can be shown that in each case, either choice of which pair is first to agglomerate leads to the same inferred tree. Without loss of generality, we choose the pair that does not include taxon t_3.

Four of the seven cases are not possible. In case 1, summing $x + z < y$ and $x + y < z$ in eqs. 17 and 18, we obtain $x < 0$. In case 4, setting $q_1 = q_2$ in eq. 17, $x + z = y$, from which we obtain $x < 0$ using $x + y < z$ in eq. 18. Similarly, in case 5, $q_1 = q_3$ in eq. 18 produces $x < 0$ using $x + z < y$ in eq. 17. In case 7, eqs. 17-19 become equalities, leading to $x = 0$ when $x + z = y$ is substituted into eq. 19. All of these cases contradict the assumption that $x > 0$.

We consider the three allowable cases (cases 2, 3, and 6). For each of the possible inferred neighbor-joining trees, denote by u the unique interior node that places taxa t_1, t_2, and t_4 in distinct subtrees (Fig. 4). Denote by $\hat{d}$ the distance between nodes on the inferred tree. In case 2, q_2 is smallest, taxa t_2 and t_4 agglomerate first, and using eqs. 2-4, we obtain

$$\hat{d}_{u,t_2} = (\lambda/4)(3x - y + z) \tag{20}$$

$$\hat{d}_{u,t_4} = (\lambda/4)(x + y - z) \tag{21}$$

$$\hat{d}_{u,t_1} = (1 - \lambda)x. \tag{22}$$

We can show that $\hat{d}_{u,t_4} < \hat{d}_{u,t_2}$ and $\hat{d}_{u,t_4} < \hat{d}_{u,t_1}$. The first of these two inequalities is equivalent to $\lambda y < \lambda(x + z)$, which holds because $\lambda > 0$, and because $y < x + z$ by eq. 17. For the second inequality, note first that $y < (1 - 2\lambda)x + z$ by eq. 19. Substituting the right-hand side in place of y in eq. 21, $\hat{d}_{u,t_4}$ is less than $[2\lambda(1 - \lambda)]x/4$, which in turn is less than $\hat{d}_{u,t_1}$ because $0 < \lambda < 1$.

In case 3, q_3 is smallest, taxa t_1 and t_4 agglomerate first, and using eqs. 2-4, we obtain

$$\hat{d}_{u,t_1} = [(1 - \lambda)/4](3x + y - z) \tag{23}$$

$$\hat{d}_{u,t_4} = [(1 - \lambda)/4](x - y + z) \tag{24}$$

$$\hat{d}_{u,t_2} = \lambda x. \tag{25}$$

Similarly to case 2, we show $\hat{d}_{u,t_4} < \hat{d}_{u,t_1}$ and $\hat{d}_{u,t_4} < \hat{d}_{u,t_2}$. The first inequality is equivalent to $(1 - \lambda)z < (1 - \lambda)(x + y)$, which holds because $\lambda < 1$, and because $z < x + y$ by eq. 18. For the second equality, $(1 - 2\lambda)x + z < y$ by eq. 19. Substituting the left-hand side in place of y in eq. 24, $\hat{d}_{u,t_4}$ is less than $[(2\lambda(1 - \lambda)]x/4$, which in turn is smaller than $\hat{d}_{u,t_2}$ because $0 < \lambda < 1$.

Finally, in case 6, q_2 and q_3 are tied with the smallest values, and either t_2 and t_4 agglomerate first as in case 2, or t_1 and t_4 agglomerate first as in case 3. Neighbor-joining produces the tree in Fig. 4C with probability 1/2, and the tree in Fig. 4B with probability 1/2. With either choice, the same arguments used to demonstrate $\hat{d}_{u,t_4} < \hat{d}_{u,t_1}$ and $\hat{d}_{u,t_4} < \hat{d}_{u,t_2}$ in cases 2 and 3 apply, except that y is equal to (instead of greater than or less than) $(1 - 2\lambda)x + z$.

