OXFORD GENETICS

GENETICS, 2021, 217(4), iyaa045
DOI: 10.1093/genetics/iyaa045
Advance Access Publication Date: 8 January 2021
Investigation

Assessing the performance of qpAdm: a statistical tool for studying population admixture

Éadaoin Harney [1,2,3,4,*] Nick Patterson,[4] David Reich,[3,4,5,6] and John Wakeley[1]

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[2]The Max Planck-Harvard Research Center for the Archaeoscience of the Ancient Mediterranean, Cambridge, MA 02138, USA
[3]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[4]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[5]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[6]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

*Corresponding author: eadaoinharney@gmail.com

Abstract

qpAdm is a statistical tool for studying the ancestry of populations with histories that involve admixture between two or more source populations. Using qpAdm, it is possible to identify plausible models of admixture that fit the population history of a group of interest and to calculate the relative proportion of ancestry that can be ascribed to each source population in the model. Although qpAdm is widely used in studies of population history of human (and nonhuman) groups, relatively little has been done to assess its performance. We performed a simulation study to assess the behavior of qpAdm under various scenarios in order to identify areas of potential weakness and establish recommended best practices for use. We find that qpAdm is a robust tool that yields accurate results in many cases, including when data coverage is low, there are high rates of missing data or ancient DNA damage, or when diploid calls cannot be made. However, we caution against co-analyzing ancient and present-day data, the inclusion of an extremely large number of reference populations in a single model, and analyzing population histories involving extended periods of gene flow. We provide a user guide suggesting best practices for the use of qpAdm.

Keywords: qpAdm; AdmixTools; admixture; simulation

Introduction

The last decade has experienced a revolution in the amount of genetic data available to study from both living and ancient organisms. Questions about the origins of populations have increased in complexity, often in an effort to understand histories that involve admixture, which are incompatible with traditional tree-like models of relatedness. qpAdm is a tool that can be used to understand the history of admixed populations in both human and nonhuman species. It has been applied to study the genetic history of human populations that would otherwise remain mysterious. For instance, the use of qpAdm was vital to studying the ancestry of the Late Bronze Age Greek culture of the "Mycenaeans" (Lazaridis *et al.* 2017)—the subjects of the Iliad and Odyssey. However, little has been done to assess qpAdm's performance under both simple and complex scenarios.

A potential drawback of many population genetic tools for studying the population history of specific groups is that they require the historical relationships of all other populations included in the analysis to be explicitly modeled (Patterson *et al.* 2012; Pickrell and Pritchard 2012). This underlying phylogeny is either specified by the user (as in qpGraph) or is calculated during the analysis (as in TreeMix). This may lead to biases or errors in inferences about admixture if mistakes are made when specifying the underlying relationships of nontarget populations (Lipson 2020). This requirement for a complete and accurate population history is especially difficult to satisfy in studies that utilize ancient DNA, which increasingly attempt to use genetic data of limited quality to analyze nuanced differences between closely related groups. However, even in cases where it is difficult to reconstruct a full population history, it is often possible to examine patterns of shared genetic drift between various populations in order to learn about their relationships to one another (Patterson *et al.* 2012). qpAdm exploits this information, enabling admixture models to be tested for plausibility and admixture proportions to be estimated.

The theory underlying qpAdm, which was introduced in Haak *et al.* (2015), builds upon a class of statistics known as *f*-statistics (Patterson *et al.* 2012). *f*-statistics analyze patterns of allele frequency correlations among populations in order to determine whether their population histories can be described using strictly tree-based models, or if more complex models, such as those involving admixture, are required to explain the genetic data. *f*-statistics have been widely used in the population genetic literature and their behavior is well understood (Reich *et al.* 2009, 2012; Patterson *et al.* 2012; Peter 2016; Soraggi and Wiuf 2019; Lipson 2020). qpAdm harnesses the power of *f*-statistics to determine

Received: October 14, 2020. **Accepted:** December 11, 2020

whether a population of interest (a target population) can be plausibly modeled as descending from a common ancestor of one or more source populations. For example, in a model with two source populations, qpAdm tests whether the target population is the product of a two-way admixture event between these source populations. The method requires users to specify a list of target and source populations and a list of additional reference populations which provide information about the relationships among the target and source populations.

Often, the target and source populations are referred to as "left" populations while the reference populations are called "right" populations. This is due to their positions as arguments in the f_4-statistics, *i.e.* f_4(target, source; ref_1, ref_2). Here, we prefer "target," "source" and "reference" and minimize our use of "left and "right." In addition, reference or "right" populations have previously been referred to as "outgroup" populations, but we also avoid this term because it suggests that reference populations should be outgroups in phylogenetic sense (*i.e.*, equally closely related to all "left" populations). In fact, if all reference populations are symmetrically related to all source populations in this way, qpAdm will not produce meaningful results. The method requires differential relatedness, meaning that at least some reference populations must be more closely related to some source populations than to others. We illustrate this in *Methods and Results*, and further describe the assumptions qpAdm makes about relationships among target, source, and reference populations with examples in the qpAdm User Guide in Supplementary Material S1.

Inferences about admixture in qpAdm are made by fitting a series of hypothetical models to a matrix of f_4-statistics computed from the data. We describe the methodological details in Supplementary Material S2, which may be summarized briefly as follows. qpAdm compares possible scenarios involving admixture to the unconstrained alternative. Each hypothesized model of admixture assumes that the target population was produced as a mixture of ancestral populations which are direct ancestors of the specified source populations. Allele frequencies in the ancestral target population are constrained to be linear combinations of allele frequencies in the ancestral source populations. In the unconstrained model, the ancestral target population is allowed to vary freely. For each constrained model, qpAdm gives a *P*-value which is used to determine whether the proposed admixture scenario is plausible or whether it is rejected in favor of the unconstrained alternative. The *P*-value is calculated using a likelihood ratio test in which the constrained model is the null hypothesis and the unconstrained model is the alternative hypothesis. A simple example in which the constrained model would obviously be rejected is when the putative target population is actually an outgroup to all source populations.

While qpAdm has been applied in numerous studies (*e.g.*, Lazaridis *et al.* 2016; Haber *et al.* 2017; Lazaridis *et al.* 2017; Skoglund *et al.* 2017; de Barros Damgaard *et al.* 2018a, 2018b; Hajdinjak *et al.* 2018; Harney *et al.* 2018; Olalde *et al.* 2018; Narasimhan *et al.* 2019), producing results that are consistent with those of other population genetic methods, very little has been done to assess the performance of the tool when the population history is known (*i.e.*, using simulated data). The only simulation-based analysis that has been previously conducted examined whether simulated populations—generated according to the model fitted by qpAdm by resampling data using the source populations and estimated admixture proportions—behaved similarly to the real target population in further statistical analyses (Lazaridis *et al.* 2017). Although this limited example supports the use of qpAdm in population genetic analyses, it did not address any of the potential limitations of the method. Here we use simulated genomic data to study the distributions of *P*-values and estimated admixture proportions from qpAdm, the potential of qpAdm to distinguish optimal from nonoptimal models of admixture for a given set of samples (where optimal models are those that do not violate any of the assumptions that qpAdm makes about the relationship between target, source and reference populations), and the performance of qpAdm in the face of more challenging demographic scenarios.

The chief purpose of qpAdm is to identify a subset of plausible models of a population's ancestry from a larger set of possible models. Users propose a series of possible models, each with different combinations of source populations contributing to a given target population, then eliminate implausible models. Models are deemed implausible if their estimated admixture proportions fall outside the biologically relevant range (0–1) or if they are rejected statistically by having a small *P*-value. Again, the proposed models are the null models in the likelihood ratio tests described above. The resulting set of plausible models are the ones which are *not* rejected, meaning they have *P*-values *greater* than the chosen significance level, which is usually 5%. As this is a nonstandard use of *P*-values, in Box 1 we provide a simple illustration of an analogous technique for identifying plausible models for the (unknown) probability of heads for a coin. We emphasize that Box 1 is not meant to illustrate the complexities of model specification and choice in qpAdm, which involves the specification of target, source and reference populations and the additional estimation of admixture proportions. In a later section (Comparing admixture models) we describe an approach for identifying optimal admixture models among the several possible models which might be deemed plausible by qpAdm.

Identical to standard statistical methods, this sort of approach may be considered to be working properly if the *P*-values generated by qpAdm follow a uniform distribution when the correct admixture model is specified. In this case, the correct model will be rejected 5% of the time when a threshold of $P < 0.05$ is applied. For other plausible but less-optimal models, the distribution of *P*-values is not expected to be uniform but should have an appreciable chance of being above the 5% cutoff. The distribution of *P*-values for implausible or incorrect models should fall largely below the 5% cutoff. While experience suggests that the *P*-values generated by qpAdm are reasonably consistent with these expectations, in this work we perform the first systematic test of these ideas.

Similarly, although the estimated admixture proportions calculated by qpAdm appear generally consistent with values generated using other statistics, the accuracy of these estimates have never been rigorously tested. Of particular interest is the accuracy of these estimates when calculated on low quality data, as qpAdm is often applied to the study of ancient DNA, which is characteristically low coverage, may have a high rate of missing data, and is susceptible to deamination of cytosine nucleotides (manifesting in sequence data as cytosines being misread as thymines). Furthermore, ancient DNA is often subject to a complex ascertainment process that could potentially bias statistical analyses. We explore the impact of each of these factors on the admixture proportions estimated by qpAdm.

Additionally, while one of the main features of qpAdm is its ability to distinguish between optimal and nonoptimal models for a group's population history, there are no formal recommendations about what strategy should be employed to compare models. We therefore consider two of the most commonly

Box 1 Coin flipping analogy

Imagine that we wish to know which of several possible models for the probability of heads best describes the behavior of a coin. The actual value is unknown and the coin may be unfair. To illustrate how *P*-values are used in qpAdm, we might specify a set of possible models, for instance with probabilities of heads constrained to fall within bins of width 0.1 (Table 1, left column). In order to determine which of these models are plausible for the coin, we flip it multiple times and count the number of heads we observe. The probability of a particular outcome would then be given by the binomial distribution.

By analogy with qpAdm, we could assess the plausibility of each model using a generalized likelihood ratio test of each constrained model against the unconstrained alternative ($0 \leq P \leq 1$). The models we are interested in are the null models in these tests, and we consider them as plausible if they are not rejected. Thus, qpAdm identifies a set of plausible models using what would be called Type II error in a standard statistical test.

For example, if we flip the coin 100 times and it comes up heads 64 times, then using a *P*-value threshold of 0.05 we can eliminate seven of the ten models for the probability of heads of the coin (middle column Table 1). Three models remain plausible. By increasing the number of flips to 1000, and assuming we observe 646 heads, we rule out two more model, corresponding to bins (0.5,0.6] and (0.7,0.8]. This leaves only the model $0.6 < P \leq 0.7$ as plausible for the probability of heads of our coin.

Table 1 Using a generalized likelihood-ratio test of H_0: $P_L < P \leq P_U$ to identify plausible models for the probability of heads of a coin

Models for the probability of heads	*P*-values for a generalized-likelihood ratio test of H_0: $P_L < P \leq P_U$ against unconstrained alternative	
Number of flips	100	1,000
Number of heads	64	646
$0.0 \leq P \leq 0.1$	<0.001	<0.001
$0.1 < P \leq 0.2$	<0.001	<0.001
$0.2 < P \leq 0.3$	<0.001	<0.001
$0.3 < P \leq 0.4$	<0.001	<0.001
$0.4 < P \leq 0.5$	0.005	<0.001
$0.5 < P \leq 0.6$	**0.411**	0.003
$0.6 < P \leq 0.7$	**1.000**	**1.000**
$0.7 < P \leq 0.8$	**0.198**	<0.001
$0.8 < P \leq 0.9$	<0.001	<0.001
$0.9 < P \leq 1.0$	<0.001	<0.001

Models that produce *P*-values ≥0.05 are indicated in bold.

employed strategies for model comparison, one using a consistent "base" set of reference populations and the other using a "rotating" set of reference populations that is dependent upon the particular source populations included in each model, highlighting their potential benefits and weaknesses.

