A geometrical framework for f -statistics

Gonzalo Oteo-García

Department of Biological Sciences, School of Applied Sciences,
University of Huddersfield, Queensgate, Huddersfield, UK
and

José-Angel Oteo*

Departamento de Física Teórica, Universidad de Valencia,
and Institute for Integrative Systems Biology (I2SysBio),
46100-Burjassot, Valencia, Spain

August 19, 2020

Abstract

A detailed derivation of the f -statistics formalism is made from a geometrical framework. It is shown that the f -statistics appear when a genetic distance matrix is constrained to describe a four population phylogenetic tree. The choice of genetic metric is crucial and plays an outstanding role as regards the tree-like-ness criterion. The case of lack of treeness is interpreted in the formalism as presence of population admixture. In this respect, four formulas are given to estimate the admixture proportions. One of them is the so-called f_4 -ratio estimate and we show that a second one is related to a known result developed in terms of the fixation index F_{ST} . An illustrative numerical simulation of admixture proportion estimates is included. Relationships of the formalism with coalescence times and pairwise sequence differences are also provided.

keywords: f -statistics, population admixture, genetic drift, fixation index F_{ST} , coalescence times

1 Introduction

In the course of time a population will be subject to drift, migration, natural selection, etc., whose recurrence eventually leads to the branching of subgroups as new populations with distinct allele frequencies. However, this might not always be the case. To

*oteo@uv.es

characterize the *tree-like-ness* or lack of this, different measures have been introduced, for example f -statistics [10, 7, 8, 6, 12].

We build up a geometrical framework where the f -statistics emerge as the constraints of a genetic distance matrix to correspond to a phylogenetic tree of four populations. The idea is to consider populations as points in an allele frequency space whose dimension equals the number of SNPs in the dataset and to interpret the three statistics f_2, f_3 , and f_4 of the f -formalism as scalar products of differences of allele frequency vectors. This approach allows also to point out some limitations of the f -formalism.

An f -statistics based criterion for tree-like-ness or lack of it is discussed and interpreted in geometric terms in the allele frequency space. In the population admixture case, four formulas are derived to estimate the admixture proportions. One of them is the already known f_4 -ratio result and another one proves to be similar to a result in [4] developed in terms of the fixation index F_{ST} [14]. Eventually, a relationship of f -formalism with two results in [8] based on coalescence times and pairwise sequence differences is established.

In Section 2 the geometrical framework is introduced. The algebraic constraints on a distance matrix to be related to a phylogenetic tree are studied in Section 3 and the subsequent derivation of f -statistics formalism is in Section 4. The formulation of the tree-like-ness criterion is given in Section 5. Section 6 contains applications of the geometrical approach in a treeness scenario and points out some limitations of the f -formalism. The population admixture case is treated in Section 7. Relationships with coalescence times, pairwise differences and fixation index F_{ST} are given in Section 8. Section 9 contains the discussion and conclusions. Two technical Appendices are included.

2 Allele frequency space S_f

Every population P_i , with $i = 1, \dots, n$, may be characterized by an array that can be represented by a point, or a vector $\vec{p}_i$, in space $\mathbb{R}^m$ of dimension m , which is the number of loci considered. For population P_i we denote

$$\vec{p}_i = (q_i^{(1)}, q_i^{(2)}, \dots, q_i^{(m)}), \quad 0 \leq q_i^{(j)} \leq 1, \quad j = 1, \dots, m; \quad i = 1, \dots, n, \quad (1)$$

where $q_i^{(j)}$ is the allele frequency of SNP j in population P_i . Every vector is bounded by $\sum_{j=1}^m q_i^{(j)} \leq m$.

Populations that are genetically close to each other appear as clustered points in this space. As a consequence of genetic evolution this picture has a dynamic character because the numerical values of frequencies $q_i^{(j)}$ do change in time. The net effect is a variation in the modulus and orientation of vector $\vec{p}_i$ which describes an erratic trajectory in $\mathbb{R}^m$ as time passes. Figure 1a gives a representation of the situation in $\mathbb{R}^3$. At an initial time three populations are represented by $P_{1_0}, P_{2_0}, P_{x_0}$. At a later time they

Figure 1: (a) Three-dimensional representation of the birth of the admixed population P_{x0} from P_{10} and P_{20} , in S_f . As time passes the three vectors describe erratic trajectories (green) and give rise to evolved populations P_1, P_x, P_2 : The segment $\overline{P_{10}P_{x0}P_{20}}$ becomes triangle $P_1P_xP_2$. (b) Enhanced illustration of the triangle in panel (a). The ratio of lengths u over v gives the admixture proportion estimate α_{ls} in (30).

are found in P_1, P_2, P_x , due to genetic evolution. The crinkly lines are mere illustrations of the paths they might have followed in the space.

Given three population vectors $\vec{p}_{10}, \vec{p}_{20}, \vec{p}_{x0}$ in $\mathbb{R}^m$, we will make extensive use of the following linear combination

$$\vec{p}_{x0} = \alpha \vec{p}_{10} + (1 - \alpha) \vec{p}_{20}. \quad (2)$$

where $0 < \alpha < 1$. It means that we build up a new population vector $\vec{p}_{x0}$ out of $\vec{p}_{10}$ and $\vec{p}_{20}$, such that every vector component i is obtained as: $q_{x0}^{(i)} = \alpha q_{10}^{(i)} + (1 - \alpha) q_{20}^{(i)}$. Figure 1a illustrates an important property: The three vectors involved in the linear combination are coplanar. Besides, $\vec{p}_{x0}$ is the inner vector in the triplet and its tip stays on the segment $\overline{P_{10}P_{20}}$. Moreover, the ratio obtained from length of segment $\overline{P_{x0}P_{20}}$ over length of segment $\overline{P_{10}P_{20}}$, equals the value α . These features will be important to assess the nature of the approximations we will build up to determine a population mixing proportion.

One further technical element that we need to introduce is the ordinary Euclidean scalar product (or dot product) of two vectors, defined as

$$\vec{p}_1 \cdot \vec{p}_2 = \sum_{i=1}^m q_1^{(i)} q_2^{(i)} = |\vec{p}_1| |\vec{p}_2| \cos \varphi, \quad (3)$$

with $|\vec{p}| = \sqrt{\vec{p} \cdot \vec{p}}$, the modulus of vector $\vec{p}$; and φ the angle between the two vectors.

Let us assume that vector $\vec{p}_2$ is of unit length, i.e. $|\vec{p}_2| = 1$. Then $\vec{p}_1 \cdot \vec{p}_2 = |\vec{p}_1| \cos \varphi$, which is the length of the orthogonal projection of vector $\vec{p}_1$ along the direction defined by the unit vector $\vec{p}_2$. This geometric interpretation of the scalar product and the two (equivalent) definitions of the dot product (3) will be instrumental in the interpretation of f -formulas.

Figure 2: Four-population (P_i) trees with branch lengths (ℓ_i) in (a) and (b). Admixture graph (c) where populations P_1, P_2 give rise to P_x . The length of admixture branches (red) is proportional to the contributor weight.

We will denote S_f the *allele frequency space* $\mathbb{R}^m$ endowed with the operations linear combination and scalar product.

