

# 1 Testing Times: Challenges in Disentangling Admixture Histories in Recent and 2 Complex Demographies

3  
4 Matthew P. Williams<sup>1\*</sup>, Pavel Flegontov<sup>2,3</sup>, Robert Maier<sup>3</sup>, Christian D. Huber<sup>1\*</sup>

5  
6 1: Pennsylvania State University, Department of Biology, University Park, PA 16802, USA

7 2: Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czechia

8 3: Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA

9  
10 \* Correspondence to Matthew P. Williams: [mkw5910@psu.edu](mailto:mkw5910@psu.edu), and Christian D. Huber: [cdh5313@psu.edu](mailto:cdh5313@psu.edu)

## 11 12 Abstract

13 Paleogenomics has expanded our knowledge of human evolutionary history. Since the 2020s, the study of  
14 ancient DNA has increased its focus on reconstructing the recent past. However, the accuracy of paleogenomic  
15 methods in answering questions of historical and archaeological importance amidst the increased demographic  
16 complexity and decreased genetic differentiation within the historical period remains an open question. We used  
17 two simulation approaches to evaluate the limitations and behavior of commonly used methods, qpAdm and the  
18  $f_3$ -statistic, on admixture inference. The first is based on branch-length data simulated from four simple  
19 demographic models of varying complexities and configurations. The second, an analysis of Eurasian history  
20 composed of 59 populations using whole-genome data modified with ancient DNA conditions such as SNP  
21 ascertainment, data missingness, and pseudo-haploidization. We show that under conditions resembling  
22 historical populations, qpAdm can identify a small candidate set of true sources and populations closely related  
23 to them. However, in typical ancient DNA conditions, qpAdm is unable to further distinguish between them,  
24 limiting its utility for resolving fine-scaled hypotheses. Notably, we find that complex gene-flow histories  
25 generally lead to improvements in the performance of qpAdm and observe no bias in the estimation of  
26 admixture weights. We offer a heuristic for admixture inference that incorporates admixture weight estimate and  
27  $P$ -values of qpAdm models, and  $f_3$ -statistics to enhance the power to distinguish between multiple plausible  
28 candidates. Finally, we highlight the future potential of qpAdm through whole-genome branch-length  $f_2$ -statistics,  
29 demonstrating the improved demographic inference that could be achieved with advancements in  $f$ -statistic  
30 estimations.

31  
32 **Keywords:** aDNA, archaeogenetics, paleogenomics, qpAdm,  $f$ -statistics, admixture  
33

## 34 Introduction

35 Beginning over a decade ago, the genome sequencing and analysis of ancient specimens, so-called ancient  
36 DNA (aDNA), spawned the field of paleogenomics and has provided novel insights into our understanding of  
37 population demographic history for a diversity of organisms and contexts (Brunson and Reich 2019; Spyrou *et al.*  
38 *et al.* 2019; De Schepper *et al.* 2019; Arning and Wilson 2020; Mitchell and Rawlence 2021; Wibowo *et al.* 2021).  
39 No species has gained deeper insights from the aDNA revolution than humans, as it has significantly unraveled  
40 our complex evolutionary and migratory histories (Haber *et al.* 2016; Slatkin and Racimo 2016; Fu *et al.* 2016;  
41 Llamas *et al.* 2017; Williams and Teixeira 2020; Liu *et al.* 2021; Ávila-Arcos *et al.* 2023). Much of the research in  
42 human paleogenomics during the early 2010s was focused on reconstructing human prehistory (dating back  
43 more than 5k years before the present (YBP)) (Figure 1A). It was during these years that many of the statistical  
44 methods and software that have since become the foundation of aDNA studies were developed and have been  
45 pivotal in defining our understanding of human prehistory. These methods range from model-free exploratory  
46 approaches such as the smartpca implementation of principal component analysis (PCA) (Patterson *et al.* 2006;  
47 Reich *et al.* 2008; McVean 2009), and the ADMIXTURE software (Alexander *et al.* 2009), to statistical tests of  
48 admixture such as  $f_3$ - and  $f_4$ -statistics (Reich *et al.* 2009; Patterson *et al.* 2012), and the related D-statistics  
49 (Green *et al.* 2010; Durand *et al.* 2011), which leverage deviations from expected allele sharing patterns to  
50 reject simple trees and suggest more complex relationships. In addition, various downstream software has been  
51 developed to elucidate more complex relationships among numerous groups, with many utilizing  $f$ -statistics.  
52 Examples include qpAdm, which models a target population as a mixture of several proxy ancestry sources  
53 (Haak *et al.* 2015; Harney *et al.* 2021); qpWave, analyzing the number of gene flow events between population  
54 sets (Reich *et al.* 2012); and qpGraph, MixMapper, TreeMix, AdmixtureBayes, and findGraphs, all creating  
55 representations of admixture histories as directed acyclic graphs (Patterson *et al.* 2012; Pickrell and Pritchard  
56 2012; Lipson *et al.* 2013, 2014; Nielsen *et al.* 2023; Maier *et al.* 2023). To a large degree, the reliance on these  
57 methods has been because of their use of allele frequencies which is suitable for pseudo-haploid aDNA  
58 whereby calling diploid genotypes is often infeasible due to its highly degraded characteristics.

59  
60 Since the 2020s there has been a shift in aDNA research to studying the more recent past (Figure 1A). As a  
61 result, aDNA is increasingly used to address questions of archaeological and historical relevance. This research  
62 field was named archaeogenetics by British archaeologist Colin Renfrew (Boyle and Renfrew 2000). The  
63 historical period, particularly in Southwest Asia, is broadly demarcated to begin somewhere around the early-  
64 mid-3rd millennium BCE (Bartash 2020) and is characterized by the invention of writing, and intermittent periods  
65 of intensified inter-regional trade, diplomacy, and human mobility (Kristiansen 2016). From this body of  
66 research, hypotheses about gene flows between ancient settlements amenable to aDNA can involve groups  
67 separated by very short periods and thought to have descended from a complex web of migration and  
68 population structure (Haak *et al.* 2015; Lazaridis *et al.* 2016, 2017, 2022a; b; Haber *et al.* 2017, 2020; de Barros  
69 Damgaard *et al.* 2018; Harney *et al.* 2018; Wang *et al.* 2019; Narasimhan *et al.* 2019; Antonio *et al.* 2019;

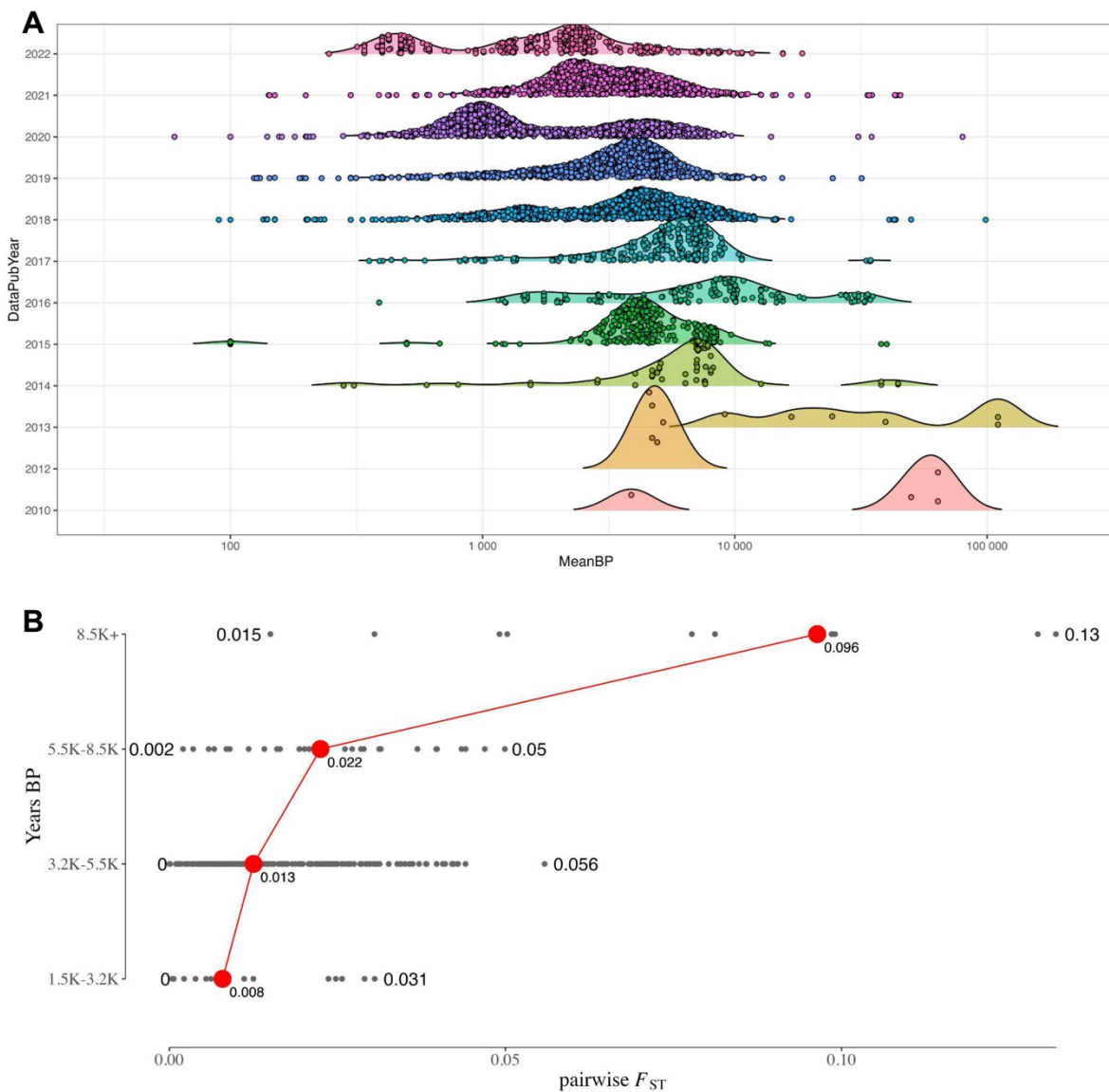
70 Fernandes *et al.* 2020; Agranat-Tamir *et al.* 2020; Skourtanioti *et al.* 2020, 2023; Clemente *et al.* 2021; Koptekin  
71 *et al.* 2023; Schmid and Schiffels 2023; Moots *et al.* 2023). These can range from questions regarding the  
72 degree of population continuity between periods of cultural change or settlement hiatus in the archaeological  
73 record to determining if cultural links between regions are indicative of inter-regional migration, and assessing if  
74 historical records of mass migrations and forced relocations result in observable signals of increased inter-  
75 regional gene flow. A common thread underlying these questions is, for a population of interest, to what extent  
76 can aDNA accurately reconstruct their genetic history, and importantly, reject false models of ancestry  
77 composed of closely related candidate populations? Moreover, what limits and possible biases emerge with the  
78 increase in demographic complexity amongst candidate source populations, a reduction in the number of  
79 generations separating aDNA samples and their ancestral admixture events, and an overall decrease in genetic  
80 differentiation indicative of the historical period? While the theoretical behavior of  $f$ - and D-statistics has been  
81 extensively tested (Patterson *et al.* 2012; Martin *et al.* 2015; Peter 2016, 2022; Harris and DeGiorgio 2017;  
82 Zheng and Janke 2018; Soraggi and Wiuf 2019; Tricou *et al.* 2022), and the performance of the commonly used  
83 software qpAdm thoroughly assessed under simple demographic models with both pulse-like and continuous  
84 migration (Ning *et al.* 2020; Harney *et al.* 2021), their behavior under varying degrees of population  
85 differentiation and complex demographic history expected of populations within the historical period remains  
86 underexplored.

87 In this study, we conducted a simulation-based evaluation of two widely used methods for reconstructing  
88 admixture histories - the "admixture"  $f_3$ -statistic and the qpAdm software (Figure 2). Our goal was to understand  
89 their effectiveness and limitations, particularly in complex scenarios that arise during the reconstruction of  
90 historical population dynamics. We started by simulating two chromosomes of combined length  $\sim 491$  Mbp  
91 under four simplistic and qualitatively different admixture graphs, aiming to explore a broad range of model  
92 parameters leading to widely varying degrees of genetic differentiation. Subsequently, we expanded our  
93 evaluation to include a complex demography representative of a model of Eurasian human history emerging  
94 from a series of recent publications, which comprised 59 populations and 41 pulse admixture events. We  
95 simulated 50 whole-genome ( $L \sim 2875$  Mbp) replicates and processed the simulated data to mimic typical aDNA  
96 conditions, including a Human-Origins-like SNP ascertainment scheme, empirical data missingness  
97 distributions, and pseudo-haploidization.

98  
99 Importantly for the study of the historical period, our findings illustrate that qpAdm converges on a small subset  
100 of plausible models for an admixed target group consisting of the true sources and closely related populations  
101 by the time the  $F_{ST}$  levels reach those observed in Bronze and Iron Age Southwest Asian populations. However,  
102 under these divergence levels and conditions typical of aDNA, we observe qpAdm has limited ability to  
103 definitively answer fine-scaled questions relevant for archaeologists and historians due to lack of power to reject  
104 all non-optimal ancestry sources minimally differentiated from true ones. Moreover, for historical populations  
105 with complex gene-flow histories, we show that whilst admixture to source populations generally improves the  
106 performance of qpAdm, the phylogenetic origin of this admixture in ancestral source groups differentially

107 impacts qpAdm accuracy and performance. We show that the number of generations post admixture has no  
108 impact on qpAdm performance or accuracy of admixture proportion (“admixture weight”) estimates. However,  
109 we observe when selecting sub-optimal ancestry sources that the admixture weights are biased in favor of the  
110 population that is most similar to the true source. We assessed several model plausibility criteria commonly  
111 used in the aDNA literature and show that each criterion impacts the performance and accuracy of qpAdm  
112 differently under various demographic conditions. Additionally, we highlight problems that users should be  
113 aware of when applying additional plausibility criteria for qpAdm models, such as negative admixture  $f_3$ -statistics  
114 or the rejection of all simpler qpAdm models, as they can lead to an increase in type II errors. Finally, we offer  
115 an interpretative heuristic guide that can enhance the power to distinguish between multiple plausible qpAdm  
116 models, thereby contributing to more robust and reliable archaeogenetic analyses.

117  
118  
119 **Figure 1**



120



Dates of published aDNA samples. (A) A per-publication-year transect of the density of the ( $\log_{10}$ ) age of published ancient genomes. The publication dates and number of samples were taken from the Allen Ancient DNA Resource (AADR) v.52.2. (B) The temporal transect of population differentiation levels in Southwest Asia. The average dates for each sample in years BP were taken from the AADR v.52.2. For the plot in panel B, they were grouped into four epochs, with 3.2k years BP approximating the start of the Iron Age, 5.5k years BP approximating the start of the Bronze Age, 8.5k years BP approximating the start of the Neolithic period, and older years representing the Paleolithic period. The  $F_{ST}$  values were calculated using the Eigensoft v8.0.0 smartpca software.