Finally, we conclude by exploring nonstandard cases where the expected behavior of qpAdm is poorly understood, such as the impact of including reference populations that violate the assumption of qpAdm that gene flow into the source population does not occur after its split with the lineage leading to the admixed target population in the proposed model, the impact of including a large number of populations in the reference population set and the behavior of qpAdm when applied to population histories that involve continuous gene flow rather than single pulses of admixture.

We show that qpAdm reliably identifies population histories involving admixture and accurately infers admixture proportions. It is robust to low coverage, high rates of missing data, DNA damage occurring at similar rates in all populations, the use of pseudo-haploid data, small sample size, and ascertainment bias. We also identify some issues with naive applications of qpAdm. One of these issues is that multiple plausible scenarios may be found most of which are not the truth because qpAdm uses nonrejection of null models as its criterion for plausibility. Another of these issues is that true models may be rejected if samples from too many populations are included in the analysis. A third is that qpAdm results may be difficult to interpret and even misleading under conditions of continuous gene flow. In order to help guard against these potential pitfalls and make this tool more accessible to users, we include an updated user guide for qpAdm (Supplementary Material S1) where we make specific recommendations for best practices for use.

Methods and results

Data generation

We used msprime version 0.7.1 (Kelleher *et al.* 2016) to simulate genome-wide data using the TreeSequence.variants() method, which provides information about the position of all mutations arising in the dataset and the alleles observed for each individual at the variant sites. We then converted this output to EIGENSTRAT format (Patterson *et al.* 2006). Parameters were chosen in order to mirror what has been estimated for humans, including a mutation rate of 1.5×10^{-8} mutations per base pair per generation (1000 Genomes Project Consortium 2010), recombination rate of 1.0×10^{-8} per base pair per generation, and effective population sizes between 2.5×10^4 and 8.0×10^5 (varying between populations and over time; see Supplementary Files S1–S5 for full details). We generated sequence data for 22 chromosomes, each of the approximate length of each of the human autosomes. We simulated $2n$ haploid individuals then combined pairs of haploid individuals to form n diploid individuals.

In order to assess the performance of qpAdm when the population history of a group is relatively simple and fully understood, we simulated genetic data according to a base population tree (Figure 1), consisting of 16 populations and two admixture events (one relatively recent and the other occurring much earlier in the population history). For the more recent admixture event, lineages 14a and 14b contribute α and $1 - \alpha$ proportion of ancestry to population 14, respectively. Unless otherwise noted, α is equal to 0.5. In the earlier admixture event, lineages 15a and 15b contribute β and $1 - \beta$ proportion of ancestry to population 15, respectively, where β is set (arbitrarily) to 0.55. This tree is an expanded version of a population tree described in Patterson *et al.* (2012), which was used to test the performance of the tool qpGraph. The parameters of this model were chosen so that the overall level of variation (total number of SNPs) and the differentiation between populations (F_{ST}) were similar to what is observed between known human groups, such as the Uyghur, French, and Han. We used the same parameters, including population sizes, as Patterson *et al.* (2012). All new populations added to the model

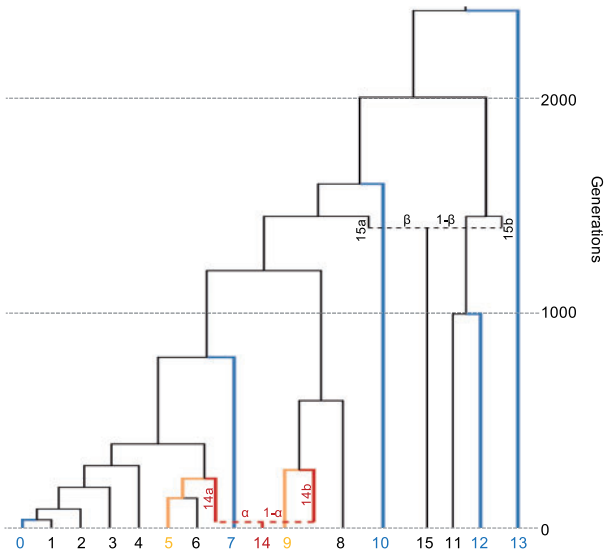

Figure 1 Population history of simulated data. Populations included in the standard model used for qpAdm models are indicated as follows: target (red), sources (yellow), and references (blue).

retain the same population sizes as the original branch from which they split. The exact simulation parameters we used are described in Supplementary File S1 and we report the pairwise F_{ST} between all populations in Supplementary Table S1.

For most simulations, we generated genomic data for samples taken from 10 (diploid) individuals from each of the 16 populations in Figure 1. The populations in Figure 1 are idealized, theoretical populations (see Winther *et al.* 2015) and are not meant to represent any particular human groups. Likewise, the mostly tree-like relationships of populations in Figure 1 simply reflect the kinds of historical scenarios qpAdm was designed to handle. We consider an example of non-tree-like structure in the section on continuous gene flow.

Unless otherwise noted, the admixture model of interest is defined as follows; population 14 is the target population (the ancestry of which is being modeled), populations 5 and 9 are defined as the sources of this admixture, while populations 0, 7, 10, 12, and 13 are designated as reference populations. As none of these reference populations are more closely related to the target population than to either of the two source populations (*i.e.*, the reference populations do not have any shared drift with the target population that is not also shared with at least one of the source populations), this model should be considered plausible. This model will be referred to as the standard model. Note that because populations 5 and 6 are symmetrically related to population 14, both represent be equally good sources of its ancestry. Unless otherwise noted, population 6 will therefore be excluded from analyses.

All qpAdm analyses were performed using qpAdm version 960, using default parameters, and the optional parameter "allsnps: YES" unless otherwise specified. See Supplementary Material S1 for a complete description of all qpAdm parameters.

We confirm that the simulated individuals share the expected genetic relationships through analysis with the population genetic tools principal components analysis (PCA) (Patterson *et al.* 2006) and ADMIXTURE (Alexander *et al.* 2009) (Supplementary Figure S1; Tables S2–S4), which reveal patterns that are consistent with the defined demographic history, including that population 14 is admixed. Notably, we cannot meaningfully distinguish between populations 0–6 using these methods, highlighting how closely related these populations are.

Distribution of *P*-values

As stated earlier, qpAdm outputs a *P*-value that is used to determine whether a specific model of population history can be considered plausible. Models are rejected, or regarded as implausible, when the *P*-value is below the chosen significance cut-off (typically, although arbitrarily, 0.05). In order for true models to be rejected properly at this nominal significance level, that is only 5% of the time, the distribution of *P*-values should be uniform when the null model is equal to the true model. However, this assumption of uniformity of *P*-values in qpAdm has never been confirmed. We therefore assessed the distribution of *P*-values produced by qpAdm by simulating 5000 replicates under our standard model (defined in Figure 1) and running qpAdm on each replicate using the target, source, and reference populations defined in the standard model. We find that the *P*-values generated by qpAdm appear uniformly distributed (Figure 2A; Supplementary Table S5). Using a Kolmogorov–Smirnov test, we fail to reject the null hypothesis that the calculated P-values are uniformly distributed ($P = 0.644$), supporting theoretical predictions for the uniform distribution of *P*-values generated by qpAdm when an accurate model is used.

As qpAdm is often used to distinguish between optimal and nonoptimal models of admixture, we also seek to confirm that the distribution of *P*-values is not uniform when an incorrect model is considered. We therefore examine the distribution of *P*-values produced when nonoptimal populations (*i.e.*, populations 1–4 and 11) are used as sources instead of population 5. As populations 1–4 share more genetic drift with reference population 0 than the true source population (and similarly because population 11 shares less drift with population 0 than the true source population), we expect that the distribution of *P*-values produced by qpAdm should be biased toward zero when these populations are used as sources (with population 11 producing the strongest bias). We ran these nonoptimal qpAdm models on the 5000 replicate datasets described above and observe a deviation from a uniform distribution. In the case of populations 1–4, models that include source populations that share the most drift with population 0 yield *P*-value distributions that are most strongly biased toward zero (Figure 2, B–F; Supplementary Table S5), and as expected, *P*-values associated with using population 11 as a source are even more strongly biased toward zero. In each case, using a Kolmogorov–Smirnov test, we reject the null hypothesis that the *P*-values are uniformly distributed.

Although the distributions of *P*-values deviate from a uniform distribution as expected, we also note that in the cases where populations 1–4 are used as potential source populations, a large proportion of these models are assigned *P*-values that would be considered plausible using 0.05 as a standard threshold. These results reflect the fact that populations 1–5 are all closely related (average pairwise F_{ST} between <0.001–0.005; Supplementary Table S1), therefore the inclusion of population 0 as the only reference population with the power to distinguish between these populations (as it is differentially related to them), may not be enough to reject models that use populations 1–4 as sources in all cases. In practice, if populations 1–5 were all proposed as potential sources and qpAdm assigned plausible *P*-values to multiple models, further analysis would be required to distinguish between these models. Furthermore, we do note that when population 0 is excluded from the reference population set, all of the tested qpAdm models using populations 1–5 as a potential

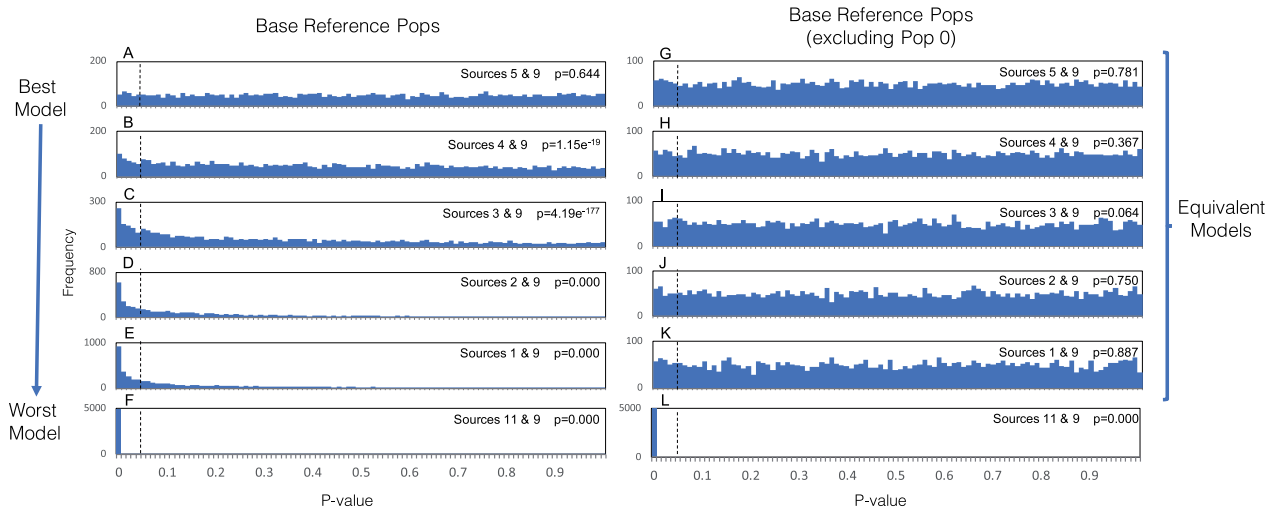

Figure 2 Distribution of *P*-values generated for various qpAdm models. The distribution of *P*-values generated by 5000 replicates of qpAdm is shown for all models of the ancestry of the admixed population (14) of interest. (A) The distribution of *P*-values produced by models using populations 5 and 9 as sources, which are the best possible sources of ancestry for population 14 out of the proposed models when the base reference population set (13, 12, 10, 7, and 0) is used (left). (B–F) The distribution of *P*-values produced by models that use increasingly inappropriate source populations, relative to the base reference population set. In contrast, when population 0 is removed from the reference population set (right), all models are considered equivalent, except for that in which population 11 is defined a source (G–L). Vertical black dotted lines indicate the *P*-value threshold of 0.05, above which qpAdm models are considered plausible. The results of a Kolmogorov–Smirnov test to determine whether the *P*-values are uniformly distributed are indicated.

sources produce approximately uniformly distributed *P*-values, as would be expected theoretically, as populations 1–5 are all symmetrically related to all other reference populations (Figure 2, G–K; Supplementary Table S6). We also observe consistent results when using qpAdm to model the ancestry of population 15, finding that population 15 can be modeled as the product of admixture between populations 8 or 9 and 11, only in cases were populations 10 and 12 are removed from the reference population set, as these populations violate the assumption of qpAdm that reference and source populations must be differentially related to the target population (Supplementary Figure S2; Tables S7 and S8).

