Performance of *qpAdm*-based screens for genetic admixture on admixture-graph-shaped histories and stepping-stone landscapes

Olga Flegontova[1,2,*], Ulaş Işıldak[1,3,*], Eren Yüncü[1,4,*], Matthew P. Williams[5], Christian D. Huber[5], Jan Kočí[1], Leonid A. Vyazov[1], Piya Changmai[1], Pavel Flegontov[1,6,§]

[1] *Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czechia*

[2] *Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, České Budějovice, Czechia*

[3] *Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Jena, Germany*

[4] *Department of Biological Sciences, Middle East Technical University, Ankara, Turkey*

[5] *Department of Biology, Eberly College of Science, Pennsylvania State University, PA, USA*

[6] *Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA*

* The authors contributed equally

§ Correspondence to: Pavel Flegontov

(pavel.flegontov@osu.cz, pavel_flegontov@fas.harvard.edu, pflegontov@gmail.com)

Abstract

qpAdm is a statistical tool that is often used in exploratory archaeogenetic studies for finding optimal admixture models of population history. Despite its popularity, *qpAdm* remains untested on histories in the form of admixture graphs of random topology or stepping-stone landscapes. We analyzed data from such simulations and found that while for admixture-graph-shaped histories there exist simple solutions (temporal stratification) for minimizing false findings of gene flow, in the case of stepping-stone landscapes the method generates results that do not appear suspect but are misleading: feasible *qpAdm* models are either accurate but simplistic in the context of landscapes, or highly inaccurate in the case of multi-component models. This is largely is due to two reasons: 1) because of complex migration networks that violate the assumptions of the method, there is poor correlation between *qpAdm p*-values and model optimality in many sections of the parameter space; 2) admixture fraction estimates between 0 and 1 are largely restricted to symmetric source configurations around targets, hence popular [0, 1] model plausibility criteria confound analyses of landscape-type demographies, unless their interpretations are explicitly spatial. For many species/regions/periods archaeogenetic sampling is very sparse and may be random with respect to population density of ancient individuals. In this situation only a specific combination of landscape properties and feasibility criteria allows to efficiently reject highly asymmetric non-optimal models most abundant in random deme sets. This problem may obscure useful signal (rare optimal models) and might be responsible for some claims about rapid long-distance migrations in the literature.

Introduction

Although a broad range of methods exists for reconstructing population history from genome-wide autosomal single nucleotide polymorphism (SNP) data, just a few methods became the cornerstone of archaeogenetic studies: principal component analysis (PCA) (Patterson et al. 2006); an unsupervised or supervised algorithm for admixture inference in individuals, *ADMIXTURE* (Alexander et al. 2009); formal tests for admixture such as f_3-statistics (Patterson et al. 2012, Peter 2016, Soraggi and Wiuf 2019, Peter 2022) and *D*-statistics (Green et al. 2010, Durand et al. 2011); and a tool for fitting two-component and more complex admixture models to populations, *qpAdm* (Haak et al. 2015, Harney et al. 2021). The popularity of these methods is explained by their relatively modest computational requirements and versatility since they can analyze unphased biallelic genotype data of various types (pseudo-haploid or diploid), generated using either targeted enrichment on a panel of sites or shotgun sequencing technologies, and low-coverage ancient genomes with high proportions of missing data (Harney et al. 2021). However, only a few studies were devoted to testing the performance of these diverse methods on simulated genetic data (Alexander et al. 2009, Harney et al. 2021, Lazaridis et al. 2017, Martin et al. 2014, McVean 2009, Moreno-Mayar et al. 2018, Ning et al. 2020b, Soraggi and Wiuf 2019, Speidel et al. 2024), and realistically complex population histories remain virtually unexplored in this respect: both those approximated best by admixture graphs and isolation by distance (stepping-stone models, see Novembre and Stephens 2008, Duforet-Frebourg and Slatkin 2016).

Isolation by distance may be more realistic than admixture-graph-shaped (AGS) histories, and PCA can generate similar results under multiple scenarios. PCA is often used by archaeogenetic studies as a first line of analysis, providing an overview of population structure, and helping to propose hypotheses about migration and admixture. Distribution of individual genomes in two- or higher dimensional spaces of principal components (PCs) does not have unambiguous interpretation since even under ideal conditions (in the absence of missing data, batch artefacts, and selection signals) it is affected by both genetic drift and admixture (McVean 2009). Nevertheless, if context information such as geographical coordinates and dates for ancient individuals is available, PCA is routinely used for revealing "genetic clines", often hypothesized as signs of admixture between originally isolated groups