## Methods and Results

### Starting simple: Insights into the behaviors of qpAdm and $f_3$ -statistic from simplistic demographic models

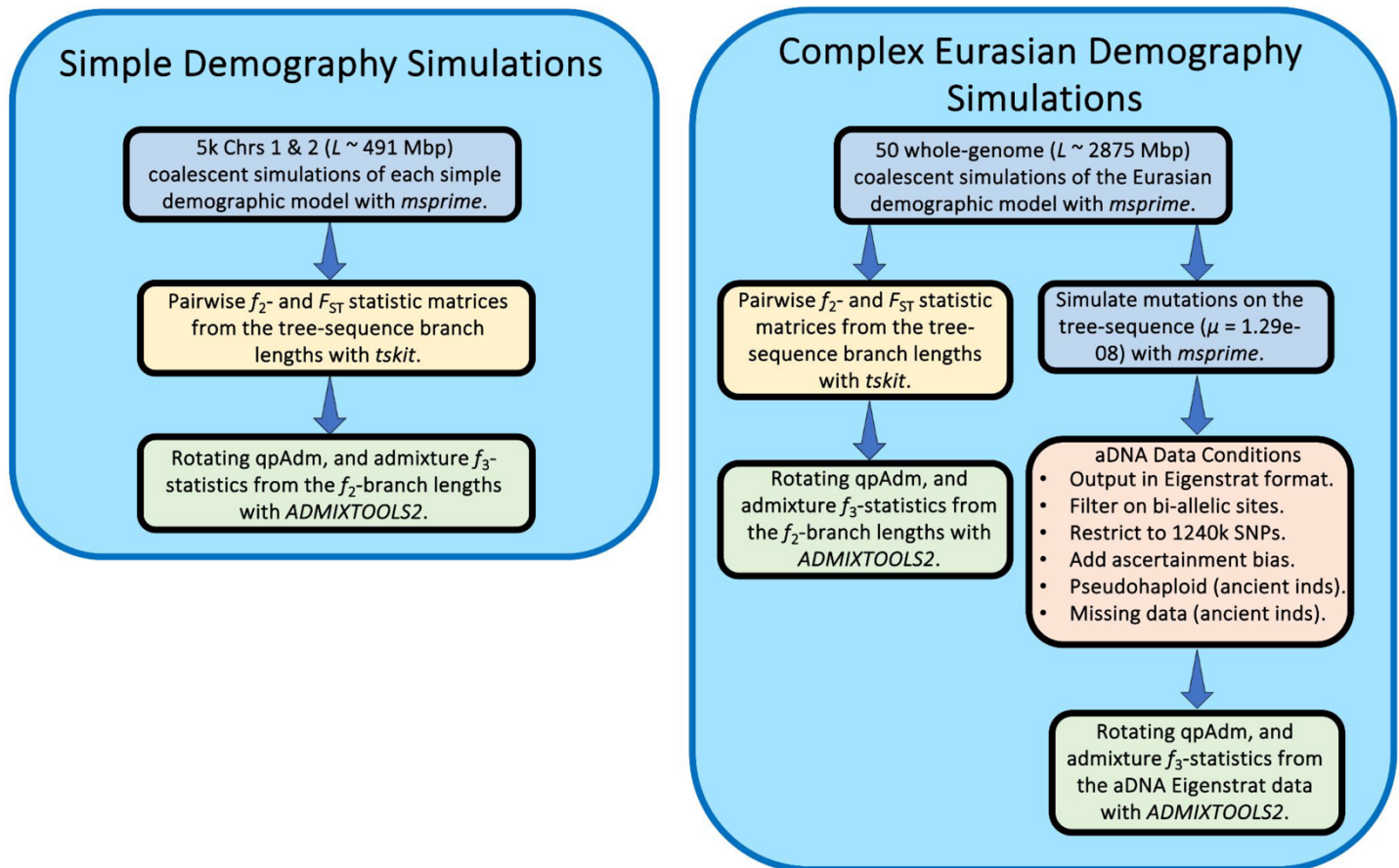
To obtain a baseline understanding of how specific demographic models and parameters impact downstream population genetic inference with qpAdm and the  $f_3$ -statistic, we formed simple bifurcating trees with varying scales of population divergence and augmented them with one to three gene flows in qualitatively different configurations (Figure 3A-D). For the simplest bifurcating demographic model with one admixture event (hereafter Model 1; Figure 3A), we randomly sampled values of five split-time parameters ( $T_1$ ,  $T_2$ ,  $T_3$ ,  $T_4$ , and  $T_{\text{admix}}$ ) from uniform distributions generated by the following framework:

- The oldest variable split-time ( $T_1$ ) was selected first from a window between four generations in the past and the fixed  $T_0$  split-time (6896 generations).
- The  $T_2$  split-time parameter was sampled between three generations in the past and the sampled  $T_1$  split-time.
- We selected the  $T_3$  and  $T_4$  split-time parameters from a window between two generations in the past and the  $T_2$  split-time parameter.
- The  $T_{\text{admix}}$  (admixture date) parameter was selected from a window between a single generation in the past and the minimum of the  $T_3$  and  $T_4$  split-time parameters.
- We randomly sampled the admixture weight parameter ( $\alpha$ ), which forms the Target population as a mixture of the Source-1 (proportion  $\alpha$ ) and Source-2 (proportion  $1 - \alpha$ ) populations, from a uniform distribution between zero and one (the distributions of simulated parameter values and scatter-plot matrices of simulation parameter correlations can be found in Supplementary Figure SI Figure S1A-B).

To assess the impact on admixture inference of more complex admixture history in one of proxy ancestry sources, we configured three additional demographic Models, each building upon the structure of Model 1 as follows:

- Model 2 includes a gene flow from an outgroup (R3 branch) into the source (S1).
- Model 3 includes admixture into the source (S1) from an internal branch ancestral to both the S2 and R2 populations (iS2R2).
- Model 4 combines the admixture events from Models 2 and 3, with no constraint on their order.

**Figure 2**



Simulation and analysis workflow in our study.

## Data generation

For each of the four simple demographic Models (Figure 3A-D), we used msprime v.1.2.0 (Kelleher *et al.* 2016; Baumdicker *et al.* 2021) to simulate 5000 iterations of succinct tree sequences without mutations with each iteration sampling demographic parameters from the schema outlined above. For the first 100 generations into the past we simulated under the Discrete Time Wright-Fisher model (DTWF) (Nelson *et al.* 2020), and then under the Standard (Hudson) coalescent model until the most recent common ancestor (MRCA). We used sequence lengths and recombination rates approximating human chromosomes one ( $L = \sim 2.49 \times 10^8$  bp, and  $r =$

~ $1.15 \times 10^{-8}$  per bp per generation) and two ( $L = 2.42 \times 10^8$ , and  $r = 1.10 \times 10^{-8}$ ) (Adrion *et al.* 2020; Elise Lauterbur *et al.* 2022), and separated each chromosome with a  $\log(2)$  recombination rate following guidelines in the msprime manual (<https://tskit.dev/msprime/docs/stable/ancestry.html#multiple-chromosomes>). For each demographic model, we fixed an upper bound split time of 200,000 years, and a generation time of 29 years, and for all populations, an effective size ( $N_e$ ) of 10,000 and a sample size of 20 diploid individuals taken at the leaves.

We generated  $f_2$  - and  $F_{ST}$  - statistic matrices directly from the tree sequences through tskit v.0.5.2 with parameters “Mode=branch”, and “span\_normalise=True”, using 5 Mbp windows. The resulting  $f_2$  - statistics matrix was used for qpAdm analyses with parameters “full\_results=TRUE”, and “fudge\_twice=TRUE”, and for calculating admixture  $f_3$ -statistics in the ADMIXTOOLS2 software (Maier *et al.* 2023). For the qpAdm rotating protocol following Harney *et al.* (2021), we included S1, S2, R1, R2, R3, and R4 as alternative sources and “outgroups” (“right” populations), resulting in six single-source, and 15 two-source models. We computed admixture  $f_3$ -statistics on pairwise combinations of the S1, S2, R1, R2, R3, and R4 populations, resulting in an  $f_3$ -statistic test for each of the 15 two-source qpAdm models. The simple simulations resulted in genetic diversity estimates that cover ranges described for all present and past populations of anatomically modern humans, with the median pairwise  $F_{ST}$  between all Source and Right populations spanning from ~0.00012 to ~0.15.

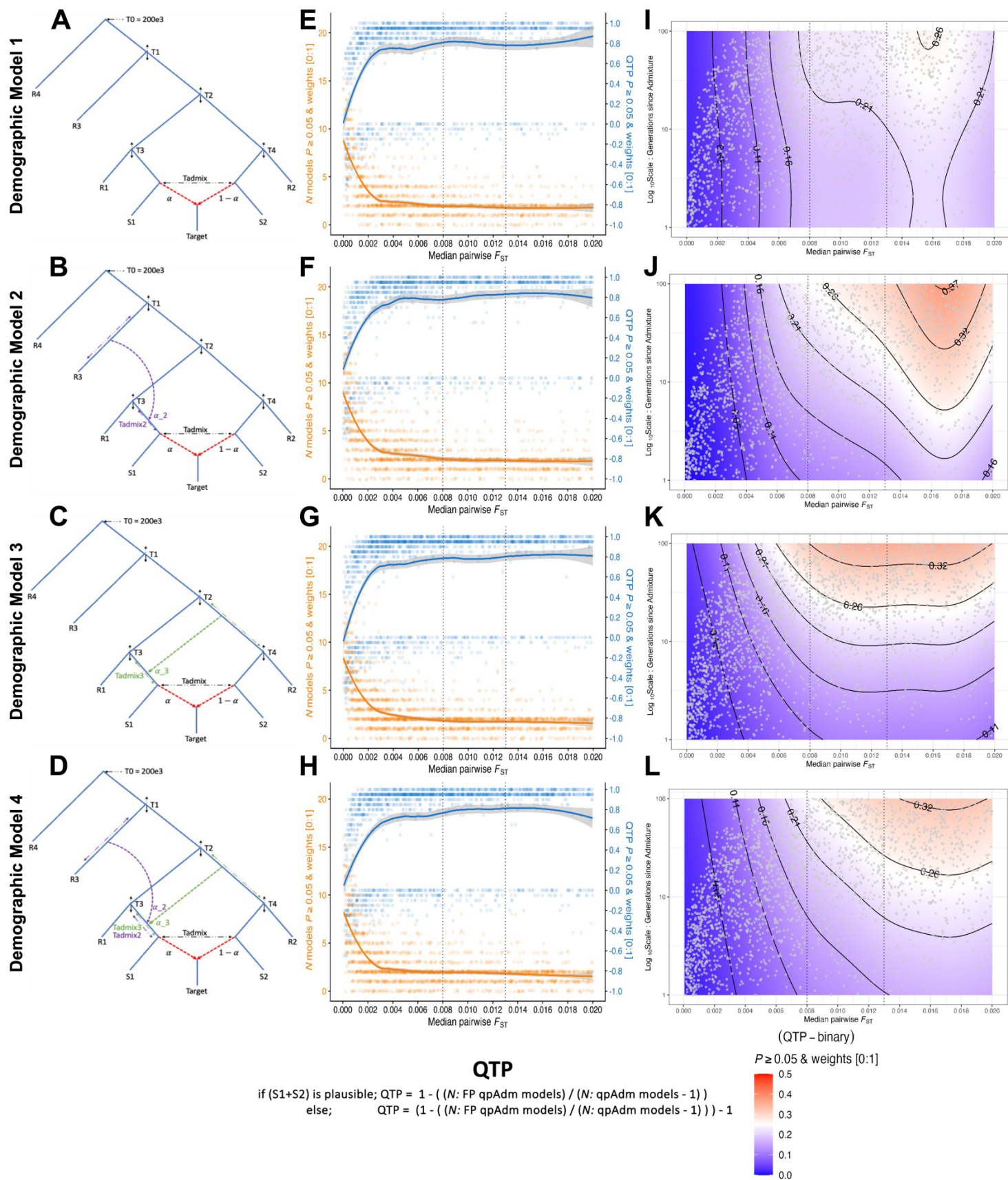
Throughout our analysis of the simple demographic Models, we refer to the pairing of the S1+S2 Source populations as the “true” model representing the ancestry of the Target population, and we refer to all other population combinations as “false” models. In evaluating the qpAdm results, unless otherwise stated, we consider plausible models to have a  $P$ -value  $\geq 0.05$  and admixture weights between zero and one ([0:1]). In addition, we configured a summary metric, “qpAdm test performance” (QTP), that conveys the precision of rotating qpAdm analyses per simulation iteration taking into account all single and two-source qpAdm models (Figure 3E-H). The range of QTP is between “+1” and “-1” where the most optimal outcome, “+1”, corresponds to the condition where all false models are rejected (single and two-source) and the true model is plausible. The worst outcome, “-1”, occurs when the true model is rejected, and all false models are considered plausible. As such, all rotating qpAdm analyses that reject the true model result in a negative QTP, and analyses that have the true model amongst the plausible qpAdm models have positive QTP values. The outcomes where all models are rejected, or all models are plausible are scored as “0”. Values between “+1” and “0” occur when both the true and false models are plausible in the same simulation, with each additional plausible false model (single and two-source) decreasing the QTP value. Likewise, values between “0” and “-1” occur when the true model is rejected, and some (but not all) false models are plausible. We also evaluated the binary QTP outcome (Figure 3I-L), whereby qpAdm either performs most optimally (i.e. rejects all false models and estimates the true model as plausible) or does not (i.e. at least one wrong model is considered plausible or the true model is rejected).



210

211

Figure 3



212

213 Simple demographic models and qpAdm test performance (QTP). (A-D) Topological structures of the four simple  
214 demographic models. (E-H) QTP and number of plausible qpAdm models across the range of median pairwise  $F_{ST}$  values  
215 calculated on the S1, S2, R1, R2, and R3 populations. For each simulation iteration we represent the counts of the number  
216 of plausible single and two-source qpAdm models (21 is the maximum possible) with orange dots and the locally estimated  
217 scatterplot smoothing (loess) computed in R and shown with the orange line. We show the QTP value for each simulation  
218 iteration with blue dots and the loess smoothing with the blue line. (I-L) Logistic GAM probability for the QTP-binary  
219 response variable with admixture date ( $T_{\text{admix}}$ ) and median pairwise  $F_{ST}$  as predictor variables. The gray dots are unique  
220 combinations of simulation parameters placed in the space of predictor variables. Vertical dotted lines in plots E-L show the  
221 median pairwise  $F_{ST}$  values at the approximate Iron (0.008), and Bronze Age (0.013) periods.  
222

## 224 The Limits of Population Differentiation for qpAdm Admixture Model Inference

225 Due to extensive admixture between ancient southwest Eurasian groups beginning around the 6th millennium  
226 BCE, populations from historical periods exhibit, on average, lower genetic differentiation than their  
227 predecessors (Figure 1B). Therefore, evaluating archaeogenetic hypotheses regarding historical migrations  
228 necessitates the ability to disentangle the admixture histories of minimally differentiated ancient groups  
229 separated by very short periods of genetic drift. To address this, we used demographic Model 1 (Figure 3A) to  
230 directly evaluate the impact and limits of population differentiation on the performance of rotating qpAdm. For all  
231 downstream demographic inference analyses, we constrained the simulated parameter space to values that  
232 approximate conditions observed amongst historical period groups such as a low median pairwise  $F_{ST}$  between  
233 0 and 0.02 computed on the S1, S2, R1, R2, and R3 populations, and  $\leq 100$  generations since the admixture  
234 event forming the Target population. Unless otherwise stated, we use this parameter range for all results  
235 described below.

### 236 *Genetic differentiation and qpAdm performance*

237 A requirement of qpAdm is that at least one right-group population is differentially related to populations in the  
238 left set (Haak *et al.* 2015; Harney *et al.* 2021) as the power of qpAdm is largely due to the right-group  
239 populations' ability to distinguish between putative ancestry sources (Harney *et al.* 2021). Consistent with this  
240 principle, we observe a general trend of increasing qpAdm performance (QTP) with larger median pairwise  $F_{ST}$   
241 values (Figure 3E). As these values approach 0.01, equivalent to that observed amongst Southwest Asian  
242 Bronze Age and older groups, we notice QTP to asymptote around 0.8 and convergence on an average of two  
243 plausible qpAdm models per simulation iteration (Figure 3E). However, as the median pairwise  $F_{ST}$  drops to  
244 values observed at the lower ends of human population differentiation ( $\sim 0.003 - 0.004$ ) we observe a sharp  
245 decline in the average QTP driven by increases in both the number of plausible false qpAdm models and  
246 rejections of the true qpAdm model (S1+S2) (Figure 3E).  
247



To analyze the distribution of plausible qpAdm models driving the QTP variation at different levels of genetic differentiation, we formed median pairwise  $F_{ST}$  bins roughly corresponding to values separating historical epochs. The smallest range,  $F_{ST}$  between 0 and 0.008, corresponds to the diversity estimated from samples dating between 1.5k to 3.2k years ago (Figure 1B) with the upper range broadly demarcating the Iron Age from the Bronze Age in Southwest Asia. The middle range,  $F_{ST}$  between 0.008 and 0.013, corresponds to the diversity estimated from samples dating between 3.2k and 5.5k years ago and encompasses the Bronze Age population diversity (Figure 1B). The upper range,  $F_{ST}$  between 0.013 and 0.02, estimated from samples dating between 5.5k and 8.5k years ago, represents the diversity present amongst populations ancestral to those of the historical period (Figure 1B). Consistent with the QTP distribution described above, the smallest  $F_{ST}$  bin contains the highest number of false plausible qpAdm models including single-source models for the target population (Figure 4A). The degree of population divergence also impacts the plausibility of the true model with larger  $F_{ST}$  bins increasing both the frequency of plausible true models (0.705, 0.859, and 0.842 for the three  $F_{ST}$  bins, respectively) and the proportion of true models out of all plausible qpAdm models (22.5%, 44.8%, and 48.7%, for the three  $F_{ST}$  bins, respectively). Notably, the increased rejection of the true model in the lowest  $F_{ST}$  bin is largely due to inaccurate estimations of the admixture weights. Approximately 50% of the true model replicates with  $P$ -values  $\geq 0.05$  are rejected due to admixture proportions outside the [0:1] range (SI Figure S2). For larger  $F_{ST}$  bins, the predominant rejection of the true model shifts to statistical significance, with the majority of true models rejected with  $P$ -values between 0.01 and 0.05 (SI Figure S2).