While the overall distributions of *P*-values differ between optimal and nonoptimal qpAdm models, we note that for individual replicates the most optimal model is not necessarily assigned the highest *P*-value. We find that the *P*-value associated with the best model (sources 5 and 9) produces the highest *P*-value in only 48% of cases (Supplementary Table S5), when the standard reference set is used (13, 12, 10, 7, and 0). In frequentist methods such as qpAdm, *P*-values below the nominal significance level are judged wrong enough to be rejected, but *P*-values do not represent probabilities of models being correct. As Figure 2 shows, qpAdm is fairly conservative in rejecting models. For example, the model which posits populations 4 and 9 as sources may be considered wrong because population 4 is more closely related to source population 0 than it is to the target population 14. Still, *P*-values under this model are almost uniformly distributed (Figure 2B) and for a given data set the *P*-value for this model could easily be larger than the *P*-value for the correct model (Figure 2A). In contrast, models that diverge strongly from the truth are always rejected, as when populations 11 and 9 are used as sources (Figure 2F). Therefore, in cases where multiple models are assigned plausible *P*-values (*i.e.*, $P \geq 0.05$), we caution that *P*-value ranking (*i.e.*, selecting the model that is assigned the highest *P*-value) should not be used to identify the best model. Methods for distinguishing between multiple models will be discussed further in the section on comparing admixture models.

Effects of varying the block jackknife size

We also explored the effect of varying the block jackknife size, which is used in qpAdm to compute standard errors (Supplementary Material S2). In order to understand how dependent qpAdm is on choosing an appropriate block jackknife size, we vary the block size between 0.0001 and 1 Morgans (default is 0.05 Morgans), and for each block size we test 500 replicates of the standard qpAdm model calculated either on the entire dataset or after randomly down-sampling to 1 million SNPs. We find that admixture proportion estimates are relatively consistent regardless of block size (Supplementary Figure S3A; Table S9) and that standard error estimates are lowest when the smallest block sizes are used (Supplementary Figure S3B). However, for the smallest and largest block sizes, we observe nonuniformly distributed *P*-values (Supplementary Figure S3, C–R), suggesting that when selecting an appropriate block jackknife size for qpAdm there is a trade-off between minimizing standard errors and calculating meaningful *P*-values. This effect also appears to depend upon the number of SNPs used, as biases in *P*-value distributions appear stronger when the full dataset is used. These results are consistent with theoretical expectations, as we expect that when the block size is too small there will be correlation between SNPs across different blocks that is uncorrected. Conversely, when the block size is too large, the standard error of the f_4 statistics used in qpAdm calculations may be poorly estimated. Despite the observation of biased *P*-value distributions, qpAdm appears relatively robust to the selected block jackknife size, as biases were only observed in cases where the block size was either 50× smaller or 20× larger than the default block jackknife size.

Accuracy of admixture proportion estimates

In addition to generating informative *P*-values, it is essential that qpAdm generates accurate admixture proportion estimates. This

has also not been formally tested using simulated data. We therefore simulate genetic data according to the population tree shown in Figure 1, varying the proportion of admixture (α) occurring in the lineage ancestral to population 14 between 0.0 and 1.0 at intervals of 0.1 with 20 replicates per interval. We find that the estimated admixture proportions are extremely close to the actual simulated admixture proportions for all values of α (Figure 3A; Supplementary Table S10). In 99.3% of cases (220 total), the estimated α is within 3 standard errors of the simulated α, consistent with theoretical expectations, with an average standard error of 0.0092 (range: 0.008–0.011). These results indicate that qpAdm accurately estimates admixture proportions, regardless of the level of admixture, and that the standard errors produced by qpAdm are well calibrated. However, we recognize that in practice, users of qpAdm have access to a much less complete dataset. Therefore, we modify the data in order to explore the performance of qpAdm when applied to data of lower coverage and quality.

Each simulation contains an average of ~30 million SNPs. In order to understand the performance of qpAdm with less data, we randomly down-sample the complete dataset to produce analysis datasets of 1 million, 100,000, and 10,000 sites. In all cases, the average admixture proportion estimate generated is extremely close to the simulated α, although we do observe an increase in variance in the individual estimates as the amount of data analyzed decreases (Figure 3A; Supplementary Table S10). Similarly, we do not observe biases in admixture proportions when using nonrandom ascertainment schemes to select sites for analysis (Figure 3B; Supplementary Table S11). The impact of nonrandom ascertainment schemes on qpAdm analyses are described in more detail in a later section. In order to increase computational efficiency and to better approximate typical analysis datasets, all subsequent analyses are performed on the data that has been randomly down-sampled to 1 million sites, unless otherwise specified.

We find that qpAdm is robust to missing data, where data from randomly selected sites in each individual is considered missing with rate 10%, 25%, 50%, 75%, or 90% (Figure 3C; Supplementary Table S12), resulting in a dataset where each individual has genetic data available for a different subset of SNPs (as opposed to the previous down-sampling test where all individuals shared a common set of SNPs of varying sizes). Additionally, we find that pseudohaploidy—a common feature of ancient DNA, where due to low sequencing coverage, a haploid genotype is determined by randomly selecting one allele at each diploid site and assigning that to be the genotype—has little impact on admixture estimates (Figure 3D; Supplementary Table S13).

Ancient DNA is also subject to deamination, resulting in C-to-T or G-to-A substitutions appearing in transition sites (Dabney *et al.* 2013). In the 1.2 million SNP sites that are commonly targeted in ancient DNA analysis, approximately 77.6% of these sites are transitions (Fu *et al.* 2015; Haak *et al.* 2015; Mathieson *et al.* 2015). We therefore randomly defined 77.6% of simulated sites to function as transitions. For each of these transition sites, in each individual, if the allele at that position is of the reference type, it was changed to the alternative type with 5% probability (representing an extreme degree of damage), mimicking the unidirectional change in allelic state caused by ancient DNA damage. We find that admixture proportion estimates produced by qpAdm are relatively robust to the presence of ancient DNA damage in cases where all populations exhibit an equal damage rate (Figure 3E; Supplementary Table S14). However, in cases where the target (population 14) and source (5 + 9) populations have a different rate of ancient damage the estimated admixture proportions are biased. This bias reflects attraction between populations on the left and right sides of the f_4 statistics calculated by qpAdm and is not unexpected, as ancient DNA damage occurring in only a subset of populations would cause allele frequencies in these populations to appear more correlated than would be expected based on their phylogenetic relationship alone due to the unidirectional shift from cytosine to thymine (or guanine to adenine) nucleotides at transition sites.

Another concern that is common among ancient DNA analyses is small sample size. We therefore explore the effect of reducing the sample size of various populations in the analysis from 10 individuals down to a single individual. We find that admixture estimates are relatively robust to this reduced sample size regardless of whether the target (14), source (5, 9, or 5 + 9), or reference (0 or 0 + 7 + 10) population set has only a single individual sampled (Figure 3F; Supplementary Table S15). Reducing the target sample size to a single individual appears to have the greatest effect out of all cases where only the sample size of a single population was reduced, maximally increasing the variance in estimates of alpha. Furthermore, we see that when only a single individual is sampled from every population, the admixture proportion estimates vary the most between replicates, however, the mean of these estimates fall close to the true α, suggesting that small sample size does not result in an upward or downward bias in the admixture proportion estimates produced by qpAdm.

In order to confirm that these results are also consistent when applied to nonsimulated data, we repeat these analyses on real population genetic data. We therefore model the ancestry of the Uyghur population as the product of admixture between populations related to the French and Han, with Adygei and Yoruba used as references populations (as defined in Patterson *et al.* [2012]), adding the Onge as an additional reference population in order to meet the requirement that there are at least as many reference populations as target and source populations. In this analysis, we replicate previous findings that the Uyghur can be modeled as ~47% French and ~53% Han. Furthermore, when we model the reduction in data quality as described for the simulated data, we find that the admixture proportion estimates are similarly robust to this reduction in data quality. While only a single replicate was performed for each condition, we note that the size of the standard errors assigned by qpAdm mirror the amount variance observed in admixture proportion estimates in the simulated analyses (Supplementary Figure S5; Table S16).

Comparing admixture models

One of the major applications of qpAdm is to identify an optimal admixture model out of a variety of proposed possible models, many of which may be deemed plausible by qpAdm. However, no formal recommendations have been made about what strategy to use when comparing models. We therefore explore two commonly employed approaches for comparing admixture models in order to make recommendations for best practices in qpAdm usage.

One of the most typical implementations of qpAdm involves the selection of a set of differentially related populations to serve as the base set of reference populations. This base set of reference populations is often chosen to represent key positions in the known population history [*i.e.*, the 'O9' reference set defined in Lazaridis *et al.* (2016)]. A nonoverlapping set of source populations is then defined, and qpAdm models involving different combinations of source populations and the base set of reference populations are tested. Using this method, multiple models may meet