Eventually, the choice of a particular genetic distance in S_f will be a crucial point in the derivation of the f -statistics formalism. The mathematical definition of distance has a number of requirements. We will pay attention to the so-called *triangle inequality* or *sub-additivity property* which establishes that the distance in S_f between P_i and P_j is always smaller or equal to the sum of distances between P_i and P_k , on one side, and P_k and P_j on another.

3 Genetic distance matrix and treeness

Given a set of n populations in S_f , it is always possible to express part of the information in terms of a symmetric *genetic distance matrix* $\mathcal{D} = (d_{ik})$, where d_{ik} stands for a genetic distance measure between populations P_i and P_k , with $i, k = 1, \dots, n$. For the time being we work with a generic genetic distance measure.

We commence by working out a formal result in [11] to provide the explicit algebraic conditions needed for $\mathcal{D}$ to correspond to a phylogenetic tree, in the cases with three and four populations. Concomitantly, an important idea introduced by f -statistics formalism consists in studying trees with three and four populations to avoid direct analysis of the whole putative tree, which is in general a cumbersome task. These results lead in a simple way to introduce the f -statistics as algebraic constraints on the distance matrix with a particular metric.

3.1 The elements of the four-population tree

The four tips and outer branch lengths in Figure 2a are denoted P_i and ℓ_i , ($i = 1, 2, 3, 4$), respectively. The inner branch is denoted ℓ_0 . Besides, the distances between populations P_i and P_k are represented by the matrix elements d_{ik} . For the time being no particular metric is chosen. The following system of six linear equations and five unknowns

is the algebraic formulation of the tree topology

$$\begin{aligned}\ell_1 + \ell_2 &= d_{12}, \\ \ell_3 + \ell_4 &= d_{34}, \\ \ell_0 + \ell_1 + \ell_3 &= d_{13}, \\ \ell_0 + \ell_1 + \ell_4 &= d_{14}, \\ \ell_0 + \ell_2 + \ell_3 &= d_{23}, \\ \ell_0 + \ell_2 + \ell_4 &= d_{24}.\end{aligned}\tag{4}$$

According to [11, Lemma 1] a necessary and sufficient condition for the matrix $\mathcal{D}$ to have a tree realization is that the system above has a solution in non-negative numbers. Higher dimension cases are solvable in terms of this one. The condition for system (4) to have a solution requires the augmented matrix of system (4)

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & d_{12} \\ 0 & 0 & 0 & 1 & 1 & d_{34} \\ 1 & 1 & 0 & 1 & 0 & d_{13} \\ 1 & 1 & 0 & 0 & 1 & d_{14} \\ 1 & 0 & 1 & 1 & 0 & d_{23} \\ 1 & 0 & 1 & 0 & 1 & d_{24} \end{pmatrix},\tag{5}$$

to be $\text{rank}(A) = 5$. This condition implies $\det(A) = 0$, and after some algebra yields the algebraic treeness condition

$$d_{14} + d_{23} - d_{13} - d_{24} = 0.\tag{6}$$

In such a case the solution reads

$$\begin{aligned}\ell_0 &= \frac{1}{2}(d_{14} + d_{23} - d_{12} - d_{34}) = \frac{1}{2}(d_{13} + d_{24} - d_{12} - d_{34}), \\ \ell_1 &= \frac{1}{2}(d_{12} + d_{13} - d_{23}) = \frac{1}{2}(d_{12} + d_{14} - d_{24}), \\ \ell_2 &= \frac{1}{2}(d_{12} + d_{23} - d_{13}) = \frac{1}{2}(d_{12} + d_{24} - d_{14}), \\ \ell_3 &= \frac{1}{2}(d_{13} + d_{34} - d_{14}) = \frac{1}{2}(d_{23} + d_{34} - d_{24}), \\ \ell_4 &= \frac{1}{2}(d_{14} + d_{34} - d_{13}) = \frac{1}{2}(d_{24} + d_{34} - d_{23}).\end{aligned}\tag{7}$$

Once evaluated, if all ℓ_i are non-negative numbers then the tree exists and is unique. More specifically, given that all $d_{ij} \geq 0$, then $\ell_0 \geq 0$ in the first line of (7) implies

$$d_{12} + d_{34} \leq d_{14} + d_{23} = d_{13} + d_{24}.\tag{8}$$

Alternatively, if (8) holds then it is $\ell_0 \geq 0$ in (7). This inequality is known as *four-point condition* [1] and guarantees, on its own, the tree to be unique. Observe that it merely establishes that the inner branch length ℓ_0 is a positive quantity.

Due to the symmetry of the distance matrix $\mathcal{D}$ and the equation system, only the trees $((P_1, P_2), P_3), P_4$, $((P_1, P_3), P_2), P_4$ and $((P_1, P_4), P_2), P_3$, have to be considered [11].

So far no mention has been made to the particular distance definition used to build up the matrix $\mathcal{D}$ and therefore all these results are generic. Any mathematical distance among the plethora of existing definitions is valid, although some of them might capture more information from the dataset at hand than others.

The equation system (4) already appears in [2, Figure 79] where three different approaches are suggested: Least squares, maximum likelihood and minimum evolution. Interestingly, it is the algebraic solution developed above which leads to the f -statistics framework, as we will show.

3.2 The elements of the three-population tree

The algebraic solution of a three-population tree goes as follows. There are three unknowns, the branch lengths ℓ_1, ℓ_2, ℓ_3 , and three equations for the tree topology

$$\begin{aligned}\ell_1 + \ell_2 &= d_{12}, \\ \ell_1 + \ell_3 &= d_{13}, \\ \ell_2 + \ell_3 &= d_{23},\end{aligned}\tag{9}$$

whose solution reads

$$\begin{aligned}\ell_1 &= \frac{1}{2}(d_{12} + d_{13} - d_{23}), \\ \ell_2 &= \frac{1}{2}(d_{12} + d_{23} - d_{13}), \\ \ell_3 &= \frac{1}{2}(d_{13} + d_{23} - d_{12}).\end{aligned}\tag{10}$$

The positive-ness of branches, $\ell_1 > 0$, $\ell_2 > 0$, $\ell_3 > 0$, should be guaranteed by the sub-additivity property of the genetic distance. However, it turns out in some practical cases that the numerical value of one branch is negative. All we can say then is that those three populations are not related by genetic drift, at least as regards the distance definition used. Hence, the lack of treeness is related to the breaking of the triangle property of the distance.

4 Derivation of f -statistics from a genetic distance matrix

The f -statistics formalism is readily obtained from the framework developed above when the following genetic distance is chosen

$$d_{ij} = \frac{1}{m} \sum_{k=1}^m \left(q_i^{(k)} - q_j^{(k)} \right)^2 = \frac{1}{m} (\vec{p}_i - \vec{p}_j)^2 \equiv f_2(P_i, P_j),\tag{11}$$

where we have introduced the notation f_2 to refer to the matrix elements of $\mathcal{D}$ with this choice of metric. The definition (11) corresponds to a squared Euclidean distance, a fact that requires important mathematical considerations. We address this issue below in Section 5.