With the recent shift of aDNA research towards reconstructing admixture histories within sub-continental regions (Ávila-Arcos et al. 2023), understanding the limits of rejecting false sources recently split from the true ancestral source is becoming increasingly pertinent. To investigate this, we explored the limits of differentiating between the sister clades of R1 and S1, and by symmetry S2 and R2, (Figure 3A) as false and true sources in qpAdm models. As expected, qpAdm has the greatest difficulty rejecting models that combine one of the (false-source) cladal populations with one of the true sources, as combinations of S1+R2 and S2+R1 account for more than 25% of all plausible qpAdm models across all  $F_{ST}$  bins (Figure 4A). As anticipated given the topological symmetry of Model 1, the two false qpAdm models are plausible at almost equal frequency. However, we less frequently observe that both false models are plausible within the same simulation (SI Figure S3), consistent with the convergence towards an average of two plausible qpAdm models described above (Figure 3E).

To assess the relationship between genetic differentiation within putative source clades and performance of rotating qpAdm, we analyzed the joint distribution of S1-R1 and S2-R2  $F_{ST}$  values for all false qpAdm models that included one of the R1 or R2 populations. As expected, we observe on average larger  $F_{ST}$  values between the S1-R1 and S2-R2 populations for rejected false models (mean = 0.004, and median = 0.002) than plausible false models (mean = 0.002, and median = 0.001) which resulted in statistically significant differences between their respective  $F_{ST}$  distributions (Mann-Whitney U  $P$ -value  $< 0.001$ ) (SI Figure S4). Thus, our simulations

suggest that under the simple topological structure of Model 1, rotating qpAdm has the power to differentiate between closely related cladal populations, albeit with more difficulty distinguishing between putative sources separated on the order of  $F_{ST} < \sim 0.002$ .

**Table 1**

qpAdm Average Performance: Historical Simulation Parameters																
Plausibility Criteria	Model 1				Model 2				Model 3				Model 4			
	FP	FDR	QTP	QTP-binary	FP	FDR	QTP	QTP-binary	FP	FDR	QTP	QTP-binary	FP	FDR	QTP	QTP-binary
<i>P</i> -value 0.01	0.318	0.844	0.644	0.000	0.301	0.806	0.658	0.003	0.330	0.842	0.639	0.000	0.321	0.839	0.647	0.001
<i>P</i> -value 0.05	0.275	0.841	0.598	0.000	0.255	0.795	0.627	0.009	0.282	0.830	0.624	0.004	0.274	0.829	0.616	0.004
<i>P</i> -value 0.01 + weights [0:1]	0.119	0.589	0.761	0.121	0.128	0.592	0.762	0.129	0.118	0.594	0.718	0.125	0.111	0.583	0.724	0.137
<i>P</i> -value 0.05 + weights [0:1]	0.098	0.574	0.699	0.148	0.107	0.574	0.712	0.157	0.097	0.566	0.683	0.166	0.092	0.566	0.676	0.155
<i>P</i> -value 0.01 + weights [0:1] ± 2s.e	0.074	0.624	0.608	0.117	0.081	0.610	0.640	0.131	0.067	0.644	0.521	0.116	0.065	0.628	0.537	0.127
<i>P</i> -value 0.05 + weights [0:1] ± 2s.e	0.060	0.610	0.554	0.143	0.066	0.592	0.597	0.159	0.054	0.614	0.495	0.152	0.052	0.615	0.495	0.142
All single-source models rejected																
<i>P</i> -value 0.01 + weights [0:1]	0.070	0.631			0.074	0.631			0.065	0.650			0.064	0.630		
<i>P</i> -value 0.05 + weights [0:1]	0.060	0.609			0.065	0.605			0.056	0.611			0.055	0.606		
<i>P</i> -value 0.01 + weights [0:1] ± 2s.e	0.067	0.642			0.070	0.630			0.062	0.663			0.060	0.643		
<i>P</i> -value 0.05 + weights [0:1] ± 2s.e	0.057	0.622			0.060	0.606			0.051	0.625			0.050	0.623		
All single-source models rejected & significant f3-statistics																
<i>P</i> -value 0.01 + weights [0:1]	0.052	0.618			0.056	0.601			0.053	0.655			0.051	0.633		
<i>P</i> -value 0.05 + weights [0:1]	0.043	0.595			0.045	0.574			0.043	0.625			0.042	0.608		
<i>P</i> -value 0.01 + weights [0:1] ± 2s.e	0.052	0.625			0.055	0.604			0.052	0.684			0.051	0.643		
<i>P</i> -value 0.05 + weights [0:1] ± 2s.e	0.043	0.603			0.045	0.576			0.042	0.634			0.041	0.621		

Performance summaries of qpAdm rotation analysis for the four demographic models and different performance metrics. Each cell contains the average of each performance metric under different qpAdm plausibility criteria. The averages are over the parameter range of the historical period (admixture generations  $\leq 100$  and median pairwise  $F_{ST} > 0$  and  $\leq 0.02$ ). The performance metrics are as follows: FP = false positive rate, FDR = false discovery rate, QTP = qpAdm test performance, and QTP-binary = qpAdm test performance provided that only the true model fits the data. Each performance metric is evaluated under a different model plausibility criteria for the four demographic Models. Their averages are printed in each cell with the color ranging from smaller (yellow) to medium (green), and larger (purple) values. For each plausibility criteria we highlight with a red box the demographic Model that performed the best for each performance metric. For example, at the plausibility criteria of  $P$ -value  $\geq 0.01$  and the FP metric, Model 2 has the smallest value and thus performed the best and a red box around its cell.

### More Complex Admixture History of Sources Affects Demographic Inference

Often, complex ancestral relationships exist among putative historical source populations (e.g., Lazaridis et al. 2016), however, whether this is detrimental to the effectiveness of identifying admixture patterns through qpAdm remains unknown. To assess the impact of both the introduction, phylogenetic origin, and number of admixture events into the source population on admixture inference we performed 5,000 simulations on each of three demographic Models that introduce admixture to the source (S1) population (Figure 3), with all other simulation parameters remaining consistent with Model 1.

310

311 Importantly, the addition of admixture events into the S1 population does not lead to significant changes to the  
312 distribution of QTP across the  $F_{ST}$  range. Results for all demographic Models converge on a maximum average  
313 QTP of  $\sim 0.8$  and an average of two plausible qpAdm models (Figure 3E-H). We do observe subtle differences  
314 in their average performance for metrics such as False Positive Rate (FPR) =  $FP / (FP+TN)$ , False Discovery  
315 Rate (FDR) =  $FP / (FP+TP)$ , QTP, and QTP-binary (Table 1). From each simulation iteration, we computed the  
316 qpAdm FPR for each demographic Model as follows: we counted the number of plausible false qpAdm models  
317 (false positives: FP) to obtain the FP qpAdm model count. To obtain the number of true negative (TN) qpAdm  
318 models, we counted the number of rejected false qpAdm models. For example, in Model 1 simulation iteration  
319 1,998, we have an FPR of 0.8 that occurred because, of the 21 total single and two-source qpAdm models, we  
320 have 16 FP qpAdm models and four false qpAdm models were rejected ( $FP / (FP+TN) = 16 / 20$ ). We computed  
321 the FDR in the same fashion. The observation of a plausible true qpAdm model represents the true positives  
322 count (TP), meaning an FDR of 1 occurs when only false qpAdm models are plausible and 0 when only the TP  
323 qpAdm model is observed and all false qpAdm models are rejected. We then averaged these metrics to  
324 generate a summary of the overall performance for each Model under historical period parameters. No single  
325 demographic Model consistently outperforms others across all performance metrics, indicating that different  
326 admixture scenarios have varying effects on qpAdm performance and accuracy. This is further supported by the  
327 observation that across multiple model plausibility criteria (discussed further below), the average QTP, QTP-  
328 binary, and FDR consistently favor demographic Model 2, while the FPR is most frequently lowest for Model 4  
329 (Table 1). However, we note that the best-performing average qpAdm metric consistently falls within one of the  
330 more complex Models 2 to 4, suggesting that, on average, the introduction of admixture to the Source  
331 population increases qpAdm rotation performance even though it decreases overall population differentiation  
332 (both median and average  $F_{ST}$  is largest in Model 1).

333

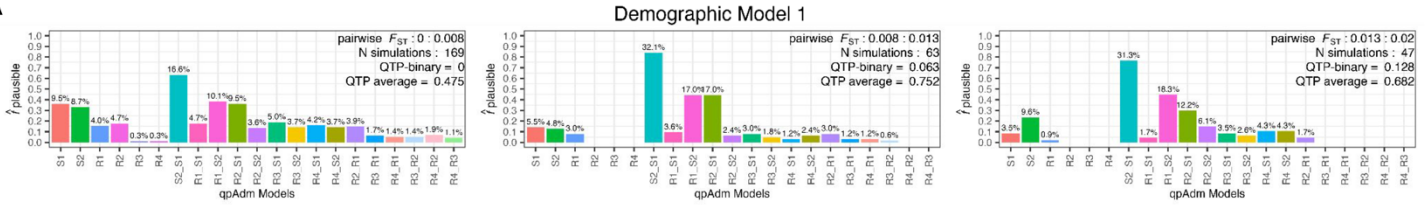
334 Similarly to Model 1, all Models with complex admixture history of S1 exhibited the highest number of false-  
335 plausible qpAdm models in the lowest range of population divergence (Figure 4). However, while we observed  
336 very similar frequencies of plausible S1-R2 and S2-R1 false qpAdm models under demographic Model 1, the  
337 introduction of admixture to the S1 population introduced an asymmetry, with the S2-R1 qpAdm model being  
338 more frequently rejected than the S1-R2 model (Figure 4B-D). Interestingly, this asymmetry is most pronounced  
339 under Model 3, which involves admixture from the common ancestor of S2 and R2 (iS2R2) to the S1 branch,  
340 and the asymmetry further increases in larger  $F_{ST}$  bins (Figure 4C). Demographic Model 4 including two  
341 admixture events in S1, displayed a distribution of false plausible models across sources intermediate between  
342 Models 2 and 3 (including one admixture event in S1) suggesting the phylogenetic source of gene flow in S1  
343 has a greater impact on the resulting plausible qpAdm models than the number of admixture events in S1  
344 (Figure 4B-D). This has important implications for the empirical study of ancient populations whose sources are  
345 themselves admixed (see Discussion).

346

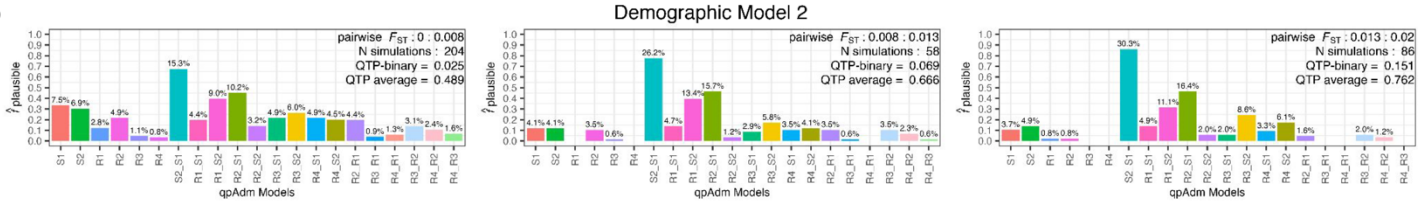
347

**Figure 4**

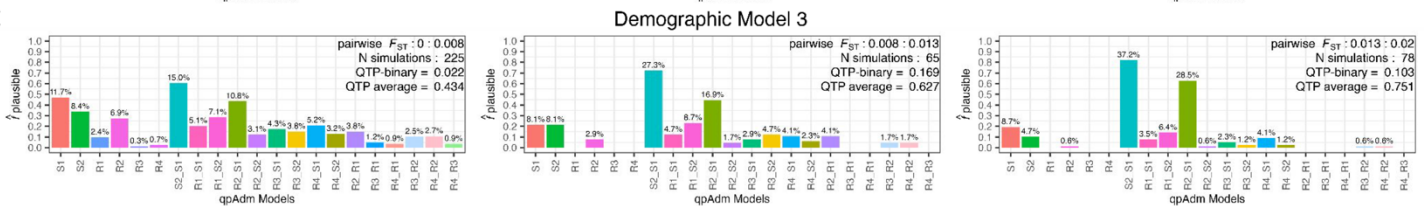
**A**



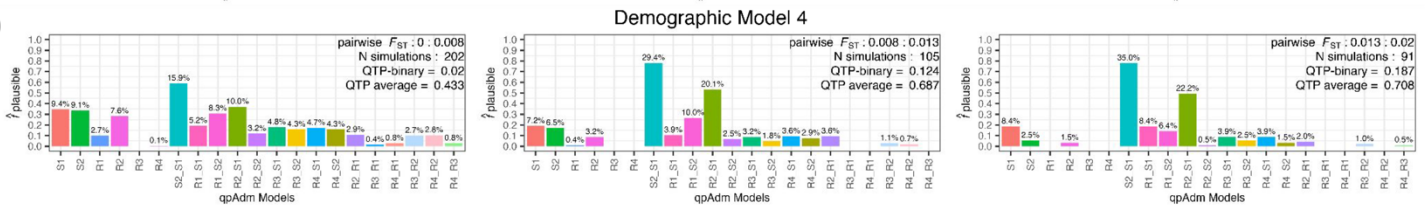
**B**



**C**



**D**



348

349

Distribution of plausible one-source and two-source qpAdm models across three population differentiation ranges (between the S1, S2, R1, R2, and R3 populations) and across the four simple demographic models. The number of generations since admixture is less than or equal to 100 in all cases. Each row represents one simulated demographic history and the columns are increasing ranges of population differentiation ( $F_{ST}$ ) corresponding to the historical period demarcations indicated in Figure 1B. The values above each barplot represent the proportion of all plausible qpAdm models within the simulation iterations for each differentiation range. The y-axis shows the frequency of each model as plausible across the total number of simulations within each differentiation range. In the top right-corner of each barplot is shown the  $F_{ST}$  range, number of simulations within that range, and the average QTP and QTP-binary.