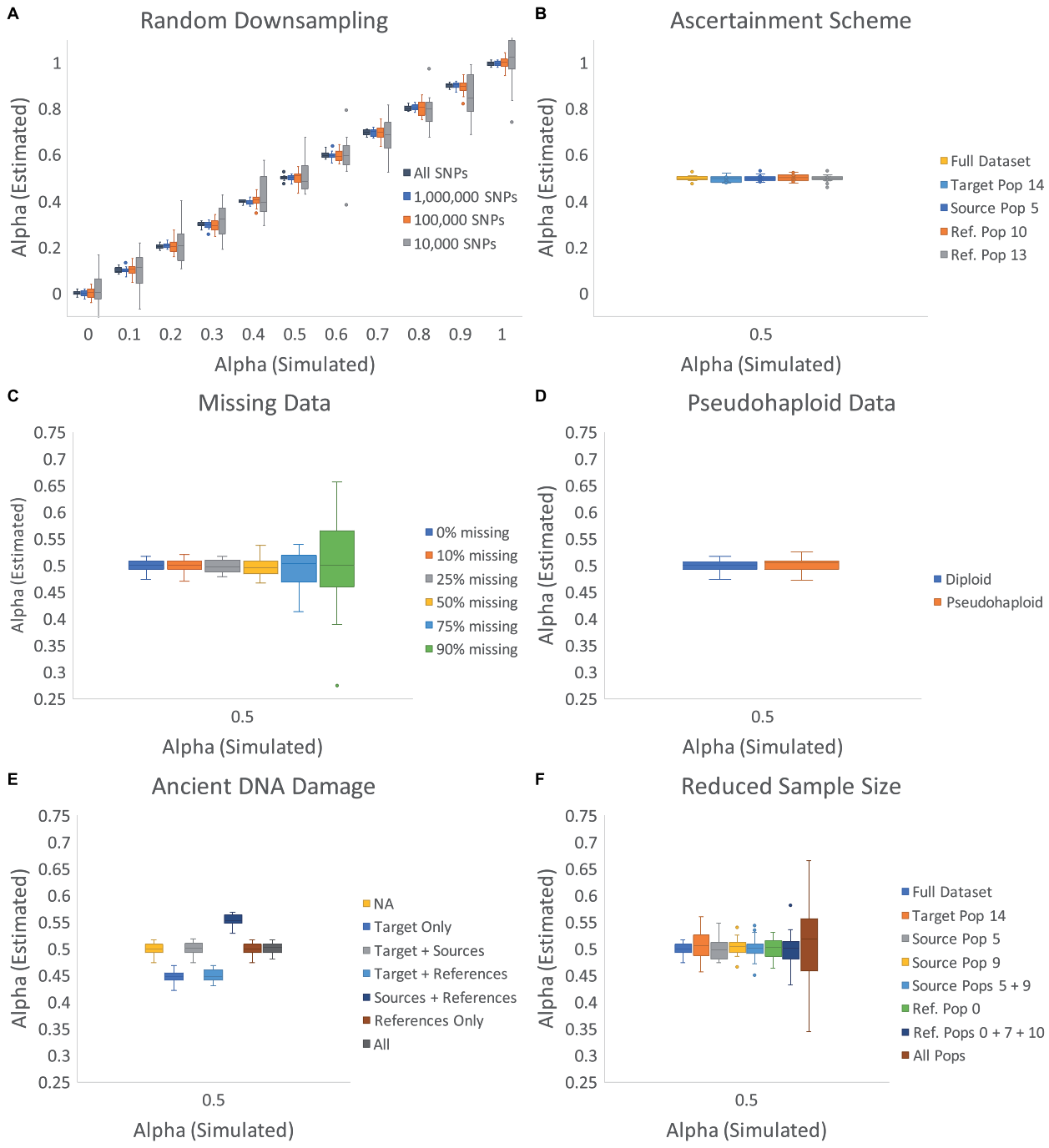

Figure 3 Accuracy of admixture proportion estimates. Box and whisker plots showing the estimated values of admixture proportion (alpha) generated by qpAdm for varying simulated alphas. Only alpha 0.5 is shown for panels B–F, however all alphas 0–1 are reported in Supplementary Figure S4. For each simulated alpha, 20 replicates of qpAdm are performed for each condition. (A) Estimates produced by qpAdm when run on the entire dataset and after randomly down-sampling to 1 million, 100 thousand, and 10 thousand SNPs. All subsequent analyses are performed on the 1 million SNP down-sampled dataset unless otherwise specified. (B) Estimates produced by qpAdm when data is ascertained on population 14, 5, 10, or 13. (C) Estimates produced by qpAdm where some proportion (0%, 10%, 25%, 50%, 75%, or 90%) of data is missing in each individual. (D) Estimates produced by qpAdm in both diploid and pseudohaploid form. (E) Estimates produced by qpAdm where 5% ancient DNA damage is simulated in a subset of populations (target, target + sources, target + references, sources + references, references only, and all populations). (F) Estimates produced by qpAdm, where only a single individual is sampled from varying populations [target, a single source (5 or 9), both sources (5 + 9), a single reference (0), multiple references (0 + 7 + 10), and all populations].

the criteria to be considered plausible, and the most optimal model is identified by adding additional reference populations to the base set of references, which are selected for their differential relatedness to one or more of the source populations in the set of potentially plausible qpAdm models.

While this strategy is relatively straightforward and widely implemented (*e.g.*, Lazaridis *et al.* 2016; Harney *et al.* 2018), it has several drawbacks. In particular, because a population cannot simultaneously serve as a source and reference population, this strategy either requires that populations that are placed in the

base set of reference populations are not considered as potential source populations (meaning it is possible that the best source population would be entirely missed if it were selected to serve in the reference population set) or that potential source populations be selectively removed from the reference population set so that they can be used as source populations for some models. This strategy results in the creation of some models that are not equivalent, and therefore are difficult to compare.

An alternative to the "base" reference set strategy that has been implemented in order to avoid these problems is to create a set of "rotating" models in which a single set of populations is selected for analysis (*e.g.*, Skoglund *et al.* 2017; Harney *et al.* 2019). From this single set of populations, a defined number of source populations are selected, and all other populations then serve in the reference population set for the model. Under this "rotating" scheme, populations are systematically moved from the set of reference populations to the set of sources. Thus, all population models are generated using a common set of principles and are therefore more easily directly compared. In order to compare the performance of these two strategies ("base" *vs* "rotating"), we again focus on the population history of population 14 (Figure 1).

For the "base" reference approach, we continue to use the base set of reference populations as previously defined (populations 0, 7, 10, 12, and 13), all other populations are considered to be potential source populations. We used qpAdm to test all possible combinations of two source populations. We ran each of these qpAdm models on the data generated using the standard population history with $\alpha = 0.50$, with 20 replicates. Among these 20 replicates, qpAdm identified the optimal model, in which populations 5 and 9 serve as source populations, as plausible in 19 cases (Figure 4A, upper triangle; Supplementary Table S17). However, there are also a large number of other population models that are consistently deemed plausible; for example when population 8 is used as a source (in conjunction with population 5) instead of population 9, 90% of the models are deemed plausible. The high rate of acceptance of this model is fully consistent with expectations, because while population 9 is more closely related to the true source population, populations 8 and 9 are symmetrically related to all of the reference populations included in the model, and therefore are indistinguishable using this approach (unless data from a population that differentially related to these two populations could be added to the model). Models that include populations 1–4 (in combination with populations 8 or 9) were also frequently identified as plausible. These results suggest that the inclusion of population 0 as a reference does not provide enough information to differentiate between these potential source populations and the true optimal source (population 5). Therefore, the next step in a qpAdm analysis that utilizes the base model approach would be to add additional reference populations that are differentially related to populations 1–5 in order to help differentiate between the remaining possible models.

In contrast, we find that under a "rotating" model, where all populations (except for population 6 because it is phylogenetically a clade with source 5) were selected to serve as either a source or a reference population, all models that included populations 5 and 9 as sources were identified as plausible. In contrast all other population models were rejected (Figure 4A, lower triangle; Supplementary Table S18). The inclusion of the optimal source populations (5 and 9) as references in all other models enables qpAdm to differentiate between models that would otherwise be indistinguishable (such as differentiating between populations 8 and 9 and between populations 1 and 5). Furthermore, in cases where optimal source populations are not available (*i.e.*, if both populations 5 and 6 are excluded from the model), qpAdm still identifies closely related models as plausible (such as those involving admixture between population 9 and populations 0–4), suggesting that this rotating approach is not overly stringent in cases where optimal sources are not available (Figure 4B;

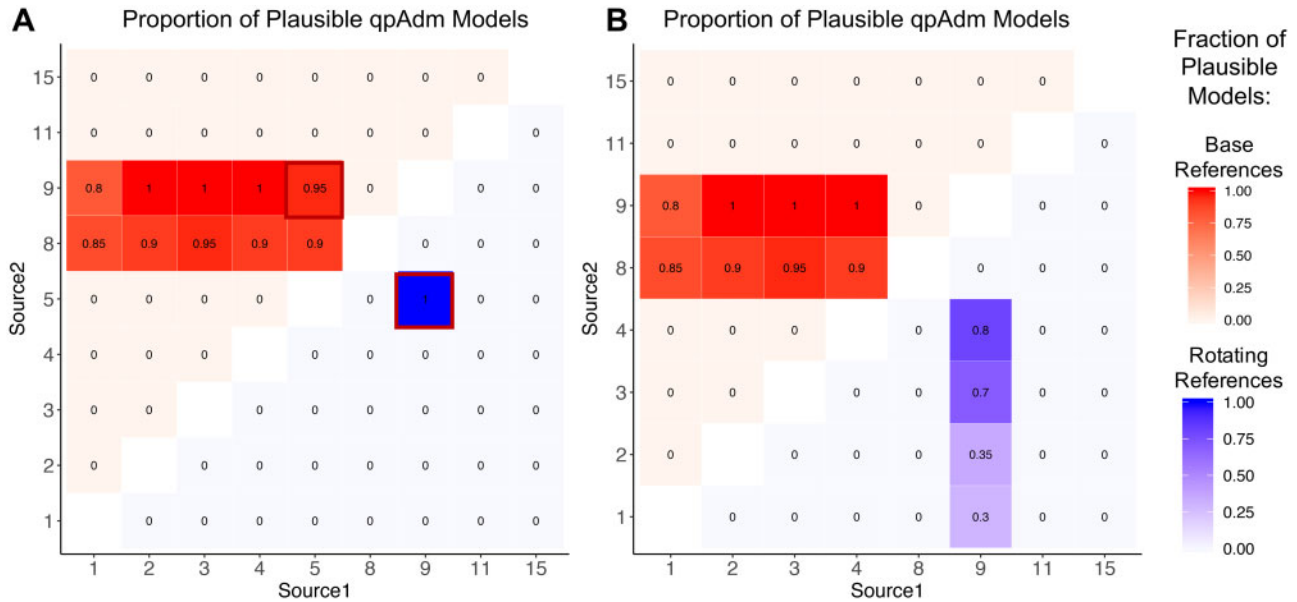

Figure 4 Comparing qpAdm models using various approaches. A heatmap showing the proportion of replicates in which the two-way admixture model generated using each combination of possible source populations is deemed plausible (*i.e.*, yielded a *P*-value > 0.05 and admixture proportion estimates between 0 and 1) by qpAdm. (A) The upper triangle (red) shows results generated using the base set of reference populations (0, 7, 10, 12, and 13), while the lower (blue) triangle shows results generated using the rotating model approach. The proportion of replicates deemed plausible is indicated by the color (darker shades indicate a higher proportion) and is written inside each square of the heatmap. The optimal admixture model for in each case is outlined in red. Only results for combinations of sources that were possible using both approaches are shown. (B) The results generated when population 5 (an optimal source) is excluded from all models.

Supplementary Table S19). Due to the relative simplicity of the rotating model approach and the increased ability to identify the optimal admixture model when using it, we recommend utilizing a rotating strategy when possible.

Ascertainment bias and "rotating" model selection

In order to understand the impact of ascertainment bias on model selection, we repeated this analysis on data that was ascertained from the full dataset using several nonrandom ascertainment strategies, in which we ascertained on (*i.e.*, restricted to) sites that were found to be heterozygous in a single individual from (1) the target (population 14), (2) a source (population 5), and (3) two populations that are uninvolved in the admixture event (population 10 and 13), mirroring the ascertainment scheme used to generate the Human Origins dataset (Patterson *et al.* 2012). The individual used for data ascertainment was excluded from subsequent analyses. In all cases, using the rotating approach previously described, only models that use populations 5 and 9 as sources are deemed plausible (Supplementary Figure S6; Table S20), suggesting that ascertainment bias is unlikely to cause users to identify inappropriate models as plausible. Furthermore, the optimal model was identified as plausible in at least 90% of replicates using all ascertainment strategies, suggesting that qpAdm is robust to ascertainment bias. These results are consistent with previous findings that f_4 statistics, which are used for all qpAdm calculations, are robust to biased ascertainment processes (Patterson *et al.* 2012).

Missing data and the "allsnps" option of qpAdm

We also explored the effect of qpAdm's "allsnps" option when working with samples with a large amount of missing data. If the default "allsnps: NO" option is selected, qpAdm only analyzes sites that are shared between all target, source, and reference populations that are included in the model. In contrast, if "allsnps: YES" is selected, every individual f_4 statistic is calculated using the intersection of SNPs that have available data for the four populations that are involved in that particular calculation, therefore every f_4 statistic is calculated using a unique set of sites. The "allsnps: YES" parameter is commonly used in cases where one or more populations in the analysis dataset has a high rate of missing data, in order to increase the number of sites analyzed. However, this causes the underlying calculations performed by qpAdm to deviate from those on which the theory is based, and the effect of this change in calculations on admixture proportions estimated by qpAdm and on optimal model identification is not well studied.

We explore the effects of this parameter, using simulated data with admixture proportion $\alpha = 0.50$ and rates of missing data equal to either 25%, 80%, 85%, or 90% for all individuals across 1 million randomly chosen SNP sites. We implemented the rotating model for both the "allsnps: YES" and "allsnps: NO" options (all previous analyses used the "allsnps: YES" option). Comparing all possible models using the rotating approach, we find that the results produced when using the "allsnps: YES" and "allsnps: NO" options are similar when the rate of missing data is low (*i.e.*, 25%) (Figure 5A; Supplementary Table S21). The optimal model (with sources 5 and 9) was identified as plausible in 95% of cases and no other models were deemed plausible for both options. Furthermore, the admixture proportion estimates produced in both cases are relatively similar, with average standard errors of 0.006 in both cases. The similar performance of the "allsnps: YES" and "allsnps: NO" options in this case is likely due to the relatively large sample size (10 individuals per population) used in the analysis. With 25% missing data, the expected number of SNPs to be included the analysis when the "allsnps: YES" option is selected is 1 million. This number is only slightly reduced, to 999,985.7, when the "allsnps: NO" option is selected.