The four population algebraic treeness condition (6) becomes the following f -formula

$$d_{14} + d_{23} - d_{13} - d_{24} = f_2(P_1, P_4) + f_2(P_2, P_3) - f_2(P_1, P_3) - f_2(P_2, P_4) = 0 \quad (12)$$

The sum of f_2 terms in this equation can be expressed as a single scalar product which, in turn, leads to the definition of the statistic f_4

$$\begin{aligned} \frac{1}{2}[f_2(P_1, P_4) + f_2(P_2, P_3) - f_2(P_1, P_3) - f_2(P_2, P_4)] &= \frac{1}{m} (\vec{p}_1 - \vec{p}_2) \cdot (\vec{p}_3 - \vec{p}_4) \\ &\equiv f_4(P_1, P_2; P_3, P_4). \end{aligned} \quad (13)$$

The treeness condition (12) reads then

$$f_4(P_1, P_2; P_3, P_4) \equiv f_4^T((P_1, P_2; P_3, P_4) = 0. \quad (14)$$

The inner branch length ℓ_0 is provided by (7) and may be written down in terms of f_4 too

$$\ell_0 = f_4(P_1, P_3; P_2, P_4) = f_4(P_1, P_4; P_2, P_3) \equiv f_4^B(P_1, P_2; P_3, P_4). \quad (15)$$

Following [8], we have introduced the superscript notation with f_4^T , for *treeness* test; and f_4^B , for inner *branch* length evaluation; in order to stave off explicit rearrangements in the arguments of f_4 . Thus, f_4 serves two purposes in the context of treeness. The evaluation of f_4^T proceeds directly as in definition (13), whereas the computation of f_4^B is preceded by an argument reordering.

The lengths of the tree external branches provided by (7) have the structure: $d_{ij} + d_{ik} - d_{jk}$, which should be a positive quantity according to the triangle inequality. Interestingly, it can be shaped as a scalar product

$$\begin{aligned} \frac{1}{2}(d_{ij} + d_{ik} - d_{jk}) &= \frac{1}{2}[f_2(P_i, P_j) + f_2(P_i, P_k) - f_2(P_j, P_k)] \\ &= \frac{1}{m}(\vec{p}_i - \vec{p}_j) \cdot (\vec{p}_i - \vec{p}_k) \equiv f_3(P_i; P_j, P_k), \quad i, j = 1, \dots, 4, \end{aligned} \quad (16)$$

where we have defined the new statistic f_3 . Therefore

$$\begin{aligned} \ell_1 &= f_3(P_1; P_2, P_3) = f_3(P_1; P_2, P_4), \\ \ell_2 &= f_3(P_2; P_1, P_3) = f_3(P_2; P_1, P_4), \\ \ell_3 &= f_3(P_3; P_1, P_4) = f_3(P_3; P_2, P_4), \\ \ell_4 &= f_3(P_4; P_1, P_3) = f_3(P_4; P_2, P_3). \end{aligned} \quad (17)$$

In Appendix B we gather a number of useful algebraic f -identities.

Figure 3: With four populations, treeness condition (12) requires the total length in red to equate that in blue.

Figure 3 illustrates the geometric meaning of treeness condition (12): For the given tree (black) the sum of lengths in red, on one side, and in blue, on another, must match. Alternatively, the four point condition (8) can be also used. In terms of f_2 it reads

$$f_2(P_1, P_2) + f_2(P_3, P_4) \leq f_2(P_1, P_3) + f_2(P_2, P_4) = f_2(P_1, P_4) + f_2(P_2, P_3). \quad (18)$$

Eventually the equations (10) for the three–population tree branch lengths become

$$\ell_1 = f_3(P_1; P_2, P_3), \quad \ell_2 = f_3(P_2; P_1, P_3), \quad \ell_3 = f_3(P_3; P_1, P_2). \quad (19)$$

5 The metric f_2 : Triangle criterion

As defined in (11), f_2 is a squared Euclidean distance in S_f , because it is the scalar product of a vector with itself, scaled by the space dimension. However, the squared Euclidean distance is not properly a metric because the triangle inequality (or sub–additivity property) that establishes $d_{ik} \leq d_{ij} + d_{jk}$, for all i, j, k ; does not hold for any combination of three vectors. It is however possible to add some constraints that may render f_2 a true distance in some practical situations. To see this let us consider three arbitrary population vectors $\vec{p}_1, \vec{p}_2, \vec{p}_3$, in S_f and work out the scalar product that appears in the definition of $f_2(\vec{p}_1, \vec{p}_2)$

$$\begin{aligned} (\vec{p}_1 - \vec{p}_2)^2 &= (\vec{p}_1 - \vec{p}_3 + \vec{p}_3 - \vec{p}_2)^2 = \\ &= (\vec{p}_3 - \vec{p}_1)^2 + (\vec{p}_3 - \vec{p}_2)^2 - 2(\vec{p}_3 - \vec{p}_1) \cdot (\vec{p}_3 - \vec{p}_2). \end{aligned} \quad (20)$$

Using (11) and (16) we get the equivalent expression

$$f_2(P_1, P_2) = f_2(P_1, P_3) + f_2(P_3, P_2) - 2f_3(P_3; P_1, P_2). \quad (21)$$

If and only if the scalar product in the rhs of (20) is positive, or $f_3(P_3; P_1, P_2) > 0$ in (21), the sub–additivity condition is preserved. According to the definition of scalar

product, the condition $(\vec{p}_3 - \vec{p}_1) \cdot (\vec{p}_3 - \vec{p}_2) > 0$, means that the angle ϕ between the difference vectors $\vec{p}_3 - \vec{p}_1$ and $\vec{p}_3 - \vec{p}_2$, lays in the interval $\phi \in [0, 90^\circ]$. Since sub-additivity must hold for all vector combinations the outcome is that the three populations must conform a triangle in S_f with all three angles acute. We will refer to this as the *triangle criterion*. Thus, in terms of f_3 the requirements for f_2 to act as a true distance in S_f are

$$\text{Triangle criterion : } \begin{cases} f_3(P_1; P_2, P_3) \geq 0, \\ f_3(P_2; P_1, P_3) \geq 0, \\ f_3(P_3; P_1, P_2) \geq 0. \end{cases} \quad (22)$$

Only in that case f_2 acts as a mathematical distance, the distance matrix $\mathcal{D}$ makes sense, and the results of Section 3 concerning treeness apply. A similar qualitative description appears in [4, Fig. 1.17.1].

Had we four populations to be compared, then the image would be that of a tetrahedron with all angles acute and therefore twelve inequalities in (22) that would secure f_2 acting as true metric within this set of four populations.

In this context, finding one value $f_3 < 0$ means that the assumed metric f_2 has lost its distance character and thus the mathematical model that studies a phylogeny does not apply to that combination of vectors. One reason may be that these populations are rather related by an admixture process. This is illustrated in Figure 1b, where the vertex P_x in the triangle is *too close* to the segment $\overline{P_1 P_2}$; and the meaning of too close is that the angle ϕ at vertex P_x is obtuse. A related discussion on the distance character of f_2 and formal results on additivity may be found in [12].

6 Applications to a treeness scenario

We review two tests used commonly in treeness cases and point out that the outcome may be conditioned by the chosen outgroup populations.

6.1 Outgroup f_3 -statistics test

The so-called outgroup f_3 -statistics test [9, 8] has been proposed to determine the closest population P_i (in the sense of treeness), among a set $\{P_1, \dots, P_n\}$, to a test population P_{test} using as reference an outgroup population P_{out} . Given the tree $((P_{test}, P_i), P_{out})$ The longer the branch from the outgroup population P_{out} to the node (P_{test}, P_i) , i.e. $f_3(P_{out}; P_i, P_{test})$, the closer P_i to P_{test} . Note that the answer is not provided by the minimum value among the two-populations distances in the set $\{f_2(P_{test}, P_i)\}$, with $i = 1, \dots, n$. The triangle criterion (22) should routinely precede the outgroup f_3 -statistics procedure in order to rule out population triplets not related by treeness.