at the ends of such clines. However, multiple studies on simulated data showed that clinal PCA patterns do not necessarily indicate mass migration and may arise even in the absence of admixture: Lissajous curves such as horseshoe-shaped clines appear in PC spaces whenever smooth gradients of various origin exist in data (Podani and Miklós 2002, Diaconis et al. 2008, Novembre and Stephens 2008, Frichot et al. 2012, Duforet-Frebourg and Slatkin 2016, Estavoyer and Francois 2022). Thus, similar clinal patterns in PCAs based on genetic data can arise for many reasons: mass migration and admixture between previously isolated groups (i.e., "admixture clines"), slower migration in the isolation-by-distance context (Novembre and Stephens 2008), isolation by time in a panmictic population (Duforet-Frebourg and Slatkin 2016), and serial founder models without any gene flow (Estavoyer and Francois 2022). It was demonstrated that the distribution of individuals in the PC space depends on the expected coalescent time (McVean 2009), hence distinct demographic models with the same expected coalescence times are expected to have the same PCA projections. Additionally, imbalanced sampling across genetically divergent populations affects PCA results substantially (McVean 2009) and may obscure smooth Lissajous curves by artefactual sharp angles (Novembre and Stephens 2008).

Other fundamental problems of PCA are its sensitivity to outliers (Nguyen and Holmes 2019, Armstrong et al. 2022) and the fact that the signal is spread over *s*-1 or *v*-1 dimensions, whichever is smaller (where *s* and *v* are counts of samples and variables, respectively), but only two or rarely three dimensions are visualized and interpreted, leading to substantial distortions of the original genetic space (Battey et al. 2021). Other methods exist that allow embedding into two- or three-dimensional spaces but are computationally expensive and non-deterministic: t-SNE (van der Maaten and Hinton 2008) and UMAP (McInnes and Healy. 2018) recover local structure (clusters of very similar individuals) accurately at the expense of global structure (Diaz-Papkovich et al. 2019, Battey et al. 2021), and neural networks such as variational autoencoders (VAE) recover both local and global structure more accurately than PCA does (Battey et al. 2021, Geleta et al. 2023). Considering this complexity, using further methods to correlate PCA/UMAP/VAE results with other lines of evidence is necessary for studying migration history (Reich et al. 2008).

Formal tests for genetic admixture such as f_3-statistics and D/f_4-statistics are often used exactly for this purpose: to test for the possibility that a certain cline spotted in PC space

103 might be a result of migration and admixture of previously isolated ancestries and does not
104 reflect isolation by distance or recurrent bottlenecks. *D*- and f_4-statistics, which are identical
105 except for the denominator and are not affected by genetic drift, test if an unrooted four-
106 population tree fits the data (Reich et al. 2009, Green et al. 2010, Durand et al. 2011,
107 Patterson et al. 2012). A statistically significant deviation of the statistic from 0 (estimated
108 using jackknife or bootstrap resampling) means that either the assumed tree topology is
109 wrong, or gene flow occurred between a pair of branches in the tree, assuming that at most
110 one mutation per locus occurred in the history of the four populations (Durand et al. 2011,
111 Patterson et al. 2012) and SNP ascertainment bias (Flegontov et al. 2023) is absent. However,
112 interpretation of these statistics is ambiguous since gene flow directionality remains
113 unknown, and two pairs of branches can be responsible for a deviation of the statistic from 0
114 (Lipson 2020). Since gene flow may be mediated by ghost groups related only distantly to the
115 sampled groups at the branch tips (Tricou et al. 2022), excluding one pair of branches due to
116 geographical and temporal plausibility of gene flow is also difficult. And interpretating
117 deviations of *D*- and f_4-statistics from 0 (or lack thereof) becomes hardly possible if both
118 branch pairs are connected by detectable gene flows, a typical situation on isolation-by-
119 distance landscapes. In general, unambiguous interpretation of non-rejection of the null
120 hypothesis (no deviation from 0) should be avoided.

121 "Admixture" f_3-statistics of the form f_3(target; proxy source$_1$, proxy source$_2$) constitute
122 another formal test for admixture (Patterson et al. 2012). Significant deviation of such a
123 statistic from 0 in the negative direction (Z-score below -3) is considered proof of admixture
124 since allele frequencies at most sites are intermediate in the target group between those in
125 the proxy sources (Patterson et al. 2012). However, "admixture" f_3-statistics are usually only
126 applicable for detection of recent admixture events since they become positive when post-
127 admixture genetic drift on the target lineage moves allele frequencies away from these
128 intermediate values (Patterson et al. 2012, Peter 2016, Peter 2022). When the sources are
129 more distant genetically, f_3-statistics remain negative for longer periods of time.