This collection of results demonstrates three phenomena for four-taxon trees built from distance matrices formed according to our admixture model. (1) The admixed taxon agglomerates with one of its two source taxa before the sources agglomerate with each other. Cases 2, 3, and 6 are the only ones allowable, and in these cases, the first neighbor-joining step agglomerates the admixed taxon t_4 with one of the sources. (2) Denoting by u the unique node for which the admixed taxon and its source taxa all lie in different subtrees, the distance on the neighbor-joining tree of the admixed taxon to u is smaller than the distances to u of both source taxa. We demonstrated this result in each of the allowed cases, and it therefore holds in general. (3) The number of edges separating the source taxa on the inferred neighbor-joining tree, for each source taxon, is greater than or equal to the number of edges separating the admixed taxon from the source taxon. Only the trees in Figs. 4B and 4C are possible outcomes of neighbor-joining in our model, and the result holds for each of these trees.

5. Properties

Using the four-taxon results, we can formally define three properties of a distance matrix and its resulting neighbor-joining tree. The properties are well-defined for arbitrary n, and it is possible to evaluate whether a given n-taxon distance matrix satisfies them when neighbor-joining is applied. All three properties are possessed by all matrices generated by the four-taxon case of our admixture model.

Property 1: antecedence of clustering. The admixed taxon clusters with one of its source taxa before the source taxa cluster together. Stated precisely, some clade containing t_1 but not t_2 or t_n merges with some clade containing t_n but not t_1 or t_2, or, some clade containing t_2 but not t_1 or t_n merges with some clade containing t_n but not t_1 or t_2, before any clade containing t_1 but not t_2 or t_n merges with any clade containing t_2 but not t_1 or t_n.

Here we allow a clade to have any size, and potentially only a single taxon. In identifying the steps at which t_1, t_2, and t_n merge into the neighbor-joining tree, as in our four-taxon case, to ensure that these taxa do not all merge simultaneously at the final stage, we adopt the convention that if a four-taxon stage is reached in which t_1, t_2, and t_n lie in separate subtrees, we choose to agglomerate two among these three subtrees rather than agglomerating the third one with the unique available subtree that does not contain t_1, t_2, or t_n.

Property 2: intermediacy of distances. The distance on the constructed neighbor-joining tree between the admixed taxon and either of its source taxa is smaller than the corresponding distance between the two source taxa. That is, $\hat{d}_{t_1,t_n} < \hat{d}_{t_1,t_2}$ and $\hat{d}_{t_2,t_n} < \hat{d}_{t_1,t_2}$. Equivalently, if u is the unique node in the constructed neighbor-joining tree for which t_1, t_2, and t_n lie in different subtrees, then $\hat{d}_{u,t_n} < \hat{d}_{u,t_1}$ and $\hat{d}_{u,t_n} < \hat{d}_{u,t_2}$.

Property 3: intermediacy of path lengths. The number of edges separating the source taxa in the constructed neighbor-joining tree is greater than or equal to the number of edges separating the admixed taxon and either source taxon. If we define $\hat{b}_{ij}$ as the number of edges in the path separating nodes i and j in the inferred tree, then $\hat{b}_{t_1,t_2} \geq \hat{b}_{t_1,t_n}$ and $\hat{b}_{t_1,t_2} \geq \hat{b}_{t_2,t_n}$.

We have already demonstrated that in our admixture model, Properties 2 and 3 hold for all distance matrices in the n-taxon additive case; for Property 2, using eqs. 8 and 9 and $0 < \lambda < 1$, $\hat{d}_{t_1,t_n} < \hat{d}_{t_1,t_2}$ and $\hat{d}_{t_2,t_n} < \hat{d}_{t_1,t_2}$. For Property 3, we have shown that for an n-taxon additive distance matrix, taxon t_n lies on the interior of the path connecting t_1 and t_2, and it is the only taxon so located. Thus, $\hat{b}_{t_1,t_2} = 2$, while $\hat{b}_{t_1,t_n} = \hat{b}_{t_2,t_n} = 1$, and Property 3 holds.

6. Discussion

We have examined neighbor-joining in a model in which an admixed taxon is produced from two source taxa, finding that for a four-taxon scenario, distance matrices and their resulting trees possess three properties: *antecedence of clustering*, in which the admixed population clusters with one of the sources before the sources cluster with each other; *intermediacy of distances*, in which the distance on the constructed tree between the admixed taxon and either source taxon is less than the distance between the sources; and *intermediacy of path lengths*, in which the number of edges separating the admixed taxon and either source taxon is no larger than the number of edges separating the sources. We have further shown that for an arbitrary number of taxa, the latter two properties hold when the distance matrix is additive.