Figure 4a illustrates this commonly used procedure. Population P_a is closer to P_{test} than population P_b , because $f_3(P_{out}; P_{test}, P_a) > f_3(P_{out}; P_{test}, P_b)$.

Figure 4: (a) Scheme for the interpretation of outgroup f_3 -statistics. The graph is the union of three-population trees having in common P_{test} and P_{out} . P_a and P_b are two populations among a larger set represented by unlabelled (grey) branches. f_3 gives the length of the branch between P_{out} and the nodes (P_{test}, P_a) (blue arrow) and (P_{test}, P_b) (red arrow). Here $f_3(P_{out}; P_{test}, P_a) > f_3(P_{out}; P_{test}, P_b)$, and so P_a is closer to P_{test} than P_b . (b) Scheme for the interpretation of the four-population test. The graph is the union of two four-population trees. Coded in colour, f_4^B gives the inner branch length of the trees.

The outgroup f_3 statistics has a drawback because the distance ranking of populations may depend on the choice of outgroup. To illustrate this we build up the following counterexample. A number of assumptions and simplifications are introduced in order to ease the explanation. It is not meant that the drawback takes place only in this singular study case.

Consider a test population P_{test} , two outgroup populations P_{out_1}, P_{out_2} , and two populations P_a, P_b . The question is: Which one of the latter two populations is closer to the test population? Let us assume that in space S_f , the tips of all five population vectors are approximately on the same plane with the distribution of vector tips on such a plane as in Figure 5. This is an unlikely situation but simplifies dramatically the explanation. The only vector tip locations related via treeness with P_{test} , and concomitantly with P_{out_1} and P_{out_2} , are outside the semi-circles. Tips inside the semi-circles break the triangle criterion at least with one outgroup population.

For the sake of computation simplicity of the scalar product involved in f_3 , let the distances be $f_2(P_{test}, P_{out_1}) = f_2(P_{test}, P_{out_2}) = 1$, in arbitrary units. Then, using the geometric interpretation of the scalar product, the orthogonal projection of the point P_a onto the segment $\overline{P_{out_1}P_{test}}$ is $f_3(P_{out_1}; P_{test}, P_a) = |\overline{A_1P_{test}}|$. We denote the length of a segment as $|\cdot|$. The three remaining projections are $|\overline{B_1P_{test}}|$, $|\overline{A_1P_{test}}|$ and $|\overline{B_2P_{test}}|$. These four quantities are related to f_3 as follows:

$$\text{using population } P_{out_1} \begin{cases} f_3(P_{out_1}; P_{test}, P_a) &= |\overline{A_1P_{test}}|, \\ f_3(P_{out_1}; P_{test}, P_b) &= |\overline{B_1P_{test}}|, \end{cases}$$

Figure 5: Outgroup f_3 -statistics test counterexample. Setup of the five population vector tips involved in. Population labels shortened for clarity.

and from the graph we get $f_3(P_{out_1}; P_{test}, P_a) > f_3(P_{out_1}; P_{test}, P_b)$. We conclude that population P_a is closer to P_{test} than P_b . Also

$$\text{using population } P_{out_2} \begin{cases} f_3(P_{out_2}; P_{test}, P_a) &= |\overline{A_2 P_{test}}| \\ f_3(P_{out_2}; P_{test}, P_b) &= |\overline{B_2 P_{test}}|, \end{cases}$$

and from the graph we get $f_3(P_{out_2}; P_{test}, P_a) < f_3(P_{out_2}; P_{test}, P_b)$. Therefore population P_b is closer to P_{test} than P_a , in contradiction with the conclusion above.

This counterexample illustrates how the choice of population outgroups may affect the outcome of the test. Conclusions based on outgroup f_3 -statistics test should be buttressed by further arguments.

6.2 Four-population test

Let us replace population P_3 in the outgroup f_3 -statistics test described above with a node so that the associated tree is $((P_{test}, P_i), P_a), P_b)$. The inner branch length is given by $f_4^B(P_{test}, P_i; P_a, P_b)$. Thus, provided $f_4^T(P_{test}, P_i; P_a, P_b) = 0$, the greater the value of $f_4^B(P_{test}, P_i; P_a, P_b)$ the closer the populations P_i and P_{test} . For instance, with populations P_1, P_2 , whenever $f_4^B(P_{test}, P_1; P_a, P_b) > f_4^B(P_{test}, P_2; P_a, P_b)$, then P_1 is closer to P_{test} than P_2 ; provided: $f_4^T(P_{test}, P_1; P_a, P_b) = f_4^T(P_{test}, P_2; P_a, P_b) = 0$. This is the instance illustrated in Figure 4b. Remind that $f_4^T(P_{test}, P_i; P_a, P_b) = f_4(P_{test}, P_i; P_a, P_b)$, whereas $f_4^B(P_{test}, P_i; P_a, P_b) = f_4(P_{test}, P_b; P_i, P_a)$.

Due to the fact that f_4 can be written down as a difference of two f_3 terms (see

property 10 in Appendix B) the same considerations about the outgroup f_3 -statistics test drawback apply.

7 Lack of treeness: Population admixture

The lack-of-treeness case is interpreted as a population admixture scenario in f -statistics. The statistic f_2 may be numerically computed, of course, but it has lost the interpretation as a genetic distance. Next, we first establish the algebraic description of a hybrid population, then give the population admixture criterion in terms of f -statistics and finally provide four different expressions to estimate the proportions of the admixture. Three of them are new results and one of them is related to the fixation index F_{ST} .

7.1 Population admixture: Mathematical description

Let P_{1_0}, P_{2_0} denote the parental and P_{x_0} the hybrid populations at the time when the admixture took place. Let P_1, P_2, P_x be the evolved populations whose allele frequencies we have experimental access to. These six vectors are represented in Figure 1a in S_f of dimension three. In the admixture process the populations P_{1_0} and P_{2_0} get combined to successfully form the new population P_{x_0} . More specifically, given a particular SNP of the hybrid, its allele frequency is formed by a fraction α of contributor P_1 and $1 - \alpha$ of contributor P_2 . And similarly for all the SNPs in the array, with the same fraction. We can then write

$$\vec{p}_{x_0} = \alpha \vec{p}_{1_0} + (1 - \alpha) \vec{p}_{2_0}, \quad (23)$$

at the time the hybrid population is born. This is just the kind of linear combination vector introduced in (2). The situation illustrated in Figure 1a shows that when the admixed population P_{x_0} is born, the three population vectors are coplanar. Subsequently, they wander in time as a consequence of genetic evolution describing three independent crinkly trajectories in S_f (curly green lines). The initial segment $\overline{P_{1_0}P_{x_0}P_{2_0}}$ becomes triangle $P_1P_xP_2$ after some time has elapsed. The goal is to estimate the fraction α from the three population vectors, $\vec{p}_1, \vec{p}_x, \vec{p}_2$. We will assume that the vector $\vec{p}_x$ of allele frequencies of the population P_x is well approximated by the linear combination

$$\vec{p}_x = \alpha \vec{p}_1 + (1 - \alpha) \vec{p}_2, \quad (24)$$

of the two contributor allele frequency vectors.