130 Considering these complications, more sophisticated tests for genetic admixture are needed.
131 The *qpAdm* method introduced by Haak *et al*. (2015) was developed with the admixture graph
132 paradigm in mind. It is based on matrices of f_4-statistics and does not require detailed
133 knowledge of population phylogeny beyond a few assumptions (Lazaridis et al. 2016, Harney

134 et al. 2021). This method tests admixture models in the form of combinations of proxy
135 ancestral groups ("sources" or "references", Lazaridis et al. 2016) for a "target" (or "test")
136 population, given a genetically diverse set of "outgroup" populations, and infers ancestry
137 proportions in the target group contributed by the lineages represented by the proxy sources
138 ("outgroups" are often termed "right" populations for convenience since they are usually not
139 outgroups in the phylogenetic sense, and they were termed "references" by Harney *et al*.
140 2021). See **Box 1** for definitions of various *qpAdm*-related terms used in this study.

141
142 **Box 1.** Terminology used in this study for describing simulations, admixture models, results of screens
143 for admixture, and *qpAdm* protocols.

admixture-graph-shaped (AGS) history	A history represented as a (usually bifurcating) directed acyclic graph, composed of episodes of population divergence and of mass migration and admixture of two source populations ("pulse admixture events").
stepping-stone landscape (SSL)	An approximation for two-dimensional isolation-by-distance landscapes ("genetic landscapes"); here we describe an implementation used in this study. A stepping-stone landscape is originally "unfolded" from a single founder deme via a serial founder process that can be represented by a bifurcating tree (Estavoyer and Francois 2022), or via multifurcation, i.e. star radiation. This stage is followed by a gene-flow era represented as an undirected graph connecting demes, where each edge represents a bidirectional gene flow, and gene-flow intensities are allowed to differ in the forward and reverse directions. Gene flows on this undirected graph happen in one or more epochs, with unidirectional gene-flow intensities sampled randomly from uniform or non-uniform distributions. In this study all landscapes are based on the triangular lattice, and node degree varies from 3 to 6.
gene-flow epoch	A period of unchanging gene-flow intensities on the landscape (they are assigned randomly at the beginning of the epoch).
gene-flow era	A collection of gene-flow epochs characterized by the same topology of the undirected gene-flow graph, but not necessarily by stable gene-flow intensities.
"right" (or outgroup) populations/groups	In a *qpAdm* protocol, these are reference populations needed for testing admixture models composed of a target and one or several proxy source groups.
target (or test) population/group	In an admixture model, this is a population/group whose genetic history is being modelled.

true (ancestry) source	In the context of simulated AGS histories, this is a population directly participating in an admixture event(s) giving rise to a target group, and its ancestral (unsampled) population before its merging with other populations deeper in time. In contrast to AGS histories, the concept of true and false ancestry sources is vague in the case of stepping-stone landscapes where all neighboring demes continuously exchange migrants; we can just rank sources into more and less appropriate.
proxy (ancestry) source	In the context of simulated admixture graphs, this is a sampled population included in an admixture model as a potential source and assumed to be cladal with one of true ancestry sources. In the context of SSL, this is any deme included in an admixture model as a potential source.
model complexity	Number of proposed proxy ancestry sources in an admixture model "target = proxy source$_1$ + proxy source$_2$ + … proxy source$_i$".
"left" populations/groups	In a *qpAdm* model, these are proxy sources and a target.
reference populations/groups	"Right" and proxy source populations combined.
qpAdm experiment	A set of all possible (and progressively more complex) models for a given target and a set of reference groups. An experiment is stopped as soon as at least one feasible model for the target is encountered.
"trailing" models	For an *n*-way admixture model, these are all possible simpler models composed of the target and/or proxy sources from that *n*-way model (but not from other *n*-way models in the experiment): "target = proxy source$_1$ + proxy source$_i$ + …" and "proxy source$_1$ = proxy source$_i$ + …".
(composite) model feasibility criterion	A set of criteria for identifying fitting (feasible) *qpAdm* models that relies on both estimated admixture proportions (or fractions, abbreviated as EAF) and *p*-values. In general terms, a composite criterion is composed of three conditions: 1) on admixture proportion estimates in an *n*-way model, 2) on the *p*-value of the *n*-way model, and 3) on *p*-values of the "trailing" simpler models. The *p*-value thresholds no. 2 and 3 can be the same or different. Here is an example applied in this study to AGS histories: 1) EAF $\pm$ 2 standard errors $\in$ (0, 1); 2) the *p*-value for an *n*-way model $\geq$0.01; 3) *p*-values for "trailing" simpler models are <0.01. Other versions of this criterion are also found in the literature: with different values of the *p*-value cut-off (usually 0.05, 0.01, or 0.001); with limits on EAF, but not on their standard errors; with or without conditions on simpler models. In the case of SSL histories, 36 composite feasibility criteria were tested.