In contrast, when the rate of missing data is elevated (*i.e.*, 80%, 85%, or 90%), a difference in performance between the "allsnps: YES" and "allsnps: NO" options was observed. In each case, when the rate of missing data increased, the number of nonoptimal models that were identified as plausible also increased (Figure 5, B–D). These changes were more dramatic when the "allsnps: NO" parameter was used, further we observe a greater increase in the standard errors associated with admixture proportion estimates produced when using the "allsnps: NO" option, with average standard errors equal to 0.025, 0.066, and 9.994 when analyzing data with 80%, 85%, and 90% missing data, respectively. In contrast, while the standard errors produced using the "allsnps: YES" option also increased, the increase was lower in magnitude in all cases, with standard errors of 0.015, 0.020, and 0.035 observed, respectively. This difference in performance is likely the result of the number of SNPs available for analysis when using each option. When using the "allsnps: YES" parameter, the expected number of SNPs used in analysis of data with 80%, 85%, and 90% missing data rates remains 1 million. However, when using the "allsnps: NO" parameter, the expected number of SNPs used in analysis with each rate of missing data is only 181,987.5, 37,303.7, and 1,610.4 SNPs, respectively. These results suggest that the increased data provided by using the "allsnps: YES" option improves the ability of qpAdm to distinguish between models, without creating biases in cases where missing data is distributed randomly throughout the genome of all individuals.

The effects of ancient DNA damage on model selection

In an earlier section, we show that admixture proportion estimates produced by qpAdm can be biased when produced using populations with differential rates of ancient DNA damage. We therefore explored the effects of damage on model comparison, using the rotating model approach. Across all cases, only models involving the optimal sources (populations 5 and 9) are deemed plausible, suggesting that ancient DNA damage, even when unevenly distributed, is unlikely to cause a user to identify a nonoptimal model as plausible (Supplementary Figure S7; Table S22). Furthermore, when damage rates are consistent between the target and optimal source populations, the optimal model is identified as plausible in at least 95% of cases. However, when the target and source populations have differential rates of damage, this optimal model is almost always deemed implausible. We do note that the ancient DNA damage simulated in this analysis (5% ancient DNA damage rate at all "transition" sites) is relatively high, as most ancient DNA damage occurs at the terminal ends of DNA molecules. Therefore, these results likely represent an extreme case. However, these results highlight the importance of considering the effect of ancient DNA damage in ancient DNA analyses. In particular, we caution against designs where both ancient and present-day populations are included in a single qpAdm model.

The effects of sample size on model selection

We also considered the impact of limited sample size when comparing models, using a rotating model approach. Using the same data shown in Figure 3E, where the sample size of the specified population(s) was reduced from 10 to 1. In cases where the population(s) with reduced sample size were not involved in the admixture event of interest the effect of sample size reduction is minimal (Supplementary Figure S8; Table S23). Similarly, the results do not appear to be significantly affected when the

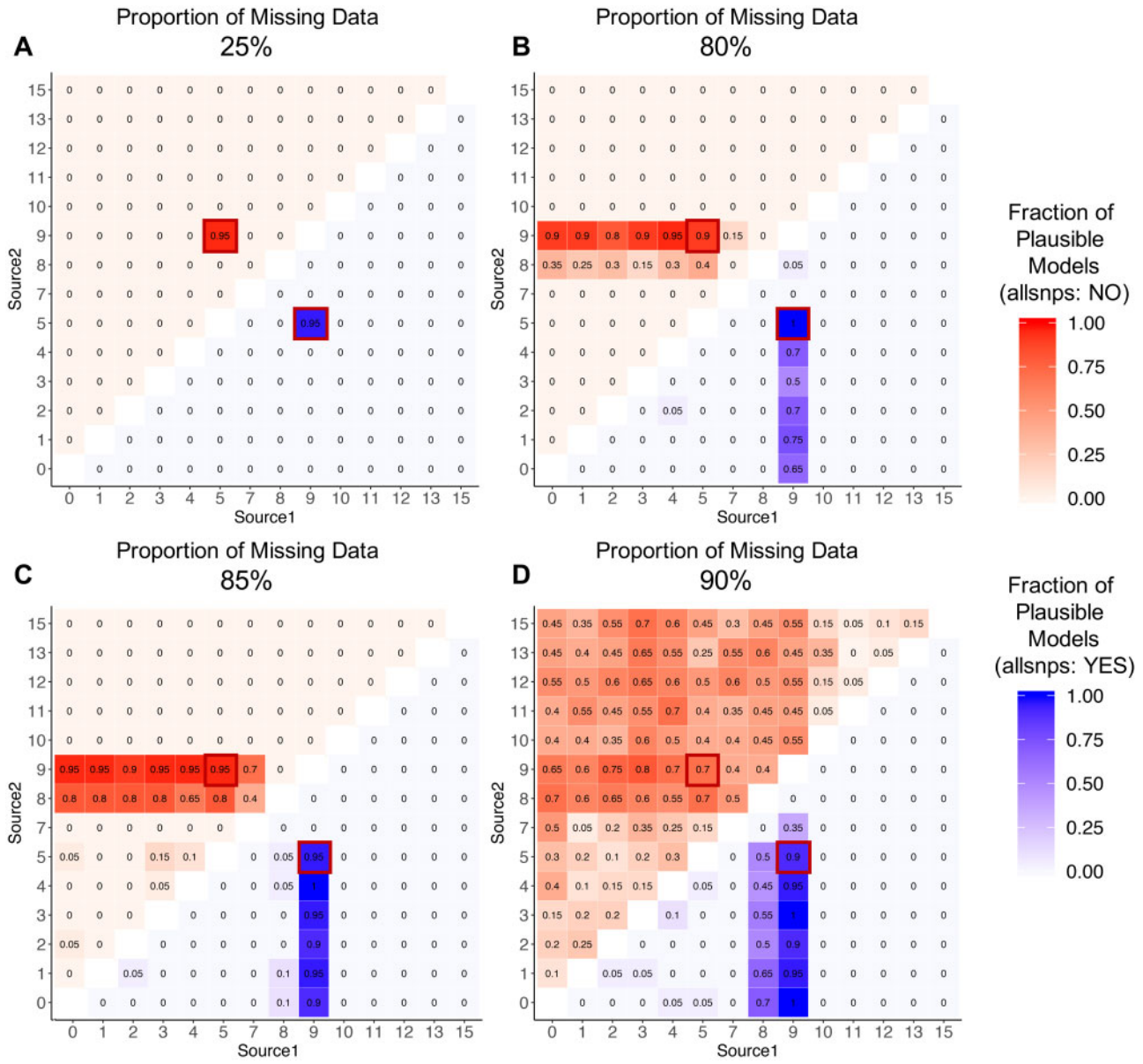

Figure 5 Effect of the allsnps parameter on qpAdm model selection. Heatmaps showing the proportion of replicates in which the two-way admixture model generated using each combination of possible source populations is deemed plausible by qpAdm (*i.e.* yielded a *P*-value > 0.05 and admixture proportion estimates between 0 and 1) on SNP data using the "allsnps: yes" (blue; lower right triangle) and "allsnps: no" parameters (red; upper left triangle), on data with (A) 25% (B) 80% (C) 85%, or (D) 90% missing data. The proportion of replicates deemed plausible is indicated by the color (darker shades indicate a higher proportion) and is written inside each square of the heatmap. The optimal admixture model for each of the approaches are highlighted in red.

sample size of population 9 (one of the optimal source populations) is reduced, suggesting that when the optimal source population is relatively differentiated from all other populations considered, reduced sample size has little effect. However, when the sample size of source population 5 is reduced to one, models using closely related populations as sources were also deemed plausible. Similarly, when the target population (14) contained only a single sampled individual, the proportion of nonoptimal models that were identified as plausible by qpAdm increased. These results suggest that when the sample size is lower, particularly for target or source populations, qpAdm has less power to reject nonoptimal models. This is likely to become an even greater issue in cases where populations included in qpAdm models contain only a single individual with large amounts of missing data. To demonstrate this, we repeat all model comparison analyses, sampling only a single individual from each population and report these results in Supplementary Figures S9–S12 (Supplementary Tables S24–S27).

Modeling unadmixed populations using qpAdm

Finally, while we know that the population history of population 14 involves admixture, the number of ancestral sources that contributed ancestry to a real target population is typically unknown. Therefore, we explored the behavior of qpAdm when modeling the population history of unadmixed and admixed populations (populations 6 and 14, respectively) under various scenarios. First, we explored models in which only a single source population contributed ancestry to the target population, using the same rotating model approach as described previously, but only selecting a single source population for each model. In the

case of the unadmixed population 6, we find that in 95% of cases, it can be modeled as forming a genetic clade with population 5, consistent with theoretical expectations (Supplementary Table S28). In contrast, population 14 is never found to form a genetic clade with any of the tested source populations (Supplementary Table S29), again consistent with expectations. However, when population 6 is modeled as the product of admixture between 2 source populations, we find that it is frequently modeled as the product of a two-way admixture between population 5 and any other source population, where population 5 is estimated to contribute the vast majority of ancestry to population 6 (Supplementary Table S30). We therefore stress the importance of testing all possible models with the lowest rank (*i.e.*, number of source populations) using qpAdm (or the related qpWave) before proceeding to test models with higher rank.

Impact of combined data quality reduction

While any observed bias produced by the factors considered here is minimal, we caution that the increase in variance caused by each form of reduced data quality is additive. Therefore, models relying on data with high rates of missingness, damage, and small sample sizes should be interpreted with particular caution. We demonstrate this by testing 12 combined models, in which we simulated data with alpha equal to 0.1, 0.5, and 0.9. For each dataset, we simulated a 5% damage rate in either the target population (14) or in all populations, made pseudohaploid genotype calls, and then down-sampled each individual to produce either 25% or 75% missingness rates, and then restricted each population sample size to either 1, 2, or 10 individuals, using the previously described methods for data quality reduction (Figure 6; Supplementary Table S31). As data quality is reduced across multiple factors, we observe an increase in variance in estimated alpha and a greater proportion of implausible models, where the estimated alpha falls outside the range of 0–1.

Additionally, we find that qpAdm is generally robust to moderate rates of data quality reduction across multiple factors, however in cases where the rate of missing data is high (75%) and sample size for all populations is low (1 or 2 individuals per population), qpAdm cannot distinguish between optimal and nonoptimal models (Figure 7; Supplementary Table S32). Notably, while differences in damage rates between populations has a large affect when the sample size is large and missing data rate is low, this appears to be less of a problem when data quality is otherwise low, suggesting that differential ancient DNA damage rates only affect model comparison when data quality is otherwise extremely high (although the bias in admixture proportion estimates is still apparent when data quality is reduced).

Challenging scenarios

While we find that qpAdm behaves as expected under standard conditions, we are also interested in identifying scenarios under which qpAdm might behave in unanticipated and undesirable ways. We therefore explore the performance of qpAdm under three challenging scenarios: when there is gene flow between source and reference populations that occurs after the admixture event of interest, when the number of reference populations is very large and when the relatedness of populations is not tree-like but rather reflects ongoing genetic exchange.

Gene flow from reference populations

One of the underlying assumptions of qpAdm is that no gene flow occurs between source and references populations following the split of the source population from the true lineage that participated in the admixture event (or population split) of interest. However, in practice this may be a difficult requirement for users to satisfy in cases where the population history is not well understood. The impact of violating this assumption on qpAdm results has not be formally studied. We therefore explored this scenario by adding gene flow events from a reference population 10 to source population 9 occurring either before (generation 350) or after (generation 200) the split of population 9 from the true admixing source lineage at generation 280. In both cases, we simulate the main admixture event with varying rates of alpha ($\alpha = 0$, 0.05, and 0.50) and the additional gene flow event at varying rates, gamma ($\gamma = 0.01$, 0.05, 0.10, and 0.25) (see Supplementary Files S2a–b for exact simulation parameters). For each scenario, we model the ancestry of population 14 using qpAdm with the standard set of source (5 and 9) and reference (0, 7, 10, 12, and 13) populations.