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

A further challenge in the use of aDNA in resolving hypotheses regarding migrations during historical periods is the increased likelihood of studying recent admixture events. Such scenarios may arise in the context of detecting shifts in genetic ancestry after episodes of human migration, where only a few generations separate the timing of admixture and the ancient human individuals sampled. Moreover, the effectiveness of qpAdm in addressing historical questions that necessitate the identification of a specific population or lineage responsible for admixture is inversely proportional to the number of plausible models it identifies. We assessed the performance of qpAdm under both of these challenges by modeling the interaction between generations since admixture and population divergence on the probability of exclusively identifying the true qpAdm model using a logistic generalized additive model (GAM) in the mgcv v.1.9.0 R package (Wood 2004) with QTP-binary as the response variable, automatic smooth terms for each of the predictor variables (median pairwise  $F_{ST}$  between the S1, S2, R1, R2, R3 populations; generations since admixture), and the model parameters were estimated using restricted maximum likelihood (REML). The model's output was in the form of log-odds, which we then



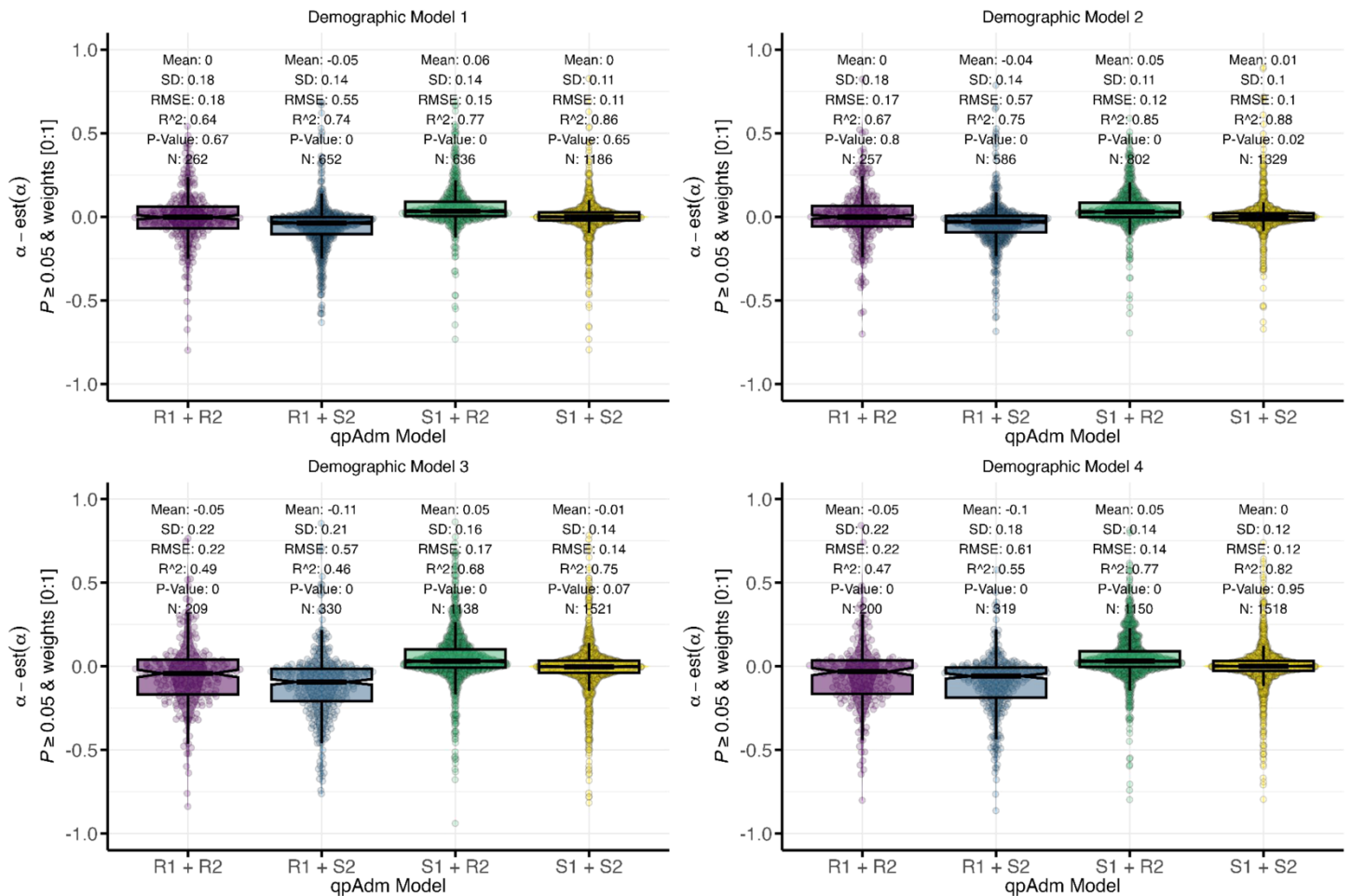
371 converted to probabilities. This conversion was done by first exponentiating the log-odds to get the odds ratio,  
372 and then dividing the odds ratio by one plus the odds ratio (i.e., Probability of QTP-binary = odds-ratio / (1 +  
373 odds-ratio)). We visualized these predicted probabilities on a grid that represents the space of the historical  
374 parameters (Figure 3I-L).

375  
376 As expected, larger median pairwise  $F_{ST}$  values resulted in increased QTP-binary probability for all simple  
377 demographic Models (Figure 3I-L), with the more complex Models 2-4 performing better than Model 1 across  
378 historical  $F_{ST}$  ranges. Counterintuitively, the model with admixture from the internal branch ancestral to both the  
379 S2 and R2 populations (iS2R2) (Model 3) performed the best at  $F_{ST}$  values both below (median pairwise  $F_{ST} <$   
380 0.008) and within (median pairwise  $0.008 < F_{ST} < 0.013$ ) ranges approximating that of historical periods (Figure  
381 3I-L). When median divergence levels reached those of populations older than the Bronze Age (median  
382 pairwise  $F_{ST} > 0.013$ ) the model with a gene flow from an outgroup to a source branch (Model 2) outperformed  
383 the others, achieving the same QTP-binary probability values with fewer generations since admixture than  
384 Models 3 and 4 (Figure 3J). It also had the highest maximum QTP-binary probability of all Models, achieving  
385 this with generations since admixture greater than ~90 (Figure 3J). In the absence of admixture events in the  
386 history of S1 (Model 1), we observed no significant impact of generations since admixture on the QTP-binary  
387 probability (Chi-sq = 2.25 and  $P$ -value = 0.089). However, all three admixed-source Models, especially Model 3,  
388 show a weak but statistically significant effect of generations since admixture on QTP-binary, with the effect  
389 appearing more pronounced for larger  $F_{ST}$  values (approximate significance of  $T_{admix}$  predictor variable smooth  
390 term: Model 2 Chi sq = 9.94 and  $P$ -value = 0.0012; Model 3 Chi sq = 23.79 and  $P$ -value < 0.001; and Model 4  
391 Chi sq = 10.17 and  $P$ -value = < 0.001.) (Figure 4I-L). The observed weak influence of generations post-  
392 admixture on the QTP-binary probability is likely a consequence of correlations between the  $T_{admix}$  and  $T_3/T_4$   
393 parameters (SI Figure S1B) rather than a decline in performance due to more recent admixture. In support of  
394 this idea is that both Models 3 and 4, which incorporate admixture from the iS2R2 branch that is delineated by  
395 the  $T_2$  and  $T_4$  split times (Figure 3C-D), have the strongest correlation between  $T_{admix}$  and  $T_4$  of all demographic  
396 Models (SI Figure S1B). Conversely, under Model 2,  $T_{admix}$  has the strongest correlation with parameter  $T_3$  (SI  
397 Figure S1B), which determines the divergence time between R1 and S1.



409

**Figure 5**



410

411

412

413

414

415

416

Deviations of estimated from simulated admixture proportions for the R1 and S1 sources in the *qpAdm* models S1+S2, R1+R2, R1+S2, and S1+R2. Median pairwise  $F_{ST}$  between the S1, S2, R1, R2, and R3 populations is between 0 and 0.02 and the number of generations since admixture is less than or equal to 100. Each panel shows results for one simple demographic model.

417

### Accuracy of Admixture Weight Estimates

418

419

420

421

422

423

424

425

426

427

428

We also evaluated if the introduction of admixture to the ancestral source population (S1) would introduce bias or increase uncertainty in the admixture weight estimation of the Target for the true *qpAdm* model. Consistent with previous studies (Harney *et al.* 2021), in the absence of ancestral admixture, we observed a delta alpha (simulated minus estimated admixture weight) mean of zero under Model 1 and an  $R^2$  of 0.86 demonstrating *qpAdm* can accurately estimate the simulated admixture weight without bias (one-sample T-test  $P$ -value = 0.65) (Figure 5). However, in the presence of admixture to the source population from an outgroup, we observe a subtle and weakly significant overestimation of the S1 contribution to the Target (Model 2, delta-alpha mean = 0.01, one-sample T-test  $P$ -value = 0.02), and an underestimation of almost equal magnitude, but not significant, when admixture in S1 is from the iS2R2 branch (Model 3, delta-alpha mean = -0.01, one sample T-test  $P$ -value = 0.07) (Figure 5). The symmetrical biases between Model 2 and Model 3 appear to cancel out under demographic Model 4, where we observe a delta-alpha mean of zero (one-sample T-test  $P$ -value = 0.95)

(Figure 5). All demographic Models exhibited similar levels of uncertainty in their admixture weight estimation (Figure 5). Whilst Model 3 performed the worst with the lowest  $R^2$ , largest delta-alpha standard deviation, and root-mean-squared error (Figure 5), the weight-estimate uncertainty is considerably smaller than expected under completely random sampling (the SD of the difference between two uniformly distributed and uncorrelated random variables is 0.408) further supporting the accuracy of qpAdm admixture estimates under these conditions.

In empirical aDNA studies, one will often include multiple closely related populations in the qpAdm candidate source list to determine which is the best representation of the Target ancestry. Therefore, we assessed how the selection of false sources and their phylogenetic relationship to the true source affected the bias and uncertainty in admixture weight estimates. Under the simplest model (Model 1), we observed that misspecified (false) models that combine the true Source populations with the sister clade of the other true Source (R1+S2 or S1+R2) resulted in an almost equal overestimation of the simulated admixture weight for the true Source (R1+S2 mean = -0.05, T-test  $P$ -value = < 0.001 whereas S1+R2 mean = 0.06, T-test  $P$ -value < 0.001) (Figure 5). However, when both sources are equally phylogenetically distant from the true admixing sources (R1+R2), we only observed an increase in the weight estimate uncertainty but no bias (SD = 0.18 and T-test  $P$ -value = 0.67) (Figure 5). We observed similar qualitative patterns in the admixed-source models (Models 2-4), with a bias in overestimating the contribution from the true source when paired with one of the false sources (S1+R2 and S2+R1 T-test  $P$ -values < 0.001) (Figure 5). Interestingly, the largest effects are observed under demographic Models 3 and 4 for the qpAdm model S2+R1, suggesting that admixture from the internal iS2R2 branch to the ancestral S1 branch increases the overestimation of the S2 contribution to the Target (Figure 5). Moreover, selecting the two symmetrical populations R1+R2 results in an overestimation of the R2 contribution under both Models 3 and 4 (T-test  $P$ -value < 0.001), but we observed no bias under Model 1 (T-test  $P$ -value = 0.67). Additionally, the Model with a gene flow from an outgroup to the ancestral S1 branch (Model 2) has no bias in admixture weight estimation (R1+R2 T-test  $P$ -value = 0.8), further supporting the impact of the admixture between ancestral source branches (iS2R2) on the qpAdm weight bias (Figure 5).

#### qpAdm Plausibility Criteria and Improving Model Inference Accuracy

A number of different qpAdm plausibility criteria are employed in empirical aDNA analysis such as  $P$ -value thresholds of 0.01 (e.g., Skoglund et al. 2017, Narasimhan et al. 2019, Lazaridis et al. 2022, Bergström et al. 2022, Koptekin et al. 2023, Skourtanioti et al. 2023) and 0.05 e.g., (Olalde et al. 2019; Sirak et al. 2021; Salazar et al. 2023), the use of two-standard error constraint (S.E.) on the admixture weights (Narasimhan et al. 2019), the requirement of the rejection of all single-source models, and favoring simpler models over more complex ones (Lazaridis et al. 2016, 2022a; Skoglund et al. 2017; Narasimhan et al. 2019; Salazar et al. 2023). Our objective was to determine how these plausibility criteria impact the performance (QTP and QTP binary) and accuracy (FPR and FDR) of qpAdm admixture model inference. Additionally, we aimed to assess whether the

465 accuracy (FPR and FDR) of qpAdm could be improved by conditioning on a significant admixture  $f_3$ -statistic for  
466 plausible two-source models, a method used to test if a target population is consistent with being formed from  
467 two putative sources (Patterson *et al.* 2012; Peter 2016, 2022).

468  
469 We observed a substantial decrease in the average error rates (FPR and FDR), and an increase in average  
470 performance metrics (QTP, and QTP-binary) across all demographic Models when introducing the admixture  
471 weight [0:1] constraint to the plausibility criteria in qpAdm (Table 1). We note that the admixture weight [0:1]  
472 constraint is the most common additional constraint on qpAdm model plausibility in the archaeogenetic  
473 literature. However, adding the additional  $\pm 2$  S.E. weight constraint, while reducing the FPR for all demographic  
474 Models, also increased the FDR for both  $P$ -value thresholds (Table 1), highlighting the trade-off between  
475 rejecting the true model and failing to reject false models when assessing accuracy. Similarly, the  $\pm 2$  S.E.  
476 weight constraint also decreased the average QTP results for both  $P$ -value thresholds across all Models and  
477 only Model 2 shows an increase in average QTP-binary (increases in QTP-binary for both  $P$ -values 0.01 and  
478 0.05) (Table 1).

479  
480 We evaluated the impact of requiring all single-source qpAdm models to be rejected on the FP and FDR error  
481 rates. We computed the FPR as follows: For each simulation iteration, if at least one false single-source qpAdm  
482 model was plausible, all two-source qpAdm models were rejected and we then computed the FPR as the FP  
483 / (FP + TN) following the guide above. Meaning, a two-source qpAdm model can only contribute to the FPR if all  
484 single-source qpAdm models are rejected in its simulation iteration. Recalling the above example, in the  
485 demographic Model 1 simulation iteration 1,998, we had an FPR of 0.8 that occurred because, of the 21 total  
486 single and two-source qpAdm models, we have 16 FP models, six of which are single-source models. However,  
487 because we conditioned on all single-source models to be rejected, we have six false positives (single-source  
488 models) and 14 true negatives (rejected two-source), giving this particular simulation iteration an FPR of 0.3.  
489 The FDR was computed following the same procedure, where all two-source qpAdm models are rejected if a  
490 single-source model is plausible in their simulation iteration. Importantly, we found that requiring all single-  
491 source models to be rejected increased the FDR for all demographic Models at both  $P$ -value thresholds (Table  
492 1). Conversely, we find that the FPR is decreased with the rejection of all single-source models for all  
493 demographic Models across all plausibility criteria (Table 1).

494  
495 In addition to rejecting all single-source qpAdm models, the further criterion of a significant admixture  $f_3$ -statistic  
496 for plausible two-source qpAdm models resulted in the lowest error rates (FPR and FDR) for all demographic  
497 Models (Table 1). The relationship between the power of the admixture  $f_3$ -statistic and demographic parameters  
498 was explored by (Peter 2016). They showed through mathematical formulae (see equation 1 below) and simple  
499 simulations similar to our Model 1, that the conditions of a negative  $f_3$ -statistic required a large number of  
500 generations between the split of the admixing sources ( $T_2$ ) and the time of admixture ( $T_{\text{admix}}$ ), a low probability of

501 lineages in the Target population coalescing before the admixture event ( $T_{\text{admix}}$ ), and the admixture proportion  
502 ( $\alpha$ ) close to 50%. As such, for any pair of true-source populations to produce a negative  $f_3$ -statistic for a target,  
503 the demographic model from which they descend must conform to the equation (1) (EQ:1) below.

$$\text{negative } f_3\text{-statistic condition} = \left( \frac{1}{(1-c_x)} \frac{T_{\text{admix}}}{T_2} < 2\alpha(1 - \alpha) \right) \quad (1)$$

504  
505 where  $c_x$  corresponds to the probability two lineages sampled in the Target population have a common ancestor  
506 before the time of admixture (Peter 2016).