The idea of using a weighted average for one particular gene frequency of the mixed population appears already in [5, eq. (47)], [3, eq. (8.3)] and [4, eq. (1.17.1)].

Equation (24) will be a good approximation if the angle ϕ at vertex P_x in Figure 1b is close to 180° . ϕ is the separation angle between the two vectors $\vec{p}_x - \vec{p}_1$ and $\vec{p}_x - \vec{p}_2$, and can be computed with the aid of (3) irrespective of the dimension of S_f

$$\cos \phi = \frac{(\vec{p}_x - \vec{p}_1) \cdot (\vec{p}_x - \vec{p}_2)}{\sqrt{(\vec{p}_x - \vec{p}_1)^2 (\vec{p}_x - \vec{p}_2)^2}} = \frac{f_3(P_x; P_1, P_2)}{\sqrt{f_2(P_1, P_x) f_2(P_2, P_x)}}, \quad (25)$$

Figure 6: Relationship between f_2, f_3 , and the admixture proportion α according to (26) (black curve). Numerical simulation taken from [8, Fig.5B] (dashed red) with confidence interval (shaded).

for it is given in terms of scalar products. The limiting value $\cos \phi = -1$ corresponds to the birth of admixture. In section 5 we have discussed that $\cos \phi < 0$ corresponds to a population admixture scenario, whereas $\cos \phi > 0$ corresponds to a treeness one. The value $\cos \phi = 0$ is the separatrix of both scenarios. When $\cos \phi = 1$, the three vector tips are indeed co-linear but it is not an admixture situation because the order of vectors is not correct.

Given a set of putative contributors we might want to determine the most likely pair that generates a hybrid population. Since the function $\cos \phi$ is bounded, $-1 \leq \cos \phi \leq 1$, it can be used to rank the contributor pairs in the set.

7.2 Admixture test

In f -statistics, whenever three populations, say P_1, P_2, P_x , do not pass the triangle criterion for treeness the triplet is considered as candidate for population admixture. The formalism is binary in this respect. One (and only one) of the three possible f_3 values is negative; say, $f_3(P_x; P_1, P_2) < 0$. This is the so-called f_3 admixture test. We prove that P_x may be considered as a population admixture of P_1 and P_2 .

Let us compute $f_3(P_x; P_1, P_2)$ with P_x the hybrid described by the allele frequency

vector $\vec{p}_x$ in the linear combination (24)

$$\begin{aligned} f_3(P_x; P_1, P_2) &= \frac{1}{m} [\alpha \vec{p}_1 + (1 - \alpha) \vec{p}_2 - \vec{p}_1] \cdot [\alpha \vec{p}_1 + (1 - \alpha) \vec{p}_2 - \vec{p}_2] \\ &= \frac{1}{m} \alpha(\alpha - 1)(\vec{p}_1 - \vec{p}_2)^2 = \alpha(\alpha - 1)f_2(P_1, P_2) \leq 0, \end{aligned} \quad (26)$$

because $f_2 \geq 0$, and $0 \leq \alpha \leq 1$ by hypothesis of population admixture. A similar computation yields

$$\begin{aligned} f_3(P_1; P_2, P_x) &= (1 - \alpha)f_2(P_1, P_2) \geq 0, \\ f_3(P_2; P_1, P_x) &= \alpha f_2(P_1, P_2) \geq 0, \end{aligned} \quad (27)$$

and so the two remaining f_3 values are positive.

We can summarize the conditions for population admixture without explicit mention to α . For a population of the form (24)

$$\begin{aligned} f_3(P_1; P_2, P_x) + f_3(P_2; P_1, P_x) &= f_2(P_1, P_2), \\ f_3(P_x; P_1, P_2) \leq 0, \quad f_3(P_1; P_2, P_x) \geq 0, \quad f_3(P_2; P_1, P_x) \geq 0. \end{aligned} \quad (28)$$

For fixed populations P_1 and P_2 , the term $f_2(P_1, P_2)$ in (26) is a constant and the functional dependency between f_3 and the admixture proportion α is given by the parabola in Figure 6 (solid black curve). Interestingly, in [8, Figure 5B] a similar parabolic curve is obtained from a numerical simulation in which f_3 has been expressed in terms of expected coalescence times and parameters have been chosen so that $f_3 < 0$. A few measurements by hand on that plot yielded the dashed curve with the shaded area as confidence interval in Figure 6.

7.3 Population admixture: Determination of proportions

Given the population vectors $\vec{p}_x, \vec{p}_1, \vec{p}_2$, four formulas are given to estimate the proportion α in (24). The first one, based on a least squares procedure, is new. The second one relies on orthogonal projections on an arbitrary direction in S_f determined by two auxiliary populations and gives rise to the so-called f_4 -ratio statistics. The last two ones are simplifications using only one auxiliary population and are new.

7.3.1 Approach 1: Least squares

The least squares solution to (23) is readily obtained using vector algebra. In this simple case no matrix algebra is needed. Let us rewrite (24) as

$$\alpha(\vec{p}_1 - \vec{p}_2) = \vec{p}_x - \vec{p}_2. \quad (29)$$

The scalar product by $\vec{p}_1 - \vec{p}_2$ of both sides yields

$$\alpha_{ls} = \frac{(\vec{p}_1 - \vec{p}_2) \cdot (\vec{p}_x - \vec{p}_2)}{(\vec{p}_1 - \vec{p}_2)^2} = \frac{f_3(P_2; P_1, P_x)}{f_2(P_1, P_2)}, \quad (30)$$

which is the least squares estimate for α . The conventional confidence interval reads

$$\alpha_{ls} \pm \rho \delta_{ls},$$

$$\delta_{ls}^2 = \frac{[\vec{p}_x - \alpha_{ls} \vec{p}_1 - (1 - \alpha_{ls}) \vec{p}_2]^2}{m(\vec{p}_1 - \vec{p}_2)^2}. \quad (31)$$

For a 95% confidence interval and m large enough, $\rho = 1.96$.

Figure 1b provides the geometric interpretation of the leftmost fraction in (30). The denominator is a squared vector modulus: $|\vec{p}_1 - \vec{p}_2|^2$. According to the geometric interpretation of the scalar product $u = (\vec{p}_x - \vec{p}_2) \cdot (\vec{p}_1 - \vec{p}_2) / |\vec{p}_1 - \vec{p}_2|$ is the length of the projection of segment $\overline{P_x P_2}$ along the direction of vector $\vec{p}_1 - \vec{p}_2$, and therefore $u = |\overline{P_a P_2}|$. The estimate in (30) is then the ratio $\alpha_{ls} = u/v$, with $v = |\vec{p}_1 - \vec{p}_2|$. At the time population P_x emerges, i.e. when $P_x \equiv P_{x_0}$, the estimate is exact.