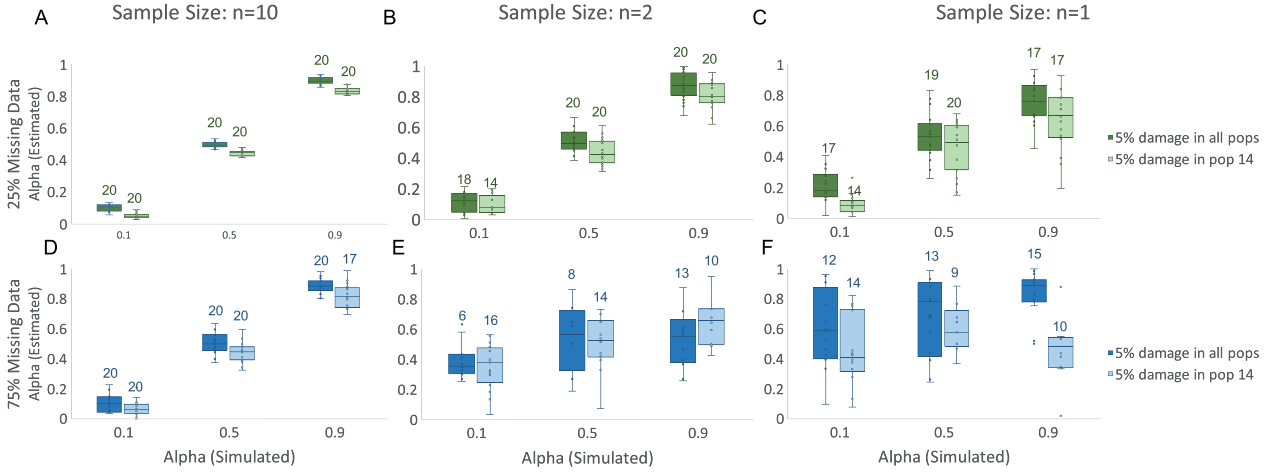

Figure 6 Combined impact of multiple factors of data quality reduction on admixture proportion estimates. Box and whisker plots showing the estimated values of admixture proportion (alpha) generated by qpAdm for varying simulated alphas when applied to data with either 25% (top) or 75% (bottom) missing data, with population sample sizes of either 10 (left), 2 (middle) or 1 (right), and 5% ancient DNA damage in either all populations or just in the target population (14). In each case, 20 replicates were simulated, but only models that produced admixture proportion estimates 0–1 are plotted. The total number of replicates included for each category is written above each box and whisker plot.

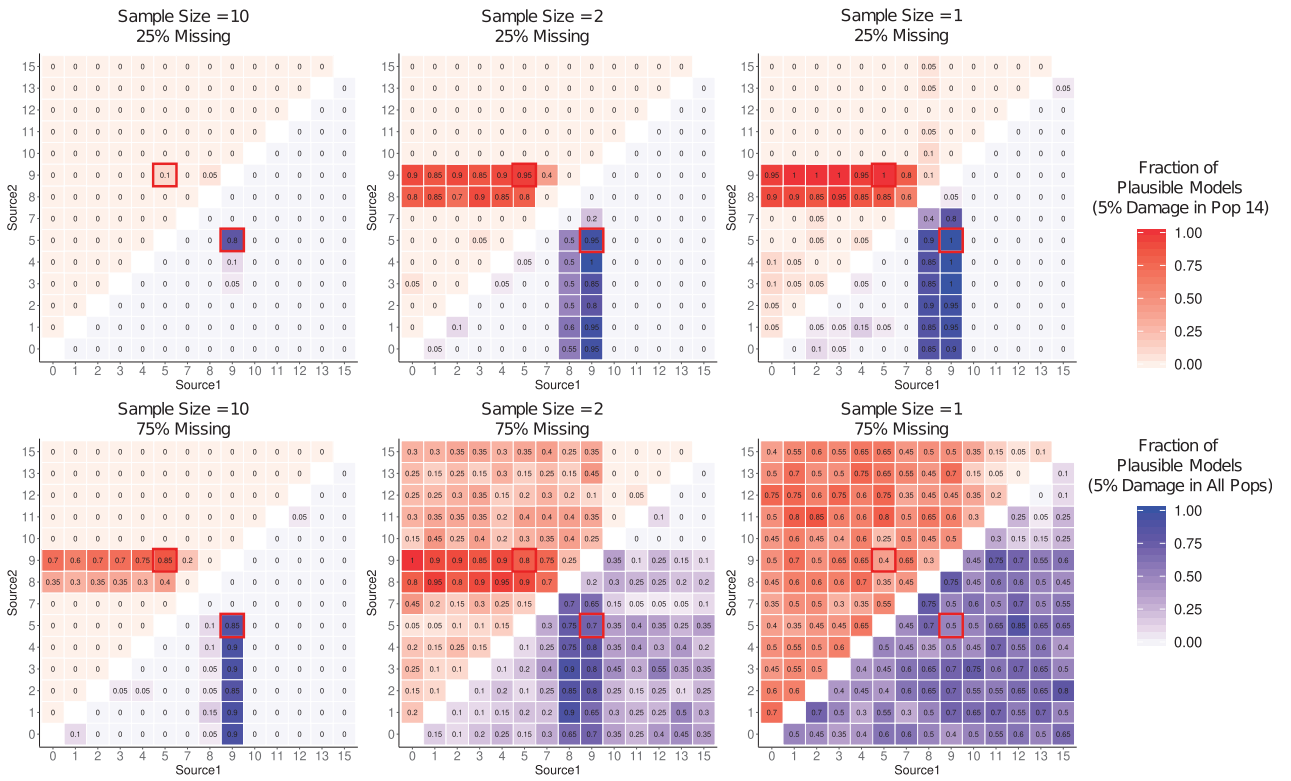

Figure 7 Combined impact of multiple factors of data quality reduction on qpAdm model selection. Heatmaps showing the proportion of replicates in which the two-way admixture model generated using each combination of possible source populations is deemed plausible by qpAdm (*i.e.*, yielded a *P*-value > 0.05 and admixture proportion estimates between 0 and 1) on SNP data using the "allsnps: yes" (blue; lower right triangle) and "allsnps: no" parameters (red; upper left triangle), on data with (A) 25%, (B) 80%, (C) 85%, or (D) 90% missing data. The proportion of replicates deemed plausible is indicated by the color (darker shades indicate a higher proportion) and is written inside each square of the heatmap. The optimal admixture model for each of the approaches are highlighted in red.

Consistent with theoretical expectations, in all cases, when the gene flow event occurs at generation 350 (*i.e.*, before the split between population 9 and the lineage involved in the admixture event of interest), no bias is observed between the simulated and estimated admixture proportions. In contrast, we observe a strong upward bias in alpha when this gene flow event occurs at generation 250 and the magnitude of this bias is correlated with the migration rate (Figure 8B; Supplementary Table S33).

As population 10 is involved in the gene flow of interest and serves as a reference population, we also considered the impact of excluding this population from the qpAdm model (Figure 8C). We observe a similar upward bias in alpha estimates, but the magnitude of this bias is reduced, indicating that any gene flow into a source population that occurs after the split with the lineage involved in the admixture event of interest causes bias, but that this bias is substantially stronger in cases where this gene flow comes from a reference population.

We also explored the impact of gene flow between reference populations (from 10 to 7) or from a source into a reference population (from 9 to 10) and did not observe any bias in admixture proportion estimates (Supplementary Figure S13; Table S33). These results indicate that when selecting populations to include in qpAdm models, users should avoid including source populations that may have experienced gene flow that is more recent than the admixture event of interest. In cases where this is unavoidable, users should make particular effort to avoid including reference populations that may have acted as sources of this gene flow and exercise particular caution when interpreting qpAdm results.

Number of reference populations

We were interested in understanding the effect of assigning an extremely large number of populations to the reference population set. While a commonly employed method for distinguishing between optimal and nonoptimal admixture models and reducing the standard errors associated with a admixture proportion estimates is to increase the number of reference populations included in qpAdm models (*e.g.*, Lazaridis *et al.* 2016; Harney *et al.* 2018), the effect of including too many reference populations in a model is unknown. As qpAdm generates f_4 statistics involving combinations of reference populations, the larger the number of reference populations is, the more poorly estimated the covariance matrix of these f_4 statistics is predicted to be. Therefore, existing guidelines for qpAdm usage recommend against assigning too many populations to the reference set, as the computed *P*-values are thought to be unreliable. However, how many reference populations is "too many" and what the effect of exceeding this number would be on the calculated *P*-values is unknown.

We therefore simulated a dataset with a large number of populations by adding two additional population branching events, occurring 50 generations apart, to all locations on the standard population tree that are marked with a star in Figure 9A, resulting in a total of 118 populations in the simulated dataset (see Supplementary File S3 for exact simulation parameters). After down-sampling the simulated data to 1 million sites, we then ran qpAdm, with population 14 as the target, and populations 5 and 9 as sources. Populations 0, 7, 10, 12, and 13 were again assigned to serve as reference populations. All other populations (excluding population 6) were added, one at a time in random order to the

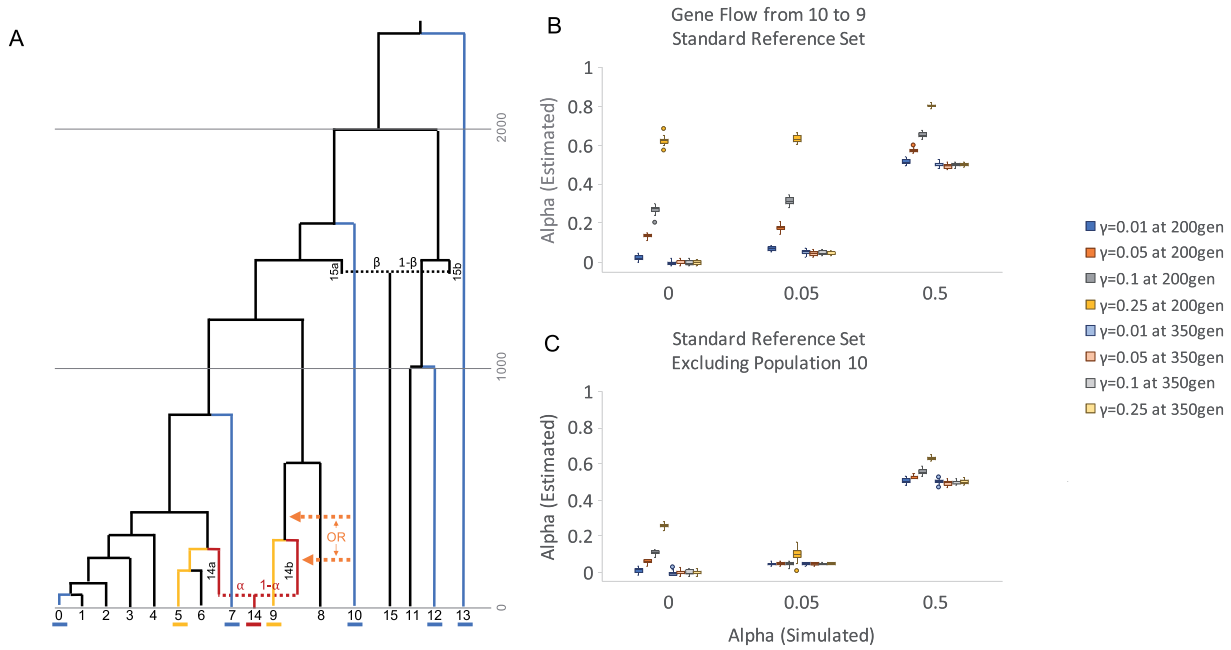

Figure 8 Gene flow from reference populations. (A) Population history where gene flow has been added between reference and source populations to the standard tree. The target, source, and reference populations underlined in red, yellow, and blue, respectively. Arrows represent one of the possible gene flow events that has been simulated, from reference population 10 into source population 9 at generation 200 or 350. (B–C) Admixture proportions estimates generated by a qpAdm model with population 14 as the target, and populations 5 and 9 as sources, applied to the simulated data described in panel A with varying alpha ($\alpha = 0$, 0.05, or 0.50). In each case, gene flow from population 10 to population 9 occurs at rate gamma ($\gamma = 0.01$, 0.05, 0.1, or 0.50). The standard set of reference populations (13, 12, 10, 7, and 0) are included in qpAdm models, while in panel (C) population 10 was excluded from the reference set. Error bars indicate 1 standard error.

reference population set, resulting in qpAdm models with between 5 and 114 reference populations. As each new reference population was added to the model, we re-ran qpAdm and recorded the *P*-value.