By a mechanistic examination, we have found that our model has features seen in empirical

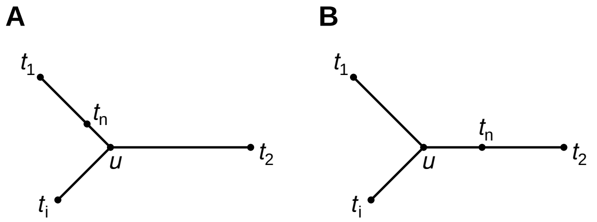

Fig. 5. Two possible placements of t_n with respect to t_1, t_2, and t_i, illustrating how $t_1, t_2, \ldots, t_n$ can have an additive distance matrix when $t_1, t_2, \ldots, t_{n-1}$ have an additive distance matrix. For convenience, we illustrate only one representative taxon t_i from among $\{t_3, t_4, \ldots, t_{n-1}\}$.

observations of neighbor-joining trees that involve admixed populations. In particular, the placement of admixed populations on short external branches incident to the paths connecting their source populations[21–25] matches the demonstration in the additive and four-taxon cases of the *intermediacy of distances* and *intermediacy of path lengths* properties. The theoretical approach validates the view that populations that are centrally located on neighbor-joining trees and that possess short external branches might be recently admixed.

Our results suggest a broader investigation of the extent to which the three properties hold with an arbitrary number of taxa. We have not reported a result regarding *antecedence of clustering* in the n-taxon additive case, nor have we commented on any of the properties for general n-taxon distance matrices that are not necessarily additive. However, we expect that Properties 1-3 will be satisfied by our admixture model considerably more often than in a model in which no special constraints are imposed on distances that involve the nth taxon. As our model also involves an nth taxon with special features, the general analysis of the model might benefit from the "rogue taxon" framework of Cueto & Matsen,[41] in which the addition of an nth taxon alters the tree produced for an initial group of $n-1$ taxa.

An additional direction is to study alternative admixture models. Distance methods are most sensible when a distance is nearly additive; however, eq. 6 severely restricts the distance matrix, as it forces a structure with two multifurcating nodes. This aspect of the model can be relaxed by assuming that the distance is additive for taxa $t_1, t_2, \ldots, t_{n-1}$, and that only distances involving t_n satisfy eq. 6. For $1 \leq i \leq n-1$, we can then apply the distance

$$d_{t_n,t_i} = \begin{cases} d_{t_1,t_i} - (1-\lambda)d_{t_1,t_2} & \text{if} \quad (1-2\lambda)d_{t_1,t_2} \leq d_{t_1,t_i} \\ d_{t_2,t_i} - \lambda d_{t_1,t_2} & \text{if} \quad (1-2\lambda)d_{t_1,t_2} \geq d_{t_1,t_i}. \end{cases} \tag{26}$$

With this distance function, t_n is simply placed on the path from t_1 to t_2 in a preexisting tree relating taxa $t_1, t_2, \ldots, t_{n-1}$ (Fig. 5). Properties 2 and 3 continue to hold.

To obtain eq. 26, we first suppose that $t_1, t_2, \ldots, t_{n-1}$ have an additive distance matrix. We wish to place taxon t_n on the tree that generates the matrix so that the matrix for $t_1, t_2, \ldots, t_n$ is additive. First, given λ, t_n is placed on the path from t_1 to t_2 such that eqs. 8 and 9 are satisfied. It remains to compute $\hat{d}_{t_n,t_i}$ for $i = 3, 4, \ldots, n-1$. Denote by u the unique node of the tree that places t_1, t_2, and t_i in distinct subtrees (Fig. 5). Then

$$\hat{d}_{t_1,u} = (\hat{d}_{t_1,t_2} + \hat{d}_{t_1,t_i} - \hat{d}_{t_2,t_i})/2 \tag{27}$$

$$\hat{d}_{t_2,u} = (\hat{d}_{t_1,t_2} + \hat{d}_{t_2,t_i} - \hat{d}_{t_1,t_i})/2 \tag{28}$$

$$\hat{d}_{t_i,u} = (\hat{d}_{t_1,t_i} + \hat{d}_{t_2,t_i} - \hat{d}_{t_1,t_2})/2. \tag{29}$$