7.3.2 Approach 2: Projection in S_f using two auxiliary populations

Consider two arbitrary populations P_a and P_b . The difference vector $\vec{p}_a - \vec{p}_b$ defines an arbitrary direction in S_f . The scalar product by $\vec{p}_a - \vec{p}_b$ of both sides in (29) gives an estimate involving five populations

$$\alpha_p = \frac{(\vec{p}_a - \vec{p}_b) \cdot (\vec{p}_x - \vec{p}_2)}{(\vec{p}_a - \vec{p}_b) \cdot (\vec{p}_1 - \vec{p}_2)} = \frac{f_4(P_a, P_b; P_x, P_2)}{f_4(P_a, P_b; P_1, P_2)}. \quad (32)$$

According to the geometric interpretation of the scalar product given in Section 2, the numerator is a length proportional to the orthogonal projection of vector $\vec{p}_x - \vec{p}_2$ along the direction defined by vector $\vec{p}_a - \vec{p}_b$. Similarly for the denominator with the difference vector $\vec{p}_1 - \vec{p}_2$. The estimate α_p is then the ratio of both lengths, provided the denominator is non-vanishing. This result is referred to as f_4 -ratio statistic and populations P_a, P_b are outgroups. Notice that, from an algebraic point of view, there is no constraint on the choice of populations P_a, P_b . The sub-index in α_p refers to the orthogonal projection involved in the method. Since we are not in a treeness scenario, no argument permutation is made in the computation of f_4 .

The interpretation of (32) is in Figure 7 where the five populations are represented by their allele frequency vectors in dimension three. The vector $\vec{p}_a - \vec{p}_b$ defines an arbitrary direction in S_f . Segments $\overline{1X}$ and $\overline{2X}$ are projected onto $\overline{ab}$. The length ratio $|\overline{aa_\perp}|$ over $|\overline{ab_\perp}|$ is α_p in (32). In practice, α_p is computed for a number of population pairs P_a, P_b , which therefore define different directions in S_f . After that, a re-sampling procedure provides a confidence interval [7].

Figure 7: Geometric representation of the elements involved in the admixture f_4 -ratio statistics (32) for S_f of dimension three. P_a and P_b are reference populations. P_x is studied as a possible mixture of P_1 and P_2 . The five frequency allele vectors have been drawn (red). In the triangle conformed by populations $\{P_1, P_2, P_x\}$ the angle at vertex P_x is obtuse. The projections on the horizontal plane are just for helping with the perspective. Segments $\overline{1X}$ and $\overline{2X}$ have been parallel transported to point a in order to illustrate their orthogonal projection onto the line $\overline{ab}$. The quotient of length $\overline{aa_\perp}$ over length $\overline{ab_\perp}$ is just α_p in (32). Population labels shortened for clarity.

Figure 8: Estimates of α_{ls} in (30), α_p in (32), α'_p in (33) and α''_p in (34), as a function of $\cos \phi$ in (25) for five different values of α used to generate the hybrids, which are indicated by the horizontal grid lines. The ε value is related to the time since the hybridization event: the greater the ε value the more time elapsed since the mixing event. Dashed lines are only for visual help.

7.3.3 Approach 3: Projection in S_f using one auxiliary population

If in (32) we choose either $P_b = P_1$ or $P_b = P_2$ we get two further formulas involving only one arbitrary auxiliary population

$$\alpha'_p = \frac{(\vec{p}_a - \vec{p}_1) \cdot (\vec{p}_x - \vec{p}_2)}{(\vec{p}_a - \vec{p}_1) \cdot (\vec{p}_1 - \vec{p}_2)} = \frac{f_4(P_1, P_a; P_x, P_2)}{f_3(P_1; P_2, P_a)} \quad (33)$$

$$\alpha''_p = \frac{(\vec{p}_a - \vec{p}_2) \cdot (\vec{p}_x - \vec{p}_2)}{(\vec{p}_a - \vec{p}_2) \cdot (\vec{p}_1 - \vec{p}_2)} = \frac{f_3(P_2; P_x, P_a)}{f_3(P_2; P_1, P_a)}. \quad (34)$$

These equations may be seen as a simplification the f_4 -ratio (32) in the sense that only one auxiliary population is involved.

7.4 An illustrative example

In order to grasp the similarities and differences among the estimates α_{ls} , α_p , α'_p and α''_p , we have carried out the following numerical simulations.

From the population dataset *v37.2.1240K_HumanOrigins* from *Reich Lab* [13] we have taken two parental populations, $P_1 = \text{English}$ and $P_2 = \text{Yoruba}$, and build up the admixed population (23), $P_{x_0} = \alpha P_1 + (1 - \alpha) P_2$ for several values of α .

Figure 9: f_4 -ratio plane for a simulation with initial $\alpha = 0.3$ in (23). The axes of the plot represent the numerator and denominator of (32). Black dots stand for different evaluations with the 36 outgroup pairs from the boxed list. The slope of the linear fit (dashed line), with fix intercept at the origin, is the average estimate for α_p in (32). Fit lines to pluses and crosses (not showed) give averages for α'_p in (33) and α''_p in (34), which involve only one auxiliary population. The solid green line has slope $\alpha = 0.3$ and is given for the sake of comparison.

To simulate finite time evolution we have added some noise of amplitude ε to the P_{x_0} allele frequencies. The rationale is that the greater the ε value the longer the time elapsed since population admixture event. The parent populations stay invariant in the simulation for the sake of simplicity. The idea is just to take the vector $\vec{p}_{x_0}$ off the plane defined by the two contributor vectors. Besides, nine auxiliary populations have been chosen from the same dataset, namely: Han, Papuan, Surui, Karitiana, Aleut, Ju hoan North, Gujarati, Iranian and Australian. The simulations have been carried out as follows.

The time evolved hybrid P_x is simulated for $\alpha = 0.1, 0.2, \dots, 0.5$, and $\varepsilon = 0.1, \dots, 0.4$. The outcomes of the simulations are in Figures 8 and 9.

As regards *Approach 1*, given the three populations P_1, P_2, P_x , equation (30) determines univocally α_{ls} . In Figure 8 the open circles stand for the α_{ls} estimates as a function of the $\cos \phi$ value reached. Irrespective of the ε value, the error estimate (31) is in the range $0.004 < \rho \delta_{ls} < 0.008$. Remind that $-1 < \cos \phi < 0$ is the admixture condition and $\cos \phi = -1$ at admixture birth. For fixed ε value, an unbalanced admixture heads toward a genetic drift scenario ($\cos \phi > 0$) faster than a balanced one in this type of simulation.

To deal with *Approach 2* we read (32) as a linear relationship between $f_4(P_a, P_b; P_x, P_{2_0})$ and $f_4(P_a, P_b; P_{1_0}, P_{2_0})$, instead of a ratio. Then a plot of pairs of both quantities, evaluated for different populations P_a, P_b , should appear as points on a straight line. This behaviour is observed in Figure 9 (black dots) as well as for the other α values used in the simulation. In the present case the slope of the fit (dashed line), which was constrained to intercept the origin, gives $\alpha_p = 0.271 \pm 0.001$. The solid line has slope 0.3 and is given for the sake of comparison. The whole set of α_p estimates is in Figure 8 (open squares).

The outcomes of *Approach 3*, which involves only one auxiliary population, are also in Figure 9. Red pluses and crosses stand for the cases where $P_{English}$ or P_{Yoruba} act as auxiliary population, respectively. The estimated slopes are $\alpha'_p = 0.270 \pm 0.001$ and $\alpha''_p = 0.310 \pm 0.006$ (fit line not showed). Further results are in Figure 8 (solid triangles).