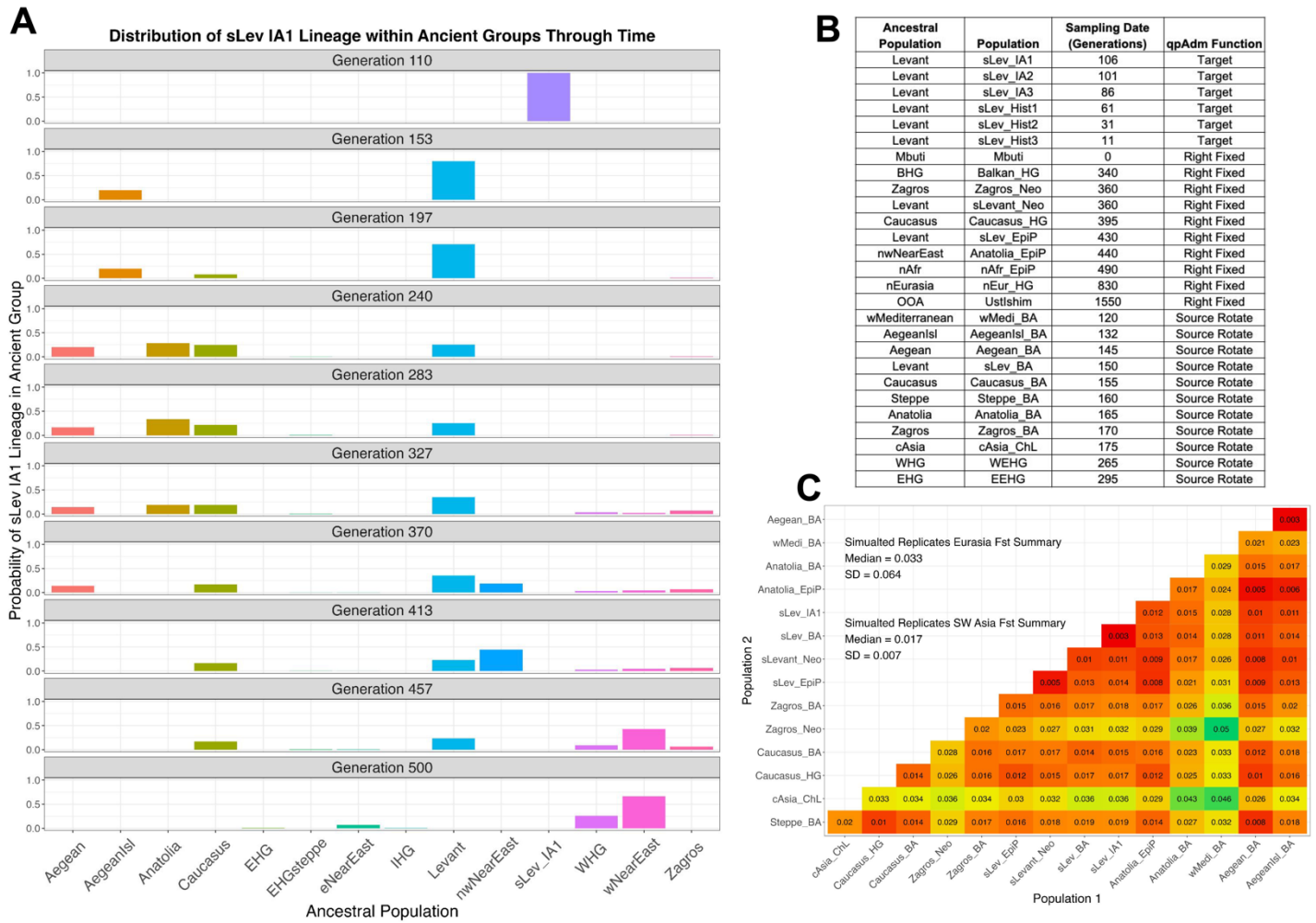
507  
508 By conditioning on simulations whose demographic parameters result in a negative  $f_3$ -statistic condition (i.e the  
509 value of EQ:1 left hand side (LHS) must be less than the value of the right-hand side (RHS)), we show that  
510 demographic Models with admixture to the source (S1) population from the iS2R2 branch (Models 3 and 4)  
511 result in the largest type II error rate (percent of simulations with  $f_3(\text{Target}; \text{S1}, \text{S2})$  Z-score  $> -3$  for Models 1, 2,  
512 3, and 4 are 34%, 30%, 44%, and 43%, respectively) (SI Figure S5A-D). As such, we show that the Models  
513 with a gene flow from the iS2R2 branch require, on average, a larger LHS to RHS difference (smaller ratio of  
514 LHS / RHS) for the negative  $f_3$ -statistic condition to generate significance (median EQ:1 LHS / RHS ratio for  
515  $f_3(\text{Target}; \text{S1}, \text{S2})$  Z-scores  $< -3$  across Models 1, 2, 3 and 4 are 0.158, 0.128, 0.085, and 0.078, respectively).  
516 Given a substantial period of independent drift between the ancestral split of the Sources and the time of  
517 admixture is a prominent factor in  $f_3$ -statistic negativity, this power reduction appears to be principally driven by  
518 the increase in the differences between  $T_{\text{admix}} / T_2$  caused by admixture between the ancestral source lineages.  
519 Importantly, all the demographic effects on  $f_3$ -statistics power described above are magnified when selecting the  
520 wrong source pairs (SI Figure S5B-D).

## 522 qpAdm model ranking by $P$ -value

523 A common application of qpAdm, and by extension qpWave, is ranking model performance via  $P$ -values (van de  
524 Loosdrecht *et al.* 2018; Oliveira *et al.* 2022; Lazaridis *et al.* 2022a; Taylor *et al.* 2023; Moots *et al.* 2023). We  
525 evaluated the use of  $P$ -values for the relative ranking of qpAdm models by assessing how frequently each of the  
526 single and two-source qpAdm models had the largest, second-largest, third, and fourth-largest  $P$ -values for  
527 each of the 5k simulations. Across all demographic Models, the true qpAdm model significantly outperformed all  
528 other qpAdm models by having the largest  $P$ -value in more than 60% of the simulations (SI Table ST1A-D).  
529 Additionally, we found that both the relative ranking and frequency of  $P$ -values reflected the underlying  
530 demography and frequency of plausible models described above (Figure 4). Under Model 1, the S1+R2 and  
531 S2+R1 models had the largest  $P$ -value with about equal frequency (0.127 and 0.137, respectively), whereas,  
532 under demographic Models 2-4, the S1+R2 qpAdm model has the largest  $P$ -value approximately 10x more  
533 frequently and is the second best performing of all qpAdm models (SI Table ST1A-D).

536

**Figure 6**



537

538

539

540

541

542

543

544

(A) Barplots showing probabilities of encountering a lineage found in the “sLev IA1” group in other simulated ancestral populations (only presenting populations with non-negative probabilities). The ancestral populations are those from which we sampled and correspond to the first column in B. (B) A table of the sampled populations used in qpAdm analysis and the and the ancestral populations they split from (corresponding to ancestral populations in A). An  $F_{ST}$  matrix (C) for the sampled simulated populations is also shown.

545

**Going complex: Admixture Inference under Complex Human History in aDNA Research**

546

547

548

549

550

551

552

553

We expanded our evaluation of admixture inference from simple topologies to a demographic model and data distribution that reflects the real-world complexities of both Eurasian human history and aDNA conditions. We framed this by simulating an archaeogenetic hypothesis on the origin of migrants to the southern Levant at the beginning of the Iron Age (the so-called Sea Peoples migration). While our demographic model and parameters are informed by the aDNA and population genetic literature, we stress that it is not designed to represent true human history, nor a proposal of the likely events associated with the Sea Peoples migration. Rather, its function is solely to capture some of the complexities surrounding the dynamics connecting populations in the historical period such as low divergence between candidate source populations, complexity of ancestral



554 population relationships, and sampling recently after admixture event. As such, it provides us a framework from  
555 which we can evaluate the behavior and limitations of admixture inference from aDNA.

556 In total, we model 59 populations and 41 pulse admixture events ([Figure 6A](#)) which are all described and  
557 referenced in [Supplementary File SF1](#). A brief summary of the model scaffold follows. The oldest split in the  
558 demography is the separation of East and Central African ancestral populations at 5,172 generations before  
559 present (Hollfelder *et al.* 2021). We model an out-of-Africa (OOA) population as separating from the East  
560 African lineage at 3,303 generations (Kamm *et al.* 2020; Marchi *et al.* 2022), and from the former lineage split  
561 East Eurasian, North Eurasian, West European Hunter-Gatherer (WEHG), and ancestral Near Eastern lineages  
562 (Kamm *et al.* 2020; Marchi *et al.* 2022). Two meta Near Eastern lineages, Eastern and Western Near East, split  
563 from the ancestral Near Eastern lineage (Marchi *et al.* 2022). The Levant lineage, from which the target  
564 southern Levant IA population (“sLev\_IA1”) largely descends, splits from the “Western Near East” lineage at  
565 483 generations (Lazaridis *et al.* 2016; Broushaki *et al.* 2016; Marchi *et al.* 2022). We model the formation of a  
566 “Northwestern Near East” lineage at 446 generations, from which the admixing source population “Aegean  
567 Island” (“AegeanIsl\_BA”) largely descends, as a mix of “Western Near East” (0.86) and WEHG (0.14) (Marchi  
568 *et al.* 2022). The Target lineage, sLev\_IA1, was modeled as a mixture of its ancestral population (southern Levant  
569 Bronze Age, “sLev\_BA”) and the AegeanIsl\_BA population (admixture fraction = 0.2) at generation 111 before  
570 present. We sampled the Target population five generations post-admixture ([Supplementary File SF1](#)). To  
571 assess the influence of post-admixture drift on admixture inference, we modeled successive step-wise splits  
572 from the Target lineage and sampled them 10, 25, 50, 80, and 100 generations post the original admixture  
573 event. From our simulated lineages, we sampled data representing the Mbuti present-day population, and 20  
574 ancient Eurasian and African populations that reflect empirical ancient groups present in many Southwest Asian  
575 aDNA analyses ([Figure 6B](#)). For all populations, we sampled 10 individuals and in all downstream analyses we  
576 defined the pairing of sLev\_BA+AegeanIsl\_BA as the true model and all others as false models. As above, we  
577 consider plausible models to have a  $P$ -value  $\geq 0.05$  and admixture weights between zero and one ([0:1]).

## 578 Data generation

579 We configured the Eurasian demographic Model ([Supplementary File SF1](#)) using the Demes graph format  
580 (Gower *et al.* 2022) and converted it to an msprime demography object through the `demography.from_demes()`  
581 function. We simulated 50 whole-genome ( $L \sim 2875$  Mbp) replicates using sequence lengths and recombination  
582 rates of chromosomes 1-22 following the HomSap ID from the `stdpopsim` library (Adrion *et al.* 2020) and  
583 separated each chromosome with a  $\log(2)$  recombination rate following msprime manual guidelines (Nelson *et al.*  
584 *et al.* 2020, Baumdicker *et al.* 2022 ). The first 25 generations into the past were simulated under the Discrete  
585 Time Wright-Fisher (DTWF) model (Nelson *et al.* 2020), and under the Standard (Hudson) coalescent model  
586 until the sequence MRCA. We applied mutations to the simulated tree sequence at a rate of  $1.29e-08$  (Jónsson  
587 *et al.* 2017) using the Jukes-Cantor mutation model (Jukes and Cantor 1969). From the mutated tree sequence,  
588 we generated Eigenstrat files through the `tskit v.0.5.2 TreeSequence.variants()` function which were passed to

589 custom R scripts to generate realistic aDNA conditions such as filtering on bi-allelic sites, adding ascertainment  
590 bias, downsampling to 1,233,013 SNPs (1240k capture), and for the simulated ancient individuals, generating  
591 pseudo-haploid data with high missing rate ([Github Repo:https://github.com/archgen/complex\\_demog\\_sims.git](https://github.com/archgen/complex_demog_sims.git)).  
592

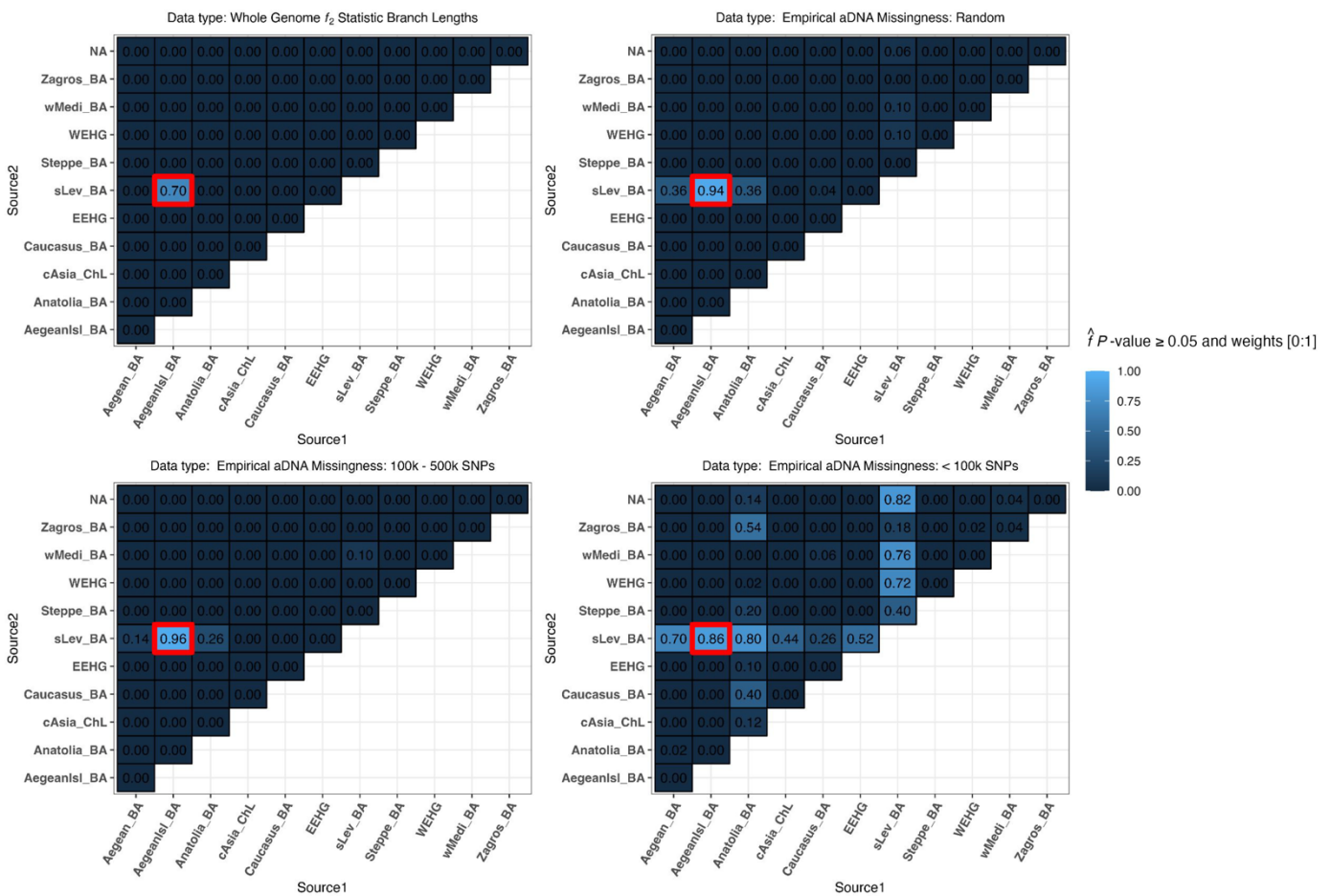
593 We configured the SNP ascertainment bias scheme replicating the general principles of the Human Origins  
594 array (Patterson *et al.* 2012). See also (Flegontov *et al.* 2023) for an overview of effects of this type of  
595 ascertainment on  $f$ -statistics and related methods. In the Eurasian demography, we defined separate lineages  
596 representing central European (CEU), East Asian (CHB), African (AFR), and South Asian (sAs) populations,  
597 sampled a single individual from these lineages at the present, and retained biallelic sites that are heterozygous  
598 in at least one of these individuals. We then downsampled the simulated data by randomly sampling 1,233,013  
599 SNP loci. For the simulated ancient samples ([Figure 6B](#)), we randomly assigned one of the two alleles as  
600 homozygous at simulated heterozygous positions mimicking what is commonly performed for low- and medium-  
601 coverage aDNA (Schuenemann *et al.* 2017). In addition, we added missing data by assigning to each ancient  
602 individual an empirical missingness distribution from a randomly selected ancient individual within the AADR  
603 v.52.2 (Mallick *et al.* 2023), which we filtered by removing related and contaminated ancient individuals and  
604 restricted to individuals from Southwest Asia (see [Supplementary File SF2 for the list of aDNA individuals](#)).  
605 Amongst the Target populations, this resulted in a range of missingness within each replication (the median  
606 standard deviations of the population missingness across the replicates ranged from 0.04 to 0.14), with the  
607 average proportion of missingness across the 50 replicates ranging from a minimum of 49% for the sLev\_IA3  
608 population to 89% in the sLev\_IA1 population (resulting in a range of approximately 130,656 to 627,948 useful  
609 SNPs, respectively) ([SI Table ST2](#)). We observe similar degrees of missingness for the other ancient  
610 populations included in the qpAdm rotation analysis ([SI Table ST2](#)). We generated two additional missing data  
611 subsets following the method above, whereby the AADR individuals were filtered to contain only low (SNPs <  
612 100k), or medium coverage (100k < SNPs < 500k) from samples across Eurasia, resulting in 50 whole-genome  
613 replicate simulations with three different degrees of missingness.  
614