If $\hat{d}_{t_1,t_n} \leq \hat{d}_{t_1,u}$, then t_n lies on the path from t_1 to u (Fig. 5A), and

$$\hat{d}_{t_n,t_i} = \hat{d}_{u,t_i} + \hat{d}_{t_1,t_2} - \hat{d}_{t_1,t_n}. \tag{30}$$

If, on the other hand, $\hat{d}_{t_1,t_n} \geq \hat{d}_{t_1,u}$, then t_n lies on the path from t_2 to u (Fig. 5B), and

$$\hat{d}_{t_n,t_i} = \hat{d}_{u,t_i} + \hat{d}_{t_1,t_2} - \hat{d}_{t_2,t_n}. \tag{31}$$

Applying eqs. 8, 9, and 27-29 together with the fact that $\hat{d} = d$ for additive distance matrices, we produce the relationship in eq. 26.

Analysis of the three properties using this modified form for the admixture model, or more generally using specific distance functions commonly employed in population genetics, will further illuminate the features of neighbor-joining in admixed populations. Such analyses might also facilitate investigations of the behavior with admixed populations of other tree-building methods, or of phylogenetic network methods[42] that are more directly designed to accommodate taxa with non-tree-like evolutionary histories.

Acknowledgments

We thank M. Feldman and four reviewers for comments. This work has been supported by National Institutes of Health grant R01 GM081441, by National Science Foundation grants BCS-1024627 and DBI-1146722, and by the Burroughs Wellcome Fund.

References

1. D. L. Swofford, G. J. Olsen, P. J. Waddell and D. M. Hillis, Phylogenetic inference, in *Molecular Systematics*, eds. D. M. Hillis, C. Moritz and B. K. Mable (Sinauer, Sunderland, MA, 1996) pp. 407–514.
2. J. Felsenstein, *Inferring Phylogenies* (Sinauer, Sunderland, MA, 2004).
3. N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
4. J. A. Studier and K. J. Keppler, *Mol. Biol. Evol.* **5**, 729 (1988).
5. D. Bryant, *J. Classif.* **22**, 3 (2005).
6. O. Gascuel and M. Steel, *Mol. Biol. Evol.* **23**, 1997 (2006).
7. R. Mihaescu, D. Levy and L. Pachter, *Algorithmica* **54**, 1 (2009).
8. O. Gascuel, Concerning the NJ algorithm and its unweighted version, UNJ, in *Mathematical Hierarchies and Biology*, eds. B. Mirkin, F. R. McMorris, F. S. Roberts and A. Rzhetsky (American Mathematical Society, Providence, 1997) pp. 149–170.
9. K. Atteson, *Algorithmica* **25**, 251 (1999).
10. N. Saitou and T. Imanishi, *Mol. Biol. Evol.* **6**, 514 (1989).
11. M. K. Kuhner and J. Felsenstein, *Mol. Biol. Evol.* **11**, 459 (1994).
12. C. A. M. Russo, N. Takezaki and M. Nei, *Mol. Biol. Evol.* **13**, 525 (1996).
13. S. T. Kalinowski, *Heredity* **102**, 506 (2009).
14. A. W. F. Edwards and L. L. Cavalli-Sforza, Reconstruction of evolutionary trees, in *Phenetic and Phylogenetic Classification*, eds. V. H. Heywood and J. McNeill (Systematics Association, London, 1964) pp. 67–76.
15. L. L. Cavalli-Sforza and A. W. F. Edwards, *Evolution* **21**, 550 (1967).
16. A. M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd and L. L. Cavalli-Sforza, *Nature* **368**, 455 (1994).
17. J. K. Pritchard, M. Stephens and P. Donnelly, *Genetics* **155**, 945 (2000).