8 Population admixture: Coalescence times, pairwise sequence differences and fixation index F_{ST}

8.1 Coalescence times

A result reported in [8], provides an alternative admixture proportion estimate on the basis that $f_2(P_i, P_j)$ can be written in terms of expected coalescence times T_{ij} of populations P_i and P_j to a common ancestor P_a . The formula in [8] is given under the assumption that all populations involved are sampled at the same time

$$\alpha = \frac{T_{a2} - T_{ax}}{T_{a2} - T_{a1}}. \quad (35)$$

According to [8], when the number of individuals in populations is large

$$f_2(P_1, P_2) = \theta \left(T_{12} - \frac{T_{11} + T_{22}}{2} \right), \quad (36)$$

with $\theta/2$ the constant rate of mutation on each branch of the tree. Writing explicitly down the f_4 -ratio (32) in terms of f_2 , and according to (13) (see also B)

$$\alpha_p = \frac{[f_2(P_a, P_2) + f_2(P_b, P_x)] - [f_2(P_a, P_x) + f_2(P_b, P_2)]}{[f_2(P_a, P_2) + f_2(P_b, P_1)] - [f_2(P_a, P_1) + f_2(P_b, P_2)]}. \quad (37)$$

Substituting (36)

$$\alpha_p = \frac{(T_{a2} + T_{bx}) - (T_{ax} + T_{b2})}{(T_{a2} + T_{b1}) - (T_{a1} + T_{b2})}. \quad (38)$$

Both formulas are not so different if the sums in parentheses are interpreted as averages of two coalescence times.

8.2 Pairwise sequence differences

A second outcome in [8] is based on the relationship between f_2 and the average number of pairwise sequence differences π_{ij} between populations P_i and P_j which reads

$$f_2(P_1, P_2) = \pi_{12} - \frac{\pi_{11} + \pi_{22}}{2}. \quad (39)$$

Substitution in (37) yields

$$\alpha_p = \frac{(\pi_{a2} + \pi_{bx}) - (\pi_{ax} + \pi_{b2})}{(\pi_{a2} + \pi_{b1}) - (\pi_{a1} + \pi_{b2})}, \quad (40)$$

to be compared with the formula in [8]

$$\alpha = \frac{\pi_{a2} - \pi_{ax}}{\pi_{a2} - \pi_{a1}}. \quad (41)$$

The comparison is similar to that of coalescence times. According to numerical simulations in [8], both estimates seem to perform well although for low admixture proportion, the latter performs better.

8.3 Fixation index F_{ST}

In [4, Sec.1.17], the populations are referred to a multidimensional *gene-frequency space* that clearly plays a role similar to S_f . Moreover, the triangle criterion is qualitatively described and the authors provide, without disclosing, a formula for the population admixture proportion in terms of the so-called F_{ST} genetic distance. We prove that such a formula and α_{ls} in (30) are related results.

The formula in [4] for the admixture proportion m reads

$$m = \frac{1}{2} + \frac{f_{BM} - f_{AM}}{2f_{AB}}, \quad (42)$$

where f_{AB} stands for the F_{ST} distance between contributors A and B ; and f_{AM} , f_{BM} , the distances between each of the two contributors and the hybrid M . Here m is the admixture proportion of population A . In the case of two populations, the difference between F_{ST} and f_2 is only an approximately constant normalization factor. Consequently, the formula (42) can be re-written as

$$m \simeq \frac{1}{2} + \frac{f_2(B, M) - f_2(A, M)}{2f_2(A, B)}. \quad (43)$$

Now, using (16) we get

$$m \simeq \frac{f_3(B; A, M)}{f_2(A, B)}, \quad (44)$$

and, by (30), the conclusion $m \simeq \alpha_{ls}$ follows.

9 Discussion and conclusions

We have introduced the f -statistics as the set of constraints needed for a genetic distance matrix to be associated to a phylogenetic tree of four populations. As far as f_2 can be interpreted as a true genetic distance in S_f , the concomitant scenario is tree-like-ness. Else, the case is interpreted as population admixture.

Populations are described as vectors in S_f evolving in time. A long enough time lapse may drive an admixture scenario into an evolutionary one, and the way around. It suffices that, as a consequence of evolution, the wandering of population vectors transforms an obtuse-angled triangle in S_f into an acute-angled one, or viceversa. An instance of this class [7, 8] points out that high population-specific drift in the admixed population could prevent a negative value for the f_3 test. In Appendix A we obtain a conservative formal lower bound to the linear distance that vectors in S_f may travel before a population admixture can be detected as an evolutionary process.

The outgroup f_3 -statistics test and the four-population test rank a population set according to a genetic distance. These tests should be preceded by the triangle criterion in order to rule out populations not related by an evolutionary process. We have noted that contradictory rankings may appear between outcomes obtained with different outgroups.

The four formulas provided to estimate the proportion of contributors in a hybrid population rely on the use of a linear combination of population vectors in S_f . This idea can be traced back to [4] where the triangle criterion is already qualitatively described. The new estimate (30) is obtained from an orthogonal projection on a fixed direction defined by the admixture contributors. It can be also interpreted as a least squares approximation and is related to a result based on the F_{ST} genetic distance. The way we have derived the formula for the f_4 -ratio (32) is based on orthogonal projections onto an arbitrary direction in S_f (defined by the difference vector of the two auxiliary populations). The two remaining new formulas are obtained by similar projections in S_f but using only one auxiliary population. The simulations carried out illustrate the behaviour of the four estimates. It seems that α_{ls} gives balanced estimates whereas α_p and α_p'' tend to underestimate and α_p' to overestimate, albeit more systematic simulations are needed to base these conclusions.

The f -statistics formalism has introduced an important tool to facilitate the study of intricate population structures. We have presented a new derivation from a geometry-based scheme, as well as some new results and relationships with other formalisms, that may ease the use and interpretation of the f -formalism in practice.

Acknowledgements: We are indebted to J.M. Bordes for illuminating discussions about the geometry of f -statistics. GOG thanks The Leverhulme Trust for financial support. JAO work was partially supported by the Spanish MINECO (grant numbers AYA2016-81065-C2-2 and PID2019-109592GB-I00).

Figure 10: Scheme to find a lower bound to reliability of f_3 admixture test.

A Lower bound to reliability of f_3 admixture test

A lower bound may be given to the Euclidean distance r travelled in S_f by three points representing an admixed population and two contributors, respectively, before they reach the right-angled triangle geometric configuration. If the rate at which allele frequencies change in the dataset were known, then an estimate of the time t elapsed since the birth of the admixture could be provided on the basis that r is proportional to $\sqrt{t}$, if we assume that the vectors describe Brownian trajectories in S_f . These bounds are of interest to assess the validity of the f_3 admixture test.

In Figure 10, the point A' represents the admixture population and A and A'' stand for the two equal-weight contributors to the linear combination that defines the birth of A' . The goal is to give just a lower bound, so we do not deal with a generic proportion. Besides, A and A'' are assumed one unit apart and so define the genetic distance scale. In S_f those points evolve in time describing a Brownian-like trajectory (green lines) and the linear distance reached, r , is proportional to $\sqrt{t}$. After a lapse of time, the trajectories (on average) cross some point of the inner circumferences. For instance, the trajectory born at initial time at A wanders in S_f until it crosses the circumference at B , and at a later time it will reach the point C .