615 From the simulated aDNA we computed rotating qpAdm analyses with the ADMIXTOOLS2 software (Maier *et al.*  
616 *et al.* 2023) using parameters typical of empirical aDNA workflows such as “allsnps = TRUE” (using all SNPs  
617 available for calculating each individual  $f_4$ -statistic), and 5 Mbp windows for calculating standard errors of  $f_4$ -  
618 statistics with the jackknife procedure. We configured the qpAdm analysis protocol in the following way: the  
619 most ancient groups are fixed in the right-group position and younger candidate populations are rotated  
620 between the left and right-group positions (Narasimhan *et al.* 2019; Lazaridis *et al.* 2022a). The Mbuti  
621 population was fixed in the first position in all qpAdm analyses, along with nine deeply divergent Eurasian and  
622 African populations fixed in the right group. We then rotated nine simulated Bronze Age and Chalcolithic, and  
623 two European hunter-gatherer populations ([Figure 6B](#)) between the left and right-group positions resulting in a  
624 total of 11 single-source models, and 66 two-source models.  
625

To evaluate the impact of aDNA conditions on admixture inference under the Eurasian human demography, we generated  $f_2$  - and  $F_{ST}$  - statistic matrices directly from the tree sequence without mutations through tskit v.0.5.2 with parameters "Mode=branch", and "span\_normalise=True", using 5 Mbp windows. The resulting  $f_2$  - statistics matrix was used to compute rotating qpAdm analyses using the same sample set as input for the aDNA application, and admixture  $f_3$ -statistic in the ADMIXTOOLS2 software (Maier *et al.* 2023).

Our simulations resulted in expected levels of population divergence given empirical observations with a median pairwise  $F_{ST}$  of 0.03 between all Eurasian populations and 0.017 amongst the Southwest Asian populations (Figure 6C). A pairwise  $F_{ST}$  matrix computed on the first replicate shows expected genetic affinities amongst the analysis populations (Figure 6C). We used the tskit lineage\_probabilities() function to further assess the relationship between the Target and analysis populations by tracking the location of lineages sampled from the Target through time amongst the remaining simulated demographic lineages (Figure 6A). The results show that between the youngest and oldest populations included in the qpAdm analysis, lineages from the Target population are principally found in the Levant, Aegean, AegeanIsl, Anatolia, and Caucasus ancestral groups (Figure 6A).

Figure 7



645 Heatmaps of the proportion of replicates with plausible  $P$ -value  $\geq 0.05$  and weights [0:1] for the complex demography  
646 (Aegean Island admixture to southern Levant) for the target population “sLev IA1”. The red box represents the most optimal  
647 true model. Results are presented for four datasets:  $f_2$ -statistics calculated on whole-genome branch lengths and the three  
648 datasets with varying SNP missing rates.

649  
650

#### 651 qpAdm performance and aDNA data quality

652 Consistent with the results previously shown by (Harney *et al.* 2021), the degree of data missingness appears to  
653 be one of the primary factors influencing the performance of rotating qpAdm. Below, we adopt the term  
654 “coverage” to represent the proportion of SNPs non-missing. Thus, the aDNA missingness sampling condition  
655 of SNPs  $< 100k$  represents the lowest-coverage dataset, the random missingness sampling condition  
656 represents the medium-coverage dataset, and the missingness condition of  $100k < \text{SNPs} < 500k$  represents the  
657 highest-coverage dataset. The lowest-coverage dataset produced the largest frequency of plausible single-  
658 source and two-source qpAdm models (Figure 7), resulting in the lowest average QTP, largest FPR and FDR,  
659 and an average QTP-binary of zero (SI Table ST3). The two higher-coverage aDNA sampled datasets resulted  
660 in very similar qpAdm performance with the highest-coverage dataset performing slightly better than the middle  
661 coverage dataset as it both rejects all single-source qpAdm models and has less total plausible qpAdm models  
662 (Figure 7), resulting in on average higher QTP and QTP-binary and the lowest FPR and FDR (SI Table ST3).  
663 Since the degree of missingness in the AADR random sampling scheme sometimes results in populations with  
664 more missingness than the lowest-coverage dataset (SI Table ST2), the relative performance of these missing  
665 data schemes highlights the importance of maximizing data coverage in all populations, not just the Target, for  
666 rejecting false qpAdm models.

667

668 Interestingly, we note that amongst the two higher-coverage datasets the plausible false qpAdm models are not  
669 arbitrarily selected as they descend from ancestral populations that are shown above to harbor the Target  
670 population lineages (Figure 6A). The single exception to this pattern is the simulated wMedi\_BA population  
671 which, when only paired with the sLev\_BA population, is plausible at 10% in each of the higher coverage  
672 datasets, and 76% in the lowest-coverage dataset (Figure 7). The simulated wMediterranean lineage, from  
673 which wMedi\_BA descends, is modeled as receiving 6% admixture from the true source population 29  
674 generations before the formation of the Target population (Supplementary File SF1). This demonstrates that  
675 correlations in allele frequencies between populations driven by admixture from a shared ancestral source, in  
676 addition to shared genetic drift, can result in false positive qpAdm results.

677

678 Interestingly, the qpAdm rotation analysis on the simulated whole-genome branch-length  $f_2$ -statistic rejected all  
679 false qpAdm models and classified the true model plausible in 70% of the replicates and rejected it in 30% of  
680 the replicates (Figure 7). We ran a receiver operating characteristic curve (ROC) analysis where we varied the  
681  $P$ -value between zero and one to assess the relationship between  $P$ -value thresholds and qpAdm performance



682 as measured by the true positive (TP) and false positive (FP) rates. In calculating the ROC, we constrained the  
683 qpAdm models to have plausible admixture weights [0:1] and performed each calculation on 3,000  $P$ -value  
684 thresholds between zero and one. These results revealed for the whole-genome branch-length  $f_2$ -statistic  
685 dataset, the qpAdm TPR converges to 100% with  $P$ -values greater than  $1 \times 10^{-3}$  and the FPR does not increase  
686 until the  $P$ -value reaches zero (SI Figure S6). All datasets with aDNA missingness exhibit a trade-off of co-  
687 varying increases/decreases in the FPR/TPR with changes to the  $P$ -value threshold (SI Figure S6) which we  
688 also observe in the distribution of  $P$ -values for qpAdm models with plausible admixture weights (SI Figure S7).  
689 Importantly, both higher-coverage aDNA datasets have greater than 89% TPR and less than 1% FPR with a  $P$ -  
690 value threshold of 0.1, suggesting an additional strategy for increasing qpAdm accuracy (SI Figure S6).

#### 691 qpAdm model ranking by $P$ -value in ancient DNA under complex demography

692 We evaluated the accuracy of determining the best-fitting qpAdm model by ranking them by their  $P$ -values given  
693 the admixture complexity of the simulated Eurasian demography. Under the higher coverage datasets, the true  
694 model has the largest  $P$ -value in more than 90% of the simulation replicates and has the largest  $P$ -value in  
695 100% of the replicates using the highest-coverage dataset (SI Table ST4). However, caution should be applied  
696 to ranking qpAdm by  $P$ -values in datasets with low coverage as in our lowest aDNA coverage dataset, the true  
697 model has the largest  $P$ -value in only 16% of the replicates, second to the false model of sLev\_BA+Anatolia\_BA  
698 (SI Table ST4). The observation of the sLev\_BA+Anatolia\_BA and sLev\_BA+Aegean\_BA source combinations  
699 as the alternative qpAdm models that possessed the largest qpAdm  $P$ -value in at least one replicate  
700 demonstrates the difficulty in rejecting closely related candidate sources ( $F_{ST}$  between the AegeanIsl\_BA or  
701 Aegean\_BA and Anatolia\_BA populations  $\sim 0.003$  and  $0.017$ , respectively). Nonetheless, the sLev\_BA  
702 population is consistently paired with alternative sources in the most frequent qpAdm models with the largest  $P$ -  
703 values, suggesting that the identification of overrepresented populations in high-ranking qpAdm models is a  
704 suitable heuristic to determine a likely true source regardless of the degree of data missingness (SI Table ST4).

#### 705 Generations since admixture and qpAdm performance and accuracy

706 We also explored the effect of post-admixture drift on qpAdm performance. Importantly, across all descendent  
707 Target populations and degrees of data missingness, we observe no significant trend in qpAdm performance or  
708 admixture weight accuracy (SI Figure S8). Under the whole-genome branch-length  $f_2$ -statistic dataset, the  
709 admixture weight S.E. appears to increase with increasing generations since admixture, however, it has no  
710 significant impact on accuracy or precision of admixture weight estimates (SI Figure S8). As expected, we  
711 observe the largest estimated admixture weight S.E. and delta-alpha values in the lowest-coverage dataset,  
712 with between 0.12 and 0.20 SD on delta-alpha (SI Figure S8). As such, caution should be given to interpreting  
713 admixture proportions from datasets of low coverage.



## 714 qpAdm plausibility criteria

715 We evaluated the impact of the different qpAdm plausibility criteria described in the simple demography section  
716 on our complex demographic aDNA simulations. In contrast to the simple demographic simulations, the  
717 introduction of the 2 S.E. constraint on admixture weight estimates consistently either reduced or maintained the  
718 FPR for all aDNA missingness conditions and either reduced or maintained the FDR in all but the lowest-  
719 coverage datasets (SI Table ST3). Of note is that each aDNA missingness dataset has a different plausibility  
720 criterion that maximizes its QTP, making the selection of single plausibility criteria to maximize QTP infeasible.  
721 We do, however, observe for all datasets, the lowest error rates (FDR and FDR) with the co-criteria of rejection  
722 of all single-source models,  $P$ -value  $\geq 0.05$ , and 2 S.E. constraint on the admixture weight estimates, albeit with  
723 greater than 0.98 FDR in the lowest coverage dataset (SI Table ST3). The plausibility criteria of  $P$ -value  $\geq 0.05$   
724 and admixture weights [0:1] results in the smallest FDR in the lowest-coverage dataset. In empirical studies with  
725 low coverage aDNA samples, this may represent the most optimal plausibility criteria as it minimizes the  
726 frequency of type II errors as evidenced by the largest QTP value (SI Table ST3). The distribution of  $P$ -values  
727 for models with plausible admixture weights [0:1] (SI Figure S7), and the ROC curve analysis (SI Figure S6)  
728 shows that increasing the  $P$ -value threshold for the low-coverage dataset does not result in a significant  
729 reduction in the FP rate without penalizing the TP rate (SI Figure S6).

730  
731 Importantly, we observe no impact from the use of significant admixture  $f_3$ -statistics as an additional plausibility  
732 criterion to increase qpAdm model inference accuracy as all pairwise combinations of qpAdm sources were not  
733 significant regardless of data quality (SI Figure S9). As described above, the power for the detection of  
734 admixture from  $f_3$ -statistics is strongly influenced by the underlying population demography and divergence of  
735 the candidate sources from the true admixing populations. Our simple demography simulations showed that  
736 both gene flow between source lineages, and increased divergence of the candidate source population from the  
737 true admixing source reduced the  $f_3$ -statistics power. Under the Eurasian demography, the two source groups,  
738 Levant and Aegean, undergo recent bi-directional gene flow after the split from their most recent common  
739 ancestral population.

740  
741 We computed the  $f_3$ -statistics negativity condition (EQ:1) for both the split-time of the Levant and Aegean  
742 sources and the date of admixture for all Target populations. As expected, we observe an increase in  $f_3$ -statistic  
743 estimates with increasing generations since admixture (SI Figure S9). Moreover, the estimated  $f_3$ -statistic  
744 negativity appears to conform closer to the  $f_3$ -statistics negativity condition (EQ:1) when computed using the  
745 date of most recent bi-directional admixture between the Levant and Aegean sources than the their split time (SI  
746 Figure S9). This further supports the impact of admixture between source lineages on decreasing the power of  
747 the admixture  $f_3$ -statistic and highlights the importance of utilizing  $f_3$ -statistic estimates as confirming plausible

748 qpAdm models rather than rejecting false models, similar to how they were originally proposed (Patterson *et al.*  
749 2012).

## 751 Data availability

752 The authors affirm that all data necessary for confirming the conclusions of the article are present within the  
753 article, figures, tables, and supplementary materials. Both simple demographic Model and complex Eurasian  
754 model simulations were written in Snakemake pipelines to facilitate reproducibility and can be accessed via our  
755 GitHub repository [https://github.com/archgen/complex\\_demog\\_sims](https://github.com/archgen/complex_demog_sims). Supplemental figures available in  
756 Supplementary Material PDF:

## 758 Discussion

759 The qpAdm software has become one of the hallmark methods in archaeogenetic analyses for reconstructing  
760 admixture histories of ancient populations (see Lazaridis *et al.* 2016, 2022a; b; Skoglund *et al.* 2017; Mathieson  
761 *et al.* 2018; Harney *et al.* 2018; Narasimhan *et al.* 2019; Antonio *et al.* 2019; Marcus *et al.* 2020; Fernandes *et al.*  
762 *et al.* 2020; Wang *et al.* 2020, 2021; Ning *et al.* 2020; Yang *et al.* 2020; Carlhoff *et al.* 2021; Papac *et al.* 2021;  
763 Librado *et al.* 2021; Sirak *et al.* 2021; Patterson *et al.* 2022; Changmai *et al.* 2022a; b; Bergström *et al.* 2022;  
764 Maróti *et al.* 2022; Lee *et al.* 2023). This is due in part to its modest computational requirements, use of allele  
765 frequency data, and minimal model assumptions (Haak *et al.* 2015; Harney *et al.* 2021). The primary motivation  
766 for this work is addressing its applicability, performance, and limits in reconstructing admixture histories under  
767 challenging scenarios that emerge when reconstructing population dynamics within the historical period. Such  
768 conditions range from identifying the true source population amongst minimally differentiated candidates, and  
769 potential biases that may arise from sources that are admixed and ancestrally connected through complex  
770 demographies. It also may involve dealing with short intervals between the admixture event of interest and the  
771 ancient sample. Additionally, we sought to determine how these challenges are impacted by missing data  
772 typical of aDNA conditions. We addressed these questions through simulations of both simple admixture-graph-  
773 like demographies exploring a wide parameter space, and whole-genome simulations of an admixture-graph-  
774 like demography that reflects the inferred complexity of Eurasian population history.