18. M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H.-C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg and A. B. Singleton, *Nature* **451**, 998 (2008).
19. G. Atzmon, L. Hao, I. Pe'er, C. Velez, A. Pearlman, P. F. Palamara, B. Morrow, E. Friedman, C. Oddoux, E. Burns and H. Ostrer, *Am. J. Hum. Genet.* **86**, 850 (2010).
20. K. Hunley and M. Healy, *Am. J. Phys. Anthropol.* **146**, 530 (2011).
21. A. Ruiz-Linares, E. Minch, D. Meyer and L. L. Cavalli-Sforza, Analysis of classical and DNA markers for reconstructing human population history, in *The Origin and Past of Modern Humans as Viewed from DNA*, eds. S. Brenner and K. Hanihara (World Scientific, Singapore, 1995) pp. 123–148.
22. A. M. Bowcock, J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd and L. L. Cavalli-Sforza, *Proc. Natl. Acad. Sci. USA* **88**, 839 (1991).
23. J. L. Mountain, A. A. Lin, A. M. Bowcock and L. L. Cavalli-Sforza, *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **337**, 159 (1992).
24. A. A. Lin, J. M. Hebert, J. L. Mountain and L. L. Cavalli-Sforza, *Gene Geog.* **8**, 191 (1994).
25. J. L. Mountain and L. L. Cavalli-Sforza, *Proc. Natl. Acad. Sci. USA* **91**, 6515 (1994).
26. J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.6 (Department of Genome Sciences, University of Washington, Seattle, 2005).
27. N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard and M. W. Feldman, *PLoS Genet.* **1**, 660 (2005).
28. S. Wang, C. M. Lewis Jr., M. Jakobsson, S. Ramachandran, N. Ray, G. Bedoya, W. Rojas, M. V. Parra, J. A. Molina, C. Gallo, G. Mazzotti, G. Poletti, K. Hill, A. M. Hurtado, D. Labuda, W. Klitz, R. Barrantes, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, L. T. Tsuneto, E. Llop, F. Rothhammer, L. Excoffier, M. W. Feldman, N. A. Rosenberg and A. Ruiz-Linares, *PLoS Genet.* **3**, 2049 (2007).
29. S. Wang, N. Ray, W. Rojas, M. V. Parra, G. Bedoya, C. Gallo, G. Poletti, G. Mazzotti, K. Hill, A. M. Hurtado, B. Camrena, H. Nicolini, W. Klitz, R. Barrantes, J. A. Molina, N. B. Freimer, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, L. T. Tsuneto, J. E. Dipierri, E. L. Alfaro, G. Bailliet, N. O. Bianchi, E. Llop, F. Rothhammer, L. Excoffier and A. Ruiz-Linares, *PLoS Genet.* **4**, e1000037 (2008).
30. E. Minch, A. Ruiz Linares, D. B. Goldstein, M. W. Feldman and L. L. Cavalli-Sforza, MICROSAT (version 1.5d): a program for calculating statistics on microsatellite data (Department of Genetics, Stanford University, Stanford, CA, 1998).
31. J. L. Mountain and L. L. Cavalli-Sforza, *Am. J. Hum. Genet.* **61**, 705 (1997).
32. O. Gascuel, *Mol. Biol. Evol.* **11**, 961 (1994).
33. J. C. Long and P. E. Smouse, *Am. J. Phys. Anthropol.* **61**, 411 (1983).
34. D. A. Fournier, T. D. Beacham, B. E. Riddell and C. A. Busack, *Can. J. Fish. Aquat. Sci.* **41**, 400 (1984).
35. N. A. Rosenberg, L. M. Li, R. Ward and J. K. Pritchard, *Am. J. Hum. Genet.* **73**, 1402 (2003).
36. H. Tang, J. Peng, P. Wang and N. J. Risch, *Genet. Epidemiol.* **28**, 289 (2005).
37. D. H. Alexander, J. Novembre and K. Lange, *Genome Res.* **19**, 1655 (2009).
38. S. M. Boca and N. A. Rosenberg, *Theor. Pop. Biol.* **80**, 208 (2011).
39. C. Semple and M. Steel, *Phylogenetics* (Oxford University Press, Oxford, 2003).
40. K. Eickmeyer and R. Yoshida, *Lect. Notes Comp. Sci.* **5147**, 81 (2008).
41. M. A. Cueto and F. A. Matsen, *Bull. Math. Biol.* **73**, 1202 (2011).
42. D. H. Huson, R. Rupp and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications* (Cambridge University Press, Cambridge, 2010).