We are interested in the situation where, given a radius, the angle at vertex B' is as small as possible. We assume the favourable case where the positions B, B', B'' , are simultaneously reached. This is the configuration depicted, where lines $\overline{B'B}$ and $\overline{B'B''}$, are tangents to a circumference. This is only an intermediate situation. We want to ascertain the radius r at which the angle at vertex B' becomes right as in C' . The proof relies on the fact that the triangles ACD and $DA'C'$ are identical. Thus, $|\overline{DA'}| = r$ and $|\overline{AD}| = |\overline{DC'}| = \sqrt{2}r$. Since $|\overline{AA'}| = 1$ we have then $2(|\overline{AD}| + |\overline{DA'}|) = 2r(1 + \sqrt{2}) = 1$, and therefore $r = (\sqrt{2} - 1)/2 \simeq 0.207$. Of course, this picture corresponds to an ex-

treme situation in which the wandering of the three trajectories ends up simultaneously at points C , C' and C'' . This is why 21% of the initial distance between contributors is just a lower bound. After that, if the dynamics leads the geometry to an acute-angled triangle the f_3 admixture test will fail.

An algebraic proof is also possible minimizing the expression of r with respect to several variables but the algebra is quite more involved. It yields the very same result for the radius r .

B Some useful algebraic identities

A listing of algebraic identities among statistics may be useful. The following one does not exhaust the list of possible relationships. All can be straightforwardly proved, for instance by expressing f_3 and f_4 in terms of f_2 . Some of them are mere symmetry properties under exchange of arguments.

1. $f_2(P_1, P_2) = f_2(P_2, P_1)$.
2. $f_3(P_a; P_1, P_2) = f_3(P_a; P_2, P_1)$.
3. $f_3(P_a; P_1, P_2) = \frac{1}{2}[f_2(P_1, P_a) + f_2(P_2, P_a) - f_2(P_1, P_2)]$.
4. $f_3(P_1; P_2, P_a) + f_3(P_2; P_1, P_a) = f_3(P_1; P_2, P_b) + f_3(P_2; P_1, P_b)$.
5. $f_4(P_a, P_b; P_1, P_2) = f_4(P_1, P_2; P_a, P_b) = f_4(P_b, P_a; P_2, P_1)$.
6. $f_4(P_a, P_b; P_1, P_2) = -f_4(P_a, P_b; P_2, P_1) = -f_4(P_b, P_a; P_1, P_2)$.
7. $f_4(P_a, P_b; P_1, P_2) = \frac{1}{2}[f_2(P_a, P_2) + f_2(P_b, P_1) - f_2(P_a, P_1) - f_2(P_b, P_2)]$.
8. $f_4(P_1, P_2; P_3, P_4) = f_4(P_1, P_3; P_2, P_4) + f_4(P_1, P_4; P_3, P_2)$.
9. $f_3(P_1; P_2, P_b) + f_3(P_2; P_1, P_a) = f_2(P_1, P_2) + f_4(P_a, P_b; P_1, P_2)$.
10. $f_4(P_a, P_b; P_1, P_2) = f_3(P_a; P_b, P_2) - f_3(P_a; P_b, P_1) = f_3(P_b; P_a, P_1) - f_3(P_b; P_a, P_2)$.
11. $f_4(P_a, P_b; P_1, P_2) = f_4(P_a, P_b; P_1, P_3) + f_4(P_a, P_b; P_3, P_2)$

They identities above are valid in both evolutionary and population admixture contexts. In an admixture scenario: $\vec{p}_x = \alpha\vec{p}_1 + (1 - \alpha)\vec{p}_2$, we have

1. $f_3(P_x; P_1, P_2) = \alpha(\alpha - 1)f_2(P_1, P_2) \leq 0$.
2. $f_3(P_1; P_2, P_x) = (1 - \alpha)f_2(P_1, P_2) \geq 0$.
3. $f_3(P_2; P_1, P_x) = \alpha f_2(P_1, P_2) \geq 0$.
4. $f_3(P_1; P_2, P_x) + f_3(P_2; P_1, P_x) = f_2(P_1, P_2)$.

$$5. \alpha = f_4(P_a, P_b; P_x, P_2) / f_4(P_a, P_b; P_1, P_2).$$

$$6. f_4(P_a, P_b; P_x, P_2) = \alpha f_4(P_a, P_b; P_1, P_2).$$

$$7. f_4(P_a, P_b; P_x, P_c) = \alpha f_4(P_a, P_b; P_1, P_c) + (1 - \alpha) f_4(P_a, P_b; P_2, P_c).$$

Eventually, remind that f_4 has a twofold role in a phylogenetic scenario:

$$1. f_4^T(P_1, P_2; P_3, P_4) = f_4(P_1, P_2; P_3, P_4).$$

$$2. f_4^B(P_1, P_2; P_3, P_4) = f_4(P_1, P_4; P_2, P_3).$$

References

- [1] J.P. Barthelemy and A. Guenoche. *Trees and proximity representations*. Wiley, 1991.
- [2] L.L. Cavalli-Sforza. Population structure and human evolution. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)*, 164:362–379, 1966.
- [3] L.L. Cavalli-Sforza and W.F. Bodmer. *The genetics of human populations*. Dover Publications, 1999.
- [4] L.L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The history and geography of human genes*. Princeton University Press, 1994.
- [5] L.L. Cavalli-Sforza and A. Piazza. Analysis of evolution: Evolutionary rates, independence and treeness. *Theoretical Population Biology*, 8:127–165, 1975.
- [6] A.M. Harris and M. DeGiorgio. Admixture and Ancestry Inference from Ancient and Modern Samples through Measures of Population Genetic Drift. *Human Biology*, 89:21, 2017.
- [7] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192:1065–93, 2012.
- [8] B.M. Peter. Admixture, Population Structure, and F-Statistics. *Genetics*, 202:1485–501, 2016.
- [9] M. Raghavan, P. Skoglund, K.E. Graf, M. Metspalu, A. Albrechtsen, I. Moltke, S. Rasmussen, T.W. Stafford Jr, L. Orlando, E. Metspalu, M. Karmin, K. Tambets, S. Rootsi, R. Mägi, P.F. Campos, E. Balanovska, O. Balanovsky, E. Khusnutdinova, S. Litvinov, L.P. Osipova, S.A. Fedorova, M.I. Voevoda, M. DeGiorgio, T. Sicheritz-Ponten, S. Brunak, S. Demeshchenko, T. Kivisild, R. Villems, R. Nielsen, M. Jakobsson, and E. Willerslev. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, 505:87–91, 2014.

- [10] D. Reich, K. Thangaraj, N. Patterson, A.L. Price, and L. Singh. Reconstructing Indian population history. *Nature*, 461:489–494, 2009.
- [11] J.M.S. Simões-Pereira. A note on the tree realizability of a distance matrix. *Journal of Combinatorial Theory*, 6:303–310, 1969.
- [12] S. Soraggi and C. Wiuf. General theory for stochastic admixture graphs and F-statistics. *Theoretical Population Biology*, 125:56–66, 2019.
- [13] David Reich Lab (Harvard University). <https://reich.hms.harvard.edu/downloadable-genotypes-present-day-and-ancient-dna-data-compiled-published-papers>. 2020.
- [14] S. Wright. The genetical structure of populations. *Annals of Eugenics*, 15:323–354, 1951.