Figure 9B shows the change in estimated *P*-value as reference populations are added to the model for 10 separate replicates (Supplementary Table S34). While the *P*-values calculated for each replicate using the original set of 5 reference populations appear to fall randomly between 0 and 1 (consistent with the uniform distribution of *P*-values observed in earlier analyses), we find that in all cases, as the number of reference populations increases the *P*-values eventually fall below the threshold of 0.05, resulting in all of the models with the maximum number of reference populations to be rejected. These results indicate that the inclusion of too many reference populations is likely to result in the rejection of qpAdm models, even in cases where the optimal source populations have been specified.

The maximum number of reference populations that can be included in a qpAdm model before this effect is observed is likely to depend on the specific population history and the total amount of data included in the analysis. In these simulations, we find that qpAdm begins to reject models that would otherwise be deemed plausible when as few as 30 additional populations are added to the outgroup set. These results support previous warnings against including too many reference populations in qpAdm models.

Continuous gene flow

An underlying assumption of qpAdm is that population admixture occurs in a single pulse over a small interval of time, during which the proportion of ancestry coming from each of the ancestral source populations can be estimated. However, real population histories often involve continuous gene flow that occurs over a prolonged period of time. In this case, although the resulting population may have received ancestry from multiple sources, estimates of admixture proportions from these sources may not be meaningful.

Continuous gene flow following an initial pulse admixture event

We first explore the effect of adding continuous migration to the standard population history described in Figure 1. In order to do this, we move the main admixture event to immediately after the split between population 5 and the lineage that directly contributed to the admixture event that produced population 14, at generation 240 (Figure 10A). Following this initial admixture event, continuous gene flow occurs from populations 5 and 9 into population 14 at varying migration rates ($m = 0.0$, 0.00001, 0.0001, 0.001, and 0.01) (see Supplementary file S4 for exact simulation parameters). We generate simulated data for two alphas ($\alpha = 0.0$ and 0.50) and model the ancestry of population 14 as the result of admixture between source populations 5 and 9 using qpAdm with the rotating reference population approach.

We find that in the case of simulated alpha = 0.5, the estimated alpha appears unbiased in all scenarios. However, in the case of simulated alpha = 0.0, as the continuous migration rate increases, the estimated alpha approaches 0.50, reflective of the symmetric migration from populations 5 and 9 into population 14 (Figure 10B; Supplementary Table S35). These results highlight the fact that qpAdm does not explicitly differentiate between pulse admixture events and continuous migration, therefore users should consider both scenarios when interpreting results.

Stepping-stone model

We further extended our exploration of the impact of continuous gene flow by considering data simulated using a stepping-stone

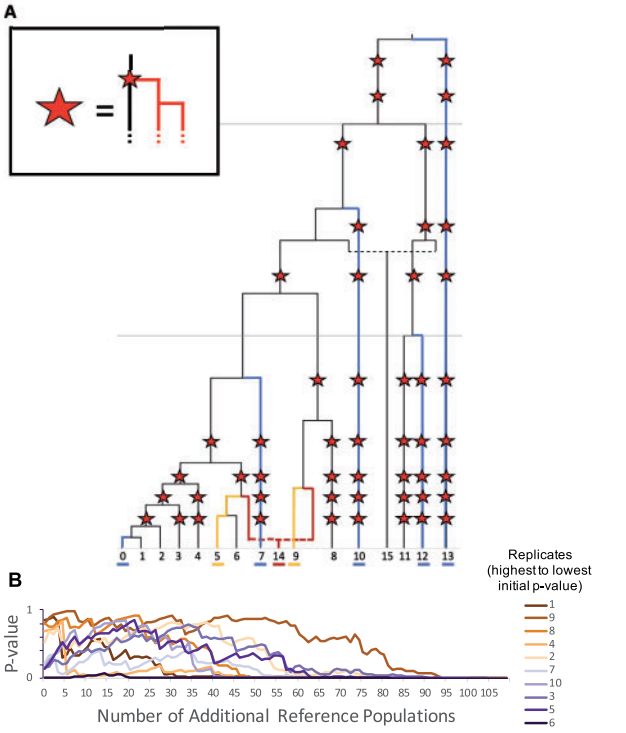

Figure 9 Inclusion of a large number of reference populations. (A) Population history of simulated data with additional populations added to tree. In all positions in the population history marked by a star, a population branching event occurs, forming an additional population. This new lineage undergoes an additional population branching event 50 generations later, resulting in two new populations created at each location marked with a star. Colors indicate the populations used in the base model, with the target in red, sources in yellow, and initial references shown in blue. (B) The change in *P*-values assigned to each model by qpAdm as additional reference populations are randomly added to the model. Each line tracks the *P*-values assigned to a single replicate [colors indicate initial *P*-value, ordered from highest (dark brown) to lowest (dark purple)], as the number of additional reference populations added to the base set of reference populations increases from 0 to 108.

model of migration, in which neighboring populations exchange migrants each generation with rate m (Kimura and Weiss 1964). We simulated a population history based on this migration model (Figure 11A), where six populations (each with an effective population size of 5,000) split from a common ancestral population 1000 generations previously, after which point migration occurred between neighboring populations. The model also includes three additional populations that are symmetrically related to these six populations, with all nine lineages splitting from a common ancestral population 2000 generations in the past (see Supplementary File S5 for exact simulation parameters).

While under this model, populations 1 and 3 have each contributed ancestry to population 2, it would be inaccurate to say that population 2 is the product of admixture between these two populations. The duration of exchange of ancestry is much longer than what is supposed in qpAdm. In addition, population 2 was formed in the same population-splitting event that formed populations 1 and 3, not as the result of admixture between distinct populations 1 and 3. Finally, by symmetry population 2 is just as much the source of populations 1 and 3 as either of these is the source of population 2.

Preliminary analyses of the relationships between these nine populations using pairwise F_{ST} (Patterson *et al.* 2006) would indicate that population 2 is closely related to both populations 1 and 3 (Figure 11B; Supplementary Table 36). Furthermore, if populations 0–5 are plotted using PCA (Figure 11C; Supplementary Table S37) (Patterson *et al.* 2006), population 2 appears to fall on a genetic cline between these two populations. These results could be interpreted as suggestions that population 2 is the product of admixture between populations 1 and 3. While it might be possible using other *f*-statistics to determine that the relationship between these populations is not well described by a pulse admixture event (Lipson 2020), there is nothing to prevent a user from attempting to model this relationship as the product of admixture using qpAdm. We therefore explore the effects of attempting to model the ancestry of population 2 (the target population) as the product of admixture between populations 1 and 3 (the source populations), with populations 0, 4, 6, 7, and 8 classified as reference populations.

We first consider the case of a very high migration rate ($m = 0.01$; equivalent to 100 migrants moving from one population to the neighboring population per generation). Out of 20 replicates, qpAdm identifies the proposed model as plausible in 90% of cases, suggesting that qpAdm cannot always distinguish between population histories that involve continuous migration and those involving pulses of admixture. Furthermore, qpAdm assigns admixture proportions of approximately 50% to each source population, which is sensible because each population does contribute roughly equal amounts of ancestry to the target population (Figure 11, D and E; Supplementary Table S38). When we consider lower migration rates ($m = 0.001$ and $m = 0.0001$), we observe similar admixture proportion estimates, but all of the *P*-values fall well below the 0.05 threshold, suggesting that with lower rates of migration, qpAdm will reject admixture as a plausible model when the actual history involves continuous migration.

These results suggest that users should be sure to consider alternative demographic models to pulse admixture, even in cases when qpAdm produces admixture proportion estimates and *P*-values that appear plausible. These continuous migration scenarios are likely not the only cases in which qpAdm identifies plausible admixture models for populations that were not formed via admixture (or entirely via admixture), therefore, we caution that users should use additional tools, in conjunction with or prior to qpAdm analysis, to determine whether admixture is a likely demographic scenario.

This stepping-stone model can also be used to highlight an interesting feature of qpAdm, which is that admixture proportion estimates that fall outside the bounds of 0–1 may also be informative about the history of the population being modeled. It has previously been suggested that in cases where the estimated admixture proportion exceeds 1, this is indicative of the target population falling in a more extreme position along a genetic cline than either of the modeled source populations (Lazaridis *et al.* 2017). We confirm this to be true by attempting to model population 1 as the product of admixture between source populations 2 and 3 (Supplementary Figure S14). In this model, an estimated alpha of 1 would indicate that population 1 could be modeled as deriving 100% of its ancestry from population 2. Instead, we observe that all of the estimates of alpha all fall outside the bounds of 0–1, instead centering around 2, supportive of population 1's more extreme position along the genetic cline that also includes populations 2 and 3.

Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures,

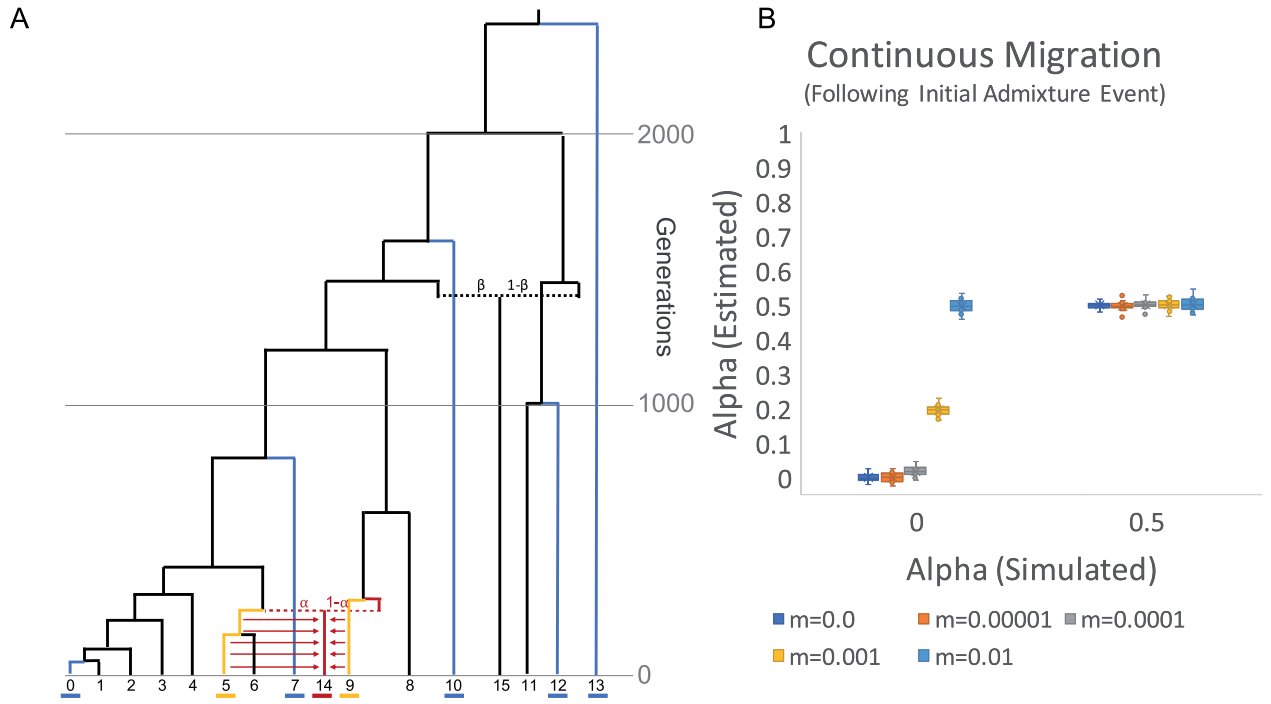

Figure 10 Continuous gene flow following an initial pulse admixture event. (A) Population history where the standard tree has been modified so that the initial admixture event occurs at generation 240, followed by continuous gene flow from populations 5 and 9 into population 14. (B) Admixture proportions estimates generated by a qpAdm model with population 14 as the target, and populations 5 and 9 as sources, with populations 13, 12, 10 7 and 0 as references when applied to the simulated data described in panel (A) with varying alpha ($\alpha = 0$ and 0.50) and subsequent gene flow occurring at varying migration rates ($m = 0.0$, 0.00001, 0.0001, 0.001, and 0.01). Error bars indicate 1 standard error.