775  
776 It is important to acknowledge that our study configures human demography as a series of discrete population  
777 splits and pulse admixture events, each separated by periods of independent genetic drift (that is why these  
778 simulations are termed “admixture-graph-like”). Thus, if the distribution of human settlements across the ancient  
779 landscape aligns more closely with temporally evolving stepping-stone models, the interpretations drawn from  
780 our study may lose some of their significance. Also of note is that all of our demographic models adhere to the

781 fundamental assumptions of qpAdm (Harney *et al.* 2021): 1) there are no gene flows connecting lineages  
782 private to candidate source populations (after their divergence from the true admixing populations) and "right-  
783 group" populations, and 2) there are no gene flows from the fully formed Target lineage to "right-group"  
784 populations (Harney *et al.* 2021). It is crucial to recognize that these assumptions might be frequently violated  
785 when investigating demographic history in the historical period and beyond it, leading to false rejections of true  
786 simple models. In turn, these prompt researchers to test more complex models which often satisfy qpAdm  
787 model plausibility criteria but are misleading when subjected to historical interpretation (Yüncü *et al.* 2023). If  
788 stringent sampling criteria, as outlined in our companion paper (Yüncü *et al.* 2023), are not diligently followed,  
789 these violations are shown to pose substantial challenges to the effective use of qpAdm in demographic  
790 inference.

791  
792 Our simple demographic simulation results show that qpAdm converges on its maximum QTP as the median  
793 pairwise  $F_{ST}$  of the sample set approaches  $\sim 0.005 - 0.008$  (Figure 3E-H), well within the diversity expected of  
794 historical period populations (Figure 1B). However, we find a much larger level of population divergence is  
795 required, with a median pairwise  $F_{ST}$  exceeding 0.015, to simultaneously reject all false models and identify the  
796 true model with a probability greater than 30% (Figure 3I-L). This finding suggests that highly specific  
797 archaeogenetic hypotheses that require the sole identification of the correct model may currently lie beyond the  
798 capabilities of qpAdm given the prevailing data conditions. Importantly however, within the set of models  
799 considered plausible by qpAdm under both the simple admixture simulations and Eurasian complex simulations,  
800 we consistently observe that one of the true sources is included in those most frequently accepted models,  
801 irrespective of the degree of population divergence or levels of data missingness (Figure 4A-D & Figure 7).

802  
803 When it comes to distinguishing between closely related cladal populations, such as the differentiation between  
804 S1 and R1 or S2 and R2 in our simple admixture simulations, our results suggest that qpAdm exhibits  
805 heightened discriminatory power when these closely related cladal populations have diverged on the order of  
806  $F_{ST} > \sim 0.002 - 0.004$  (SI Figure S4). A similar result emerges from our complex Eurasian demographic  
807 simulations. For instance, the candidate source Aegean\_BA, which is modeled as having recently split from the  
808 true source AegeanIsl\_BA, is differentiated at a median  $F_{ST}$  of 0.003 and is frequently included in plausible  
809 qpAdm models at all levels of data missingness (Figure 7). However, it's worth noting that in the complex  
810 Eurasian demographic simulations, population divergence alone does not exclusively determine the probability  
811 of a false source appearing in a plausible qpAdm model. For instance, we frequently observe the Anatolian\_BA  
812 population in plausible qpAdm models (Figure 7) whilst it is both approximately equally divergent from the true  
813 sources (median  $F_{ST}$  Aegean\_BA = 0.016 and sLev\_BA = 0.015) (Figure 6C). This is likely driven by  
814 demographic factors analogous to the conditions of simple demographic Models C-D (Figure 3 C-D) whereby  
815 the Eurasian demographic model includes bi-directional gene flow between the ancestral Levant and Anatolian  
816 populations (30% Levant to Anatolia, and 40% Anatolia to the Levant in generations 305 and 224, respectively)

817 resulting in a substantial likelihood that lineages from the Target population are present within the Anatolian  
818 population ([Figure 6A](#)).

819  
820 A key discovery with relevance for archaeogenetic research in regions with complex migration histories is that  
821 introducing admixture into the source population (but not violating the topological assumptions described above)  
822 can notably improve qpAdm's performance, especially when considering conditions that resemble the typical  
823 levels of divergence observed during the historical period. Notably, the phylogenetic origin of the ancestral  
824 admixture differently impacts qpAdm accuracy (FPR and FDR) and performance (QTP). For instance, when the  
825 gene flow originates from an outgroup to the Target, true Sources, and candidate source populations, it yields  
826 the highest average QTP performance and lowest FDR among all demographic models. In contrast, we observe  
827 lower FP rates under demographic Models that include admixture between sources (Model 3) than from an  
828 outgroup (Model 2), and the lowest FP rate when both ancestral source admixture events occur (Model 4)  
829 ([Table 1](#)). Overall, this trend appears to be primarily driven by the increased differential relatedness between left  
830 and right-set qpAdm populations, irrespective of the decrease in average population divergence.

831  
832 This observation is consistent with theoretical expectations regarding the way qpAdm uses  $P$ -values to reject  
833 candidate models (Haak *et al.* 2015; Harney *et al.* 2021). When the Target and right-group populations share  
834 genetic drift distinct from the shared ancestry between the Target and the putative left-group sources, this will  
835 result in the rejection of the left-group sources as an admixture model of the Target population given a certain  
836  $P$ -value threshold. As such, the ancestry inherited by the Target from a source that is itself admixed increases  
837 the number of populations that it uniquely shares drift with. This is evident in the increased power to reject the  
838 false R1+S2 qpAdm models with the introduction of admixture to the S1 ancestral source lineage ([Figure 4A-D](#)).  
839 Consequently, these observations underscore the importance in empirical aDNA studies of pre-screening  
840 qpAdm right-groups to optimize genetic differentiation and differential relatedness with potential source  
841 populations for maximizing qpAdm performance, as originally proposed in (Haak *et al.* 2015).

842  
843 We also note that as long as the correct source populations are chosen, usage of source populations with  
844 complex admixture history does not introduce bias in the estimation of admixture weights ([Figure 5](#)). However,  
845 we do observe a reduction in accuracy when admixture to a source lineage occurs from another source branch,  
846 in contrast to an outgroup ([Figure 5](#)). Most notably, the selection of source populations appears to be more  
847 critical for accurately estimating admixture contributions. We observed a significant bias towards the population  
848 closest to the true source, leading to an overestimation of admixture proportions for this population ([Figure 5](#)).  
849 This phenomenon is present in all simple demographic models but appears to be more pronounced in models  
850 with ancestral admixture to the source ([Figure 5](#)). In cases where both populations are equidistant from the true  
851 admixing sources, the bias is only evident in models that include an admixture event between source branches  
852 ([Figure 5](#)).



854 We have observed that two additional criteria significantly enhance the accuracy of qpAdm model inference.  
855 The first involves considering two-source (admixture) qpAdm models only when all single-source models are  
856 rejected. The second criterion involves deeming these models as plausible when the source pairs generate a  
857 significantly negative admixture  $f_3$ -statistic (Z-score < -3). While these criteria have proven effective in reducing  
858 bias (FPR and FDR) across a wide range of demographic parameters, in empirical studies it is crucial to assess  
859 the anticipated parameters of each demographic model being evaluated before applying these criteria  
860 universally. This is because certain demographic conditions can increase the FDR. For example, we observe an  
861 increase of plausible single-source models when the admixture weight is close to 1 (SI Figure S10), which, if  
862 requiring all single-source models to be rejected, will result in the more frequent false rejection of the true  
863 model.

864  
865 As for the criterion of a significant admixture  $f_3$ -statistic, under conditions where there is only a short period  
866 between the split of the admixing source populations and their admixture to form the Target, and when the  
867 admixture proportions deviate significantly from 0.5, the power of the  $f_3$ -statistic to detect an admixture event  
868 diminishes, increasing type II errors (SI Figure S5). We observe this scenario in our Eurasian demographic  
869 simulations where admixture between the ancestral sources after their split decreased the power of the  
870 admixture  $f_3$ -statistic, resulting in a 100% type II error rate (SI Figure S9). Moreover, the simple simulations  
871 reveal this effect is exacerbated by the divergence between the tested candidate population and the true  
872 admixing source, making the conditions for negativity of the  $f_3$ -statistic even more stringent (SI Figure S5).  
873 Therefore, we suggest that the  $f_3$ -statistic should be used as a confirmation and ranking tool for plausible  
874 qpAdm models, rather than as a criterion for rejecting them (i.e. favoritism is given to plausible models with a  
875 significant  $f_3$ -statistic over those without). This aligns with the original interpretative guidance when using the  $f_3$ -  
876 statistic as a formal admixture test (Patterson *et al.* 2012).

877  
878 With respect to the procedure of ranking feasible qpAdm models based on their  $P$ -values, the initial suggestions  
879 from Harney *et al.* in 2021 raised a notable concern and advised against  $P$ -value ranking to identify the best  
880 model. Their argument is based on the observation that  $P$ -values under false models closely related to the true  
881 model are almost uniformly distributed, and that in cases when multiple models are plausible any one false  
882 model could easily have a larger  $P$ -value than the true model (Harney *et al.* 2021). However, our findings from  
883 simple admixture demographic simulations show that in over 60% of the simulations, the true model has the  
884 largest  $P$ -value (SI Table ST1). Similarly, in the complex Eurasian simulations under the condition of high  
885 coverage, we found that in over 90% of the replications, the true model had the largest  $P$ -value (SI Table ST4).  
886 In both simple admixture and complex Eurasian simulations, we also found that the relative ranking of  $P$ -values  
887 accurately reflected how closely a false model represented the true ancestry of the Target population.  
888 Therefore, provided an empirical dataset has good coverage, we propose that ranking qpAdm models by  $P$ -  
889 values can offer valuable additional information for determining the true model.

891 A surprising finding from the complex Eurasian demographic simulations was when performing the qpAdm  
892 rotation analysis with  $f_2$ -statistics computed from the whole-genome branch lengths we obtained a 100% true  
893 positive rate and a 0% false positive rate with a  $P$ -value threshold of 0.001 (SI Figure S6). This outcome  
894 underscores the remarkable potential inherent in the principles underlying qpAdm, while also highlighting the  
895 constraints imposed by data scale. In light of this, we suggest there is room for significant enhancements in the  
896 qpAdm protocol from methods that can leverage more accurate estimations of  $f_2$ -statistics. Possible avenues for  
897 this improvement could be developing innovative techniques for extracting information from ancestral  
898 recombination graphs (ARG) within the context of aDNA, or conditioning on the site frequency spectrum (SFS)  
899 for the enrichment of rare alleles. As such, it is clear from these results that further improvements would make  
900 qpAdm a powerful tool for accurately reconstructing the genetic histories of populations under the most complex  
901 scenarios.

## 903 Acknowledgments

904 We are grateful for the advice and helpful suggestions from Yassine Souilmi, Raymond Tober, and Bastien  
905 Llamas at an initial stage of the project. All simulations were performed on the computational infrastructure at  
906 the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer. We  
907 note that our content is solely the responsibility of the authors and does not necessarily represent the views of  
908 the Institute for Computational and Data Sciences.

## 910 Funding

911 C.D.H. and M.P.W were funded by the National Institute of Health under award number R35GM146886. P.F.  
912 was supported by the Czech Science Foundation (project no. 21-27624S led by P.F.), the Czech Ministry of  
913 Education, Youth and Sports (program ERC CZ, project no. LL2103), the John Templeton Foundation (grant no.  
914 61220 to David Reich), and by a gift from Jean-Francois Clin.

## 916 Conflicts of interest

917 None declared.

921

## 922 Literature cited

- 923 Adrion J. R., C. B. Cole, N. Dukler, J. G. Galloway, A. L. Gladstein, *et al.*, 2020 A community-maintained  
924 standard library of population genetic models.
- 925 Agranat-Tamir L., S. Waldman, M. A. S. Martin, D. Gokhman, N. Mishol, *et al.*, 2020 The Genomic History of the  
926 Bronze Age Southern Levant. *Cell* 181: 1146–1157.e11.
- 927 Alexander D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated  
928 individuals. *Genome Res.* 19: 1655–1664.
- 929 Antonio M. L., Z. Gao, H. M. Moots, M. Lucci, F. Candilio, *et al.*, 2019 Ancient Rome: A genetic crossroads of  
930 Europe and the Mediterranean. *Science* 366: 708–714.
- 931 Arning N., and D. J. Wilson, 2020 The past, present and future of ancient bacterial DNA. *Microb Genom* 6.
- 932 Ávila-Arcos M. C., M. Raghavan, and C. Schlebusch, 2023 Going local with ancient DNA: A review of human  
933 histories from regional perspectives. *Science* 382: 53–58.
- 934 Barros Damgaard P. de, R. Martiniano, J. Kamm, J. V. Moreno-Mayar, G. Kroonen, *et al.*, 2018 The first horse  
935 herders and the impact of early Bronze Age steppe expansions into Asia. *Science* 360.
- 936 Bartash V., 2020 *The Early Dynastic Near East*, in Oxford University Press.
- 937 Baumdicker F., G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, *et al.*, 2021 Efficient ancestry and  
938 mutation simulation with msprime 1.0. *Genetics* 220: iyab229.
- 939 Bergström A., D. W. G. Stanton, U. H. Taron, L. Frantz, M.-H. S. Sinding, *et al.*, 2022 Grey wolf genomic history  
940 reveals a dual ancestry of dogs. *Nature* 607: 313–320.
- 941 Boyle K., and C. Renfrew, 2000 *Archaeogenetics : DNA and the population prehistory of Europe*, in *Published in*  
942 **2000** *in Cambridge by McDonald institute for archaeological research*, Cambridge : McDonald Institute for

- 943 Archaeological Research.
- 944 Broushaki F., M. G. Thomas, V. Link, S. López, L. van Dorp, *et al.*, 2016 Early Neolithic genomes from the  
945 eastern Fertile Crescent. *Science* 353: 499–503.
- 946 Brunson K., and D. Reich, 2019 The Promise of Paleogenomics Beyond Our Own Species. *Trends in Genetics*  
947 35: 319–329.
- 948 Carlhoff S., A. Duli, K. Nägele, M. Nur, L. Skov, *et al.*, 2021 Genome of a middle Holocene hunter-gatherer from  
949 Wallacea. *Nature* 596: 543–547.
- 950 Changmai P., K. Jaisamut, J. Kampuansai, W. Kutanan, N. E. Altınışık, *et al.*, 2022a Indian genetic heritage in  
951 Southeast Asian populations. *PLoS Genet.* 18: e1010036.
- 952 Changmai P., R. Pinhasi, M. Pietrusewsky, M. T. Stark, R. M. Ikehara-Quebral, *et al.*, 2022b Ancient DNA from  
953 Protohistoric Period Cambodia indicates that South Asians admixed with local populations as early as 1st-  
954 3rd centuries CE. *Sci. Rep.* 12: 22507.
- 955 Clemente F., M. Unterländer, O. Dolgova, C. E. G. Amorim, F. Coroado-Santos, *et al.*, 2021 The genomic  
956 history of the Aegean palatial civilizations. *Cell* 184: 2565–2586.e21.
- 957 De Schepper S., J. L. Ray, K. S. Skaar, H. Sadatzki, U. Z. Ijaz, *et al.*, 2019 The potential of sedimentary ancient  
958 DNA for reconstructing past sea ice evolution. *ISME J.* 13: 2566–2577.
- 959 Durand E. Y., N. Patterson, D. Reich, and M. Slatkin, 2011 Testing for ancient admixture between closely  
960 related populations. *Mol. Biol. Evol.* 28: 2239–2252.
- 961 Elise Lauterbur M., M. I. A. Cavassim, A. L. Gladstein, G. Gower, N. S. Pope, *et al.*, 2022 Expanding the  
962 stdpopsim species catalog, and lessons learned for realistic genome simulations. *bioRxiv*  
963 2022.10.29.514266.
- 964 Fernandes D. M., A. Mitnik, I. Olalde, I. Lazaridis, O. Cheronet, *et al.*, 2020 The spread of steppe and Iranian-  
965 related ancestry in the islands of the western Mediterranean. *Nat Ecol Evol* 4: 334–345.