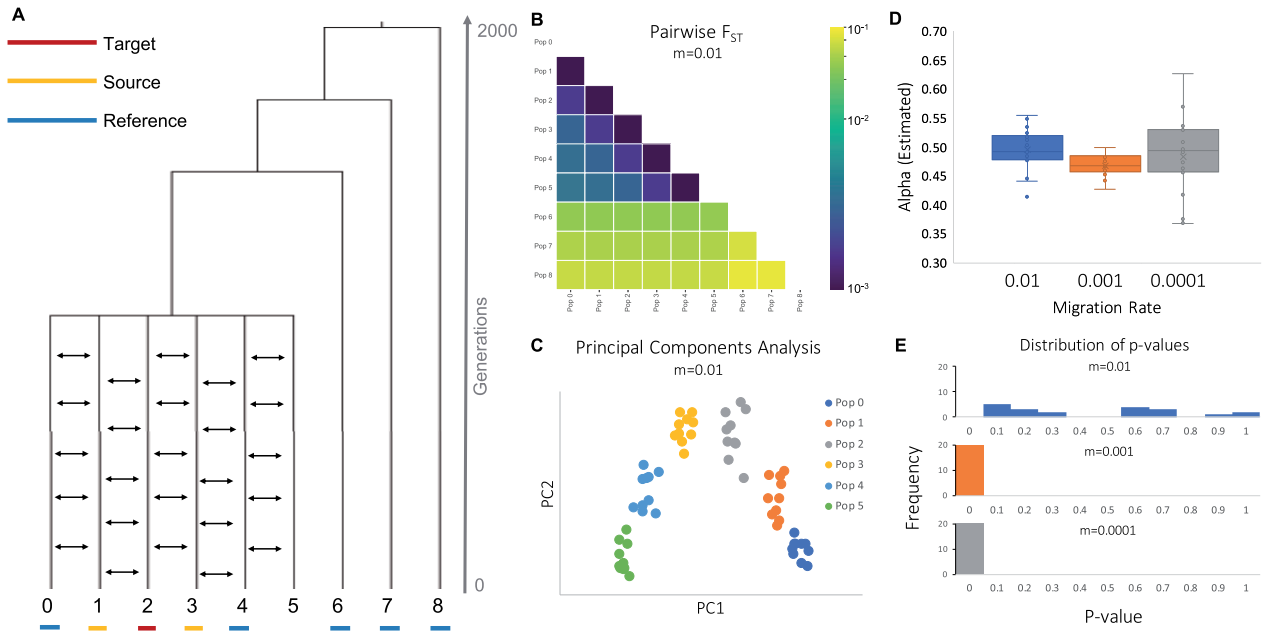

Figure 11 Continuous migration models. (A) Population history involving continuous migration. The target, source, and reference populations underlined in red, yellow, and blue, respectively. (B) A heatmap showing average pairwise FST between each population for 20 replicates (C) A PCA plot showing the relationship between all populations, calculated using a single replicate (D) Admixture proportions assigned by qpAdm for a model with population 2 as the target, and populations 1 and 3 as sources at varying migration rates. (E) Histograms showing the frequency of *P*-values produced by this qpAdm model at varying migration rates.

tables, and supplementary materials. Code used to generate the simulated data is provided in Supplementary Files S1–S5.

Supplemental material available at figshare DOI: https://doi.org/10.25386/genetics.13403225.

Discussion

We find that qpAdm can accurately identify plausible admixture models and estimate admixture proportions when applied to simulated data, matching previous theoretical expectations (Haak *et al.* 2015). When an appropriate admixture model is suggested, qpAdm calculates *P*-values that follow a uniform distribution, suggesting that a cut-off value of 0.05 will result in the acceptance of a correct model in 95% of cases. Additionally, qpAdm estimates admixture proportions with high accuracy, even when calculated on datasets with a limited number of SNPs, high rates of missingness or damage (when occurring at similar rates in all populations), or when analyses are performed on pseudo-haploid data or on data that is subject to strong ascertainment bias. Additionally, while the use of populations with small sample sizes does increase the variance in admixture proportion estimates, admixture proportion estimates appear unbiased.

Furthermore, we tested two commonly used strategies for identifying the best admixture model using qpAdm—base and rotating—and find that both strategies can distinguish between plausible and implausible models. However, the rotating strategy is better able to distinguish between plausible and implausible models, particularly when the potential source populations are closely related. We therefore recommend users implement a rotating model comparison strategy when possible. It is important to note that the results from qpAdm are always going to depend on the availability of samples. Thus, even if the rotating strategy points to one particular model as the optimal model for a given dataset, this should not be taken as proof that the source populations identified are the true sources populations. For example, in Figure 1, if data were available from population 8 and not from population 9, the rotating model would identify populations 5 and 8 as the optimal sources of population 14. This would be correct, given the samples available, but it would come as no surprise if data from population 9 subsequently became available and it was deemed a better source than population 8. A number of examples exist in which previously identified qpAdm models have been refined when ancient DNA from new populations has become available, including in the Levant (Haber *et al.* 2017; Harney *et al.* 2018) and Sardinia (Haak *et al.* 2015; Chiang *et al.* 2018; Fernandes *et al.* 2020; Marcus *et al.* 2020).

While qpAdm's ability to identify the optimal admixture model is affected by data quality, including the amount of missing data, the number of individuals in an analysis population, and the rate of ancient DNA damage, none of these factors ever bias qpAdm toward accepting a nonoptimal model and rejecting the optimal model. Instead, we find that high rates of missing data or small sample size may make it more likely for qpAdm to accept multiple models, particularly in cases where both of these factors are present. On the other hand, ancient DNA damage appears to cause qpAdm to be too stringent when it occurs at differential rates in the target and optimal source populations, often rejecting models that should be considered optimal, and resulting in biased admixture proportion estimates. While these results show that improving data quality and carefully curating data prior to analysis should be a priority of qpAdm users, they are promising as they suggest that data quality issues are unlikely to causes users to infer an incorrect model of admixture using qpAdm.

Although we find that the performance of qpAdm matches theoretical predictions under standard conditions, we also highlight several cases in which users should exercise caution. For instance, we show that users should attempt to avoid choosing source populations that have experienced gene flow since their split with the lineage that contributed admixture to the target population, or in cases where this is unavoidable, avoid including the populations that contributed to this gene flow event as reference populations. They should also limit the number of reference populations included in a qpAdm model, as the inclusion of too many reference populations may result in lowered *P*-values. Furthermore, we show that qpAdm may produce plausible admixture proportion estimates and *P*-values in cases where the population of interest was not formed via admixture, such as the case of continuous migration, therefore users should be careful to consider whether alternative demographic models may better explain their data.

Overall, we find that qpAdm is a useful tool for identifying plausible admixture models and estimating admixture proportions, and that its performance matches theoretical expectations. qpAdm is particularly useful because it can be used in cases where the underlying population history of all the populations included in the analysis is difficult to determine and can therefore be used in cases where it may not be possible to use other tools for modeling population histories that involve admixture, like qpGraph and TreeMix. We include an updated user guide for qpAdm in Supplementary Material S1 in order to make this method more accessible to future users.

Acknowledgments

We are grateful for Dan Hartl, who raised critical questions about qpAdm that inspired and guided this project. We also thank Iosif Lazaridis and Mark Lipson for their advice and helpful suggestions. Additionally, we are grateful to the three reviewers whose comments substantially increased the quality of this manuscript.

Funding

E.H. was supported by a graduate student fellowship from the Max Planck-Harvard Research Center for the Archaeoscience of the Ancient Mediterranean (MHAAM). D.R. was funded by NIH grant GM100233 and John Templeton Foundation grant 61220, and is an Investigator of the Howard Hughes Medical Institute.

Conflicts of interest

None declared.

Literature cited

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature. 467: 1061–1073.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664.

Chiang CW, Marcus JH, Sidore C, Biddanda A, Al-Asadi H, *et al.* 2018. Genomic history of the Sardinian population. Nat Genet. 50: 1426–1434.

Dabney J, Meyer M, Pääbo S. 2013. Ancient DNA damage. Cold Spring Harbor Perspect Biol. 5:a012567.

de Barros Damgaard P, Marchi N, Rasmussen S, Peyrot M, Renaud G, *et al.* 2018a. 137 ancient human genomes from across the Eurasian steppes. Nature. 557:369–374.

de Barros Damgaard P, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, *et al.* 2018b. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science. 360: eaar7711.

Fernandes DM, Mittnik A, Olalde I, Lazaridis I, Cheronet O, *et al.* 2020. The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean. Nat Ecol Evol. 4:334–345.

Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, *et al.* 2015. An early modern human from Romania with a recent Neanderthal ancestor. Nature. 524:216–219.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, *et al.* 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 522:207–211.

Haber M, Doumet-Serhal C, Scheib C, Xue Y, Danecek P, *et al.* 2017. Continuity and admixture in the last five millennia of Levantine history from ancient Canaanite and present-day Lebanese genome sequences. Am J Hum Genet. 101:274–282.

Hajdinjak M, Fu Q, Hübner A, Petr M, Mafessoni F, *et al.* 2018. Reconstructing the genetic history of late Neanderthals. Nature. 555:652–656.

Harney É, May H, Shalem D, Rohland N, Mallick S, *et al.* 2018. Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation. Nat Commun. 9:3336.

Harney É, Nayak A, Patterson N, Joglekar P, Mushrif-Tripathy V, *et al.* 2019. Ancient DNA from the skeletons of Roopkund Lake reveals Mediterranean migrants in India. Nature Commun. 10:1–10.

Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol. 12:e1004842.

Kimura M, Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics. 49:561–576.

Lazaridis I, Mittnik A, Patterson N, Mallick S, Rohland N, *et al.* 2017. Genetic origins of the Minoans and Mycenaeans. Nature. 548: 214–218.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, *et al.* 2016. Genomic insights into the origin of farming in the ancient Near East. Nature. 536:419–424.

Lipson M. 2020. Applying f4 -statistics and admixture graphs: Theory and examples. Mol Ecol Resour. **20**:1658–1667.

Marcus JH, Posth C, Ringbauer H, Lai L, Skeates R, *et al.* 2020. Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia. Nat Commun. 11:1–14.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, *et al.* 2015. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 528:499–512.

Narasimhan VM, Patterson NJ, Moorjani P, Lazaridis I, Mark L, *et al.* 2019. The formation of human populations in South and Central Asia. Science. 365:eaat7487.

Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, *et al.* 2018. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature. 555:190–196.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, *et al.* 2012. Ancient admixture in human history. Genetics. 192:1065–1093.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.

Peter BM. 2016. Admixture, population structure, and F-statistics. Genetics. 202:1485–1501.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8: e1002967.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, *et al.* 2012. Reconstructing native American population history. Nature. 488: 370–374.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature. 461:489–494.

Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, *et al.* 2017. Reconstructing prehistoric African population structure. Cell. 171:59–71.e21.

Soraggi S, Wiuf C. 2019. General theory for stochastic admixture graphs and F-statistics. Theor Popul Biol. 125:56–66.

Winther RG, Giordano R, Edge MD, Nielsen R. 2015. The mind, the lab, and the field: three kinds of populations in scientific practice. Stud Hist Philos Biol Biomed Sci. 52:12–21.

Communicating editor: J. Novembre