- 966 Flegontov P., U. Işıldak, R. Maier, E. Yüncü, P. Changmai, *et al.*, 2023 Modeling of African population history  
967 using f-statistics is biased when applying all previously proposed SNP ascertainment schemes. *PLoS*  
968 *Genet.* 19: e1010931.
- 969 Fu Q., C. Posth, M. Hajdinjak, M. Petr, S. Mallick, *et al.*, 2016 The genetic history of Ice Age Europe. *Nature*  
970 534: 200–205.
- 971 Gower G., A. P. Ragsdale, G. Bisschop, R. N. Gutenkunst, M. Hartfield, *et al.*, 2022 Demes: a standard format  
972 for demographic models. *Genetics* 222.
- 973 Green R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, *et al.*, 2010 A draft sequence of the Neandertal  
974 genome. *Science* 328: 710–722.
- 975 Haak W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, *et al.*, 2015 Massive migration from the steppe was  
976 a source for Indo-European languages in Europe. *Nature* 522: 207–211.
- 977 Haber M., M. Mezzavilla, Y. Xue, and C. Tyler-Smith, 2016 Ancient DNA and the rewriting of human history: be  
978 sparing with Occam’s razor. *Genome Biol.* 17: 1.
- 979 Haber M., C. Doumet-Serhal, C. Scheib, Y. Xue, P. Danecek, *et al.*, 2017 Continuity and Admixture in the Last  
980 Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome  
981 Sequences. *Am. J. Hum. Genet.* 101: 274–282.
- 982 Haber M., J. Nassar, M. A. Almarri, T. Saupe, L. Saag, *et al.*, 2020 A Genetic History of the Near East from an  
983 aDNA Time Course Sampling Eight Points in the Past 4,000 Years. *Am. J. Hum. Genet.*
- 984 Harney É., H. May, D. Shalem, N. Rohland, S. Mallick, *et al.*, 2018 Ancient DNA from Chalcolithic Israel reveals  
985 the role of population mixture in cultural transformation. *Nat. Commun.* 9: 3336.
- 986 Harney É., N. Patterson, D. Reich, and J. Wakeley, 2021 Assessing the performance of qpAdm: a statistical tool  
987 for studying population admixture. *Genetics*.
- 988 Harris A. M., and M. DeGiorgio, 2017 Admixture and Ancestry Inference from Ancient and Modern Samples

- 989 through Measures of Population Genetic Drift. *Hum. Biol.* 89: 21–46.
- 990 Hollfelder N., G. Breton, P. Sjödin, and M. Jakobsson, 2021 The deep population history in Africa. *Hum. Mol.*  
991 *Genet.* 30: R2–R10.
- 992 Jónsson H., P. Sulem, B. Kehr, S. Kristmundsdóttir, F. Zink, *et al.*, 2017 Parental influence on human germline  
993 de novo mutations in 1,548 trios from Iceland. *Nature* 549: 519–522.
- 994 Jukes T. H., and C. R. Cantor, 1969 CHAPTER 24 - Evolution of Protein Molecules, pp. 21–132 in *Mammalian*  
995 *Protein Metabolism*, edited by Munro H. N. Academic Press.
- 996 Kamm J., J. Terhorst, R. Durbin, and Y. S. Song, 2020 Efficiently inferring the demographic history of many  
997 populations with allele count data. *J. Am. Stat. Assoc.* 115: 1472–1487.
- 998 Kelleher J., A. M. Etheridge, and G. McVean, 2016 Efficient Coalescent Simulation and Genealogical Analysis  
999 for Large Sample Sizes. *PLoS Comput. Biol.* 12: e1004842.
- 000 Koptekin D., E. Yüncü, R. Rodríguez-Varela, N. E. Altınışik, N. Psonis, *et al.*, 2023 Spatial and temporal  
001 heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean. *Curr.*  
002 *Biol.* 33: 41–57.e15.
- 003 Kristiansen K., 2016 Interpreting Bronze Age Trade and Migration, pp. 154–180 in *Human Mobility and*  
004 *Technological Transfer in the Prehistoric Mediterranean*, Cambridge University Press.
- 005 Lazaridis I., D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, *et al.*, 2016 Genomic insights into the origin of  
006 farming in the ancient Near East. *Nature* 536: 419–424.
- 007 Lazaridis I., A. Mitnik, N. Patterson, S. Mallick, N. Rohland, *et al.*, 2017 Genetic origins of the Minoans and  
008 Mycenaeans. *Nature* 548: 214–218.
- 009 Lazaridis I., S. Alpaslan-Roodenberg, A. Acar, A. Açikkol, A. Agelarakis, *et al.*, 2022a The genetic history of the  
010 Southern Arc: A bridge between West Asia and Europe. *Science* 377: eabm4247.
- 011 Lazaridis I., S. Alpaslan-Roodenberg, A. Acar, A. Açikkol, A. Agelarakis, *et al.*, 2022b A genetic probe into the

- 012 ancient and medieval history of Southern Europe and West Asia. *Science* 377: 940–951.
- 013 Lee J., B. K. Miller, J. Bayarsaikhan, E. Johannesson, A. Ventresca Miller, *et al.*, 2023 Genetic population  
014 structure of the Xiongnu Empire at imperial and local scales. *Sci Adv* 9: eadf3904.
- 015 Librado P., N. Khan, A. Fages, M. A. Kusliy, T. Suchan, *et al.*, 2021 The origins and spread of domestic horses  
016 from the Western Eurasian steppes. *Nature* 598: 634–640.
- 017 Lipson M., P.-R. Loh, A. Levin, D. Reich, N. Patterson, *et al.*, 2013 Efficient moment-based inference of  
018 admixture parameters and sources of gene flow. *Mol. Biol. Evol.* 30: 1788–1802.
- 019 Lipson M., P.-R. Loh, N. Patterson, P. Moorjani, Y.-C. Ko, *et al.*, 2014 Reconstructing Austronesian population  
020 history in Island Southeast Asia. *Nat. Commun.* 5: 4689.
- 021 Liu Y., X. Mao, J. Krause, and Q. Fu, 2021 Insights into human history from the first decade of ancient human  
022 genomics. *Science* 373: 1479–1484.
- 023 Llamas B., E. Willerslev, and L. Orlando, 2017 Human evolution: a tale from ancient genomes. *Philos. Trans. R.*  
024 *Soc. Lond. B Biol. Sci.* 372.
- 025 Loosdrecht M. van de, A. Bouzouggar, L. Humphrey, C. Posth, N. Barton, *et al.*, 2018 Pleistocene North African  
026 genomes link Near Eastern and sub-Saharan African human populations. *Science* 360: 548–552.
- 027 Maier R., P. Flegontov, O. Flegontova, U. Isildak, P. Changmai, *et al.*, 2023 On the limits of fitting complex  
028 models of population history to f-statistics.
- 029 Mallick S., A. Micco, M. Mah, H. Ringbauer, I. Lazaridis, *et al.*, 2023 The Allen Ancient DNA Resource (AADR):  
030 A curated compendium of ancient human genomes. *bioRxiv* 2023.04.06.535797.
- 031 Marchi N., L. Winkelbach, I. Schulz, M. Brami, Z. Hofmanová, *et al.*, 2022 The genomic origins of the world's  
032 first farmers. *Cell* 185: 1842–1859.e18.
- 033 Marcus J. H., C. Posth, H. Ringbauer, L. Lai, R. Skeates, *et al.*, 2020 Genetic history from the Middle Neolithic  
034 to present on the Mediterranean island of Sardinia. *Nat. Commun.* 11: 939.

- 035 Maróti Z., E. Neparáczi, O. Schütz, K. Maár, G. I. B. Varga, *et al.*, 2022 The genetic origin of Huns, Avars, and  
036 conquering Hungarians. *Curr. Biol.* 32: 2858–2870.e7.
- 037 Martin S. H., J. W. Davey, and C. D. Jiggins, 2015 Evaluating the use of ABBA-BABA statistics to locate  
038 introgressed loci. *Mol. Biol. Evol.* 32: 244–257.
- 039 Mathieson I., S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, *et al.*, 2018 The genomic  
040 history of southeastern Europe. *Nature* 555: 197–203.
- 041 McVean G., 2009 A genealogical interpretation of principal components analysis. *PLoS Genet.* 5: e1000686.
- 042 Mitchell K. J., and N. J. Rawlence, 2021 Examining Natural History through the Lens of Palaeogenomics.  
043 *Trends Ecol. Evol.* 36: 258–267.
- 044 Moots H. M., M. Antonio, S. Sawyer, J. P. Spence, V. Oberreiter, *et al.*, 2023 A genetic history of continuity and  
045 mobility in the Iron Age central Mediterranean. *Nat Ecol Evol* 7: 1515–1524.
- 046 Narasimhan V. M., N. Patterson, P. Moorjani, N. Rohland, R. Bernardos, *et al.*, 2019 The formation of human  
047 populations in South and Central Asia. *Science*.
- 048 Nelson D., J. Kelleher, A. P. Ragsdale, C. Moreau, G. McVean, *et al.*, 2020 Accounting for long-range  
049 correlations in genome-wide simulations of large cohorts. *PLoS Genet.* 16: e1008619.
- 050 Nielsen S. V., A. H. Vaughn, K. Leppälä, M. J. Landis, T. Mailund, *et al.*, 2023 Bayesian inference of admixture  
051 graphs on Native American and Arctic populations. *PLoS Genet.* 19: e1010410.
- 052 Ning C., T. Li, K. Wang, F. Zhang, T. Li, *et al.*, 2020 Ancient genomes from northern China suggest links  
053 between subsistence changes and human migration. *Nat. Commun.* 11: 2700.
- 054 Olalde I., S. Mallick, N. Patterson, N. Rohland, V. Villalba-Mouco, *et al.*, 2019 The genomic history of the Iberian  
055 Peninsula over the past 8000 years. *Science* 363: 1230–1234.
- 056 Oliveira S., K. Nägele, S. Carlhoff, I. Pugach, T. Koesbardiati, *et al.*, 2022 Ancient genomes from the last three  
057 millennia support multiple human dispersals into Wallacea. *Nat Ecol Evol* 6: 1024–1034.



- 058 Papac L., M. Ernée, M. Dobeš, M. Langová, A. B. Rohrlach, *et al.*, 2021 Dynamic changes in genomic and  
059 social structures in third millennium BCE central Europe. *Sci Adv* 7.
- 060 Patterson N., A. L. Price, and D. Reich, 2006 Population Structure and Eigenanalysis. *PLoS Genet.* 2: e190.
- 061 Patterson N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, *et al.*, 2012 Ancient admixture in human history.  
062 *Genetics* 192: 1065–1093.
- 063 Patterson N., M. Isakov, T. Booth, L. Büster, C.-E. Fischer, *et al.*, 2022 Large-scale migration into Britain during  
064 the Middle to Late Bronze Age. *Nature* 601: 588–594.
- 065 Peter B. M., 2016 Admixture, Population Structure, and F-Statistics. *Genetics* 202: 1485–1501.
- 066 Peter B. M., 2022 A geometric relationship of F2, F3 and F4-statistics with principal component analysis.  
067 *PHILOSOPHICAL TRANSACTIONS B*.
- 068 Pickrell J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele  
069 frequency data. *PLoS Genet.* 8: e1002967.
- 070 Reich D., A. L. Price, and N. Patterson, 2008 Principal component analysis of genetic data. *Nat. Genet.* 40:  
071 491–492.
- 072 Reich D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, 2009 Reconstructing Indian population history.  
073 *Nature* 461: 489–494.
- 074 Reich D., N. Patterson, D. Campbell, A. Tandon, S. Mazieres, *et al.*, 2012 Reconstructing Native American  
075 population history. *Nature* 488: 370–374.
- 076 Salazar L., R. Burger, J. Forst, R. Barquera, J. Nesbitt, *et al.*, 2023 Insights into the genetic histories and  
077 lifeways of Machu Picchu’s occupants. *Sci Adv* 9: eadg3377.
- 078 Schmid C., and S. Schiffels, 2023 Estimating human mobility in Holocene Western Eurasia with large-scale  
079 ancient genomic data. *Proc. Natl. Acad. Sci. U. S. A.* 120: e2218375120.
- 080 Schuenemann V. J., A. Peltzer, B. Welte, W. P. van Pelt, M. Molak, *et al.*, 2017 Ancient Egyptian mummy

- 081 genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nat. Commun.* 8:  
082 15694.
- 083 Sirak K. A., D. M. Fernandes, M. Lipson, S. Mallick, M. Mah, *et al.*, 2021 Social stratification without genetic  
084 differentiation at the site of Kulubnarti in Christian Period Nubia. *Nat. Commun.* 12: 7283.
- 085 Skoglund P., J. C. Thompson, M. E. Prendergast, A. Mitnik, K. Sirak, *et al.*, 2017 Reconstructing Prehistoric  
086 African Population Structure. *Cell* 171: 59–71.e21.
- 087 Skourtanioti E., Y. S. Erdal, M. Frangipane, F. Balossi Restelli, K. A. Yener, *et al.*, 2020 Genomic History of  
088 Neolithic to Bronze Age Anatolia, Northern Levant, and Southern Caucasus. *Cell* 181: 1158–1175.e28.
- 089 Skourtanioti E., H. Ringbauer, G. A. Gneccchi Ruscone, R. A. Bianco, M. Burri, *et al.*, 2023 Ancient DNA reveals  
090 admixture history and endogamy in the prehistoric Aegean. *Nat Ecol Evol* 7: 290–303.
- 091 Slatkin M., and F. Racimo, 2016 Ancient DNA and human history. *Proc. Natl. Acad. Sci. U. S. A.* 113: 6380–  
092 6387.
- 093 Soraggi S., and C. Wiuf, 2019 General theory for stochastic admixture graphs and F-statistics. *Theor. Popul.*  
094 *Biol.* 125: 56–66.
- 095 Spyrou M. A., K. I. Bos, A. Herbig, and J. Krause, 2019 Ancient pathogen genomics as an emerging tool for  
096 infectious disease research. *Nat. Rev. Genet.* 20: 323–340.
- 097 Taylor W. T. T., P. Librado, C. J. American Horse, C. Shield Chief Gover, J. Arterberry, *et al.*, 2023 Early  
098 dispersal of domestic horses into the Great Plains and northern Rockies. *Science* 379: 1316–1323.
- 099 Tricou T., E. Tannier, and D. M. de Vienne, 2022 Ghost Lineages Highly Influence the Interpretation of  
100 Introgression Tests. *Syst. Biol.* 71: 1147–1158.
- 101 Wang C.-C., S. Reinhold, A. Kalmykov, A. Wissgott, G. Brandt, *et al.*, 2019 Ancient human genome-wide data  
102 from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nat. Commun.* 10:  
103 590.

- 104 Wang K., I. Mathieson, J. O'Connell, and S. Schiffels, 2020 Tracking human population structure through time  
105 from whole genome sequences. *PLoS Genet.* 16: e1008552.
- 106 Wang C.-C., H.-Y. Yeh, A. N. Popov, H.-Q. Zhang, H. Matsumura, *et al.*, 2021 Genomic insights into the  
107 formation of human populations in East Asia. *Nature* 591: 413–419.
- 108 Wibowo M. C., Z. Yang, M. Borry, A. Hübner, K. D. Huang, *et al.*, 2021 Reconstruction of ancient microbial  
109 genomes from the human gut. *Nature* 594: 234–239.
- 110 Williams M., and J. Teixeira, 2020 A genetic perspective on human origins. *Biochem.* 42: 6–10.
- 111 Wood S. N., 2004 Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive  
112 Models. *J. Am. Stat. Assoc.* 99: 673–686.
- 113 Yang M. A., X. Fan, B. Sun, C. Chen, J. Lang, *et al.*, 2020 Ancient DNA indicates human population shifts and  
114 admixture in northern and southern China. *Science* 369: 282–288.
- 115 Yüncü E., U. Işıldak, M. P. Williams, C. D. Huber, O. Flegontova, *et al.*, 2023 False discovery rates of qpAdm-  
116 based screens for genetic admixture. *bioRxiv* 2023.04.25.538339.
- 117 Zheng Y., and A. Janke, 2018 Gene flow analysis method, the D-statistic, is robust in a wide parameter space.  
118 *BMC Bioinformatics* 19: 10.
- 119