feasible (or fitting) *qpAdm* model	A *qpAdm* model satisfying the feasibility criterion above.
positive/negative admixture model	An admixture model supported/not supported by one or several analytical tools such as *qpAdm*, PCA, *ADMIXTURE*.
positive rate	The fraction of all tested *qpAdm* models that are positive.
false discovery rate (FDR)	The fraction of admixture models of a certain complexity (e.g., two-way) satisfying the *qpAdm* model feasibility criteria but classified as false considering the simulated graph topology and simulated admixture proportions. For topological criteria used for classifying two-way admixture models into true and false ones see the Results and Methods sections. The term FDR is also applied to outcomes of more complex admixture inference pipelines composed of several methods.
false omission rate (FOR)	The fraction of feasible *qpAdm* models that are classified as true considering the simulated graph topology and simulated admixture proportions but are not supported by another method (*qpAdm* model competition, PCA, or *ADMIXTURE*) or a combination of methods.
"rotating" *qpAdm* protocol	A protocol having the following feature: a large subset of reference populations or all of them are distributed between the "right" and "left" sets according to the principle "whatever is not in the left set is in the right set", testing all possible bisections of this sort for a given rotated set and a given range of model complexities. In the most extreme case, target groups are also included in this rotation. Model testing starts from the simplest one-way models and moves on to the next complexity level if all simple models for a given target are rejected according to the composite feasibility criterion. The goal of this approach, as compared to "non-rotating" protocols, is to increase the power of the method to reject non-optimal proxy ancestry sources.
"non-rotating" *qpAdm* protocol	This protocol we also term "standard" or "basic": all models are tested with one or few fixed sets of "right" groups which usually pre-date "left" groups or are contemporaneous with them. In practice, modern populations genetically divergent from the target are often included in such a "right" set if ancient reference groups are unavailable. Model testing starts from the simplest one-way models and moves on to the next complexity level if all simple models for a given target are rejected according to the composite feasibility criterion.
"distal" *qpAdm* model	A model where the target post-dates or is contemporaneous with all proxy sources, in other words, they are temporally stratified.
"proximal" *qpAdm* model	A model where the target pre-dates at least one proxy source.

"distal" *qpAdm* protocol	A rotating or non-rotating *qpAdm* protocol considering distal models only.
"proximal" *qpAdm* protocol	A rotating or non-rotating *qpAdm* protocol without temporal stratification attempted, that is considering both proximal and distal models. Such a protocol should not be confused with a protocol deliberately considering proximal models only.
model competition	A *qpAdm* protocol which starts from sets of alternative feasible *qpAdm* models of a certain complexity level for a given target, outcomes of a non-rotating protocol. Alternative proxy sources from these models form a rotated set. Rotation can be performed in two ways (groups from the rotated set are placed in the "left" set one by one, and the other groups from the rotated set are placed in the "right" set; or groups from the rotated set are placed in both "left" and "right" sets one by one), and the composite feasibility criterion is applied. The goal of this approach is to increase the power of the method to reject non-optimal proxy ancestry sources.
proximal model competition *qpAdm* protocol	As introduced by Narasimhan *et al*. (2019), this is a non-rotating *qpAdm* protocol with temporal stratification of "right" and "left" sets, with a subsequent model competition step, and with no (or very limited) temporal stratification of targets and proxy sources at both steps.
maximal absolute difference between estimated admixture fractions and equal fractions	This metric, abbreviated as $\max\lvert EAF–EF\rvert$, measures how far <u>estimated</u> admixture fractions (EAF) in an admixture model deviate from equal fractions ($\frac{1}{n}$ for an *n*-way models).
maximal standard error of EAF	This is a maximal standard error reported by *qpAdm* for admixture fractions in a model.
$AEH_T - av.\ AEH_S$	Difference between number of admixture events in target's history and average number of admixture events in sources' histories (In the context of an AGS history).

Nine years later we find *qpAdm*-based protocols routinely employed in large-scale screens of ancient human or animal populations for admixture (often between closely related sources) and used as formal tests for admixture (see Lazaridis et al. 2016, Skoglund et al. 2017, Harney et al. 2018, Mathieson et al. 2018, Antonio et al. 2019, Narasimhan et al. 2019, Fernandes et al. 2020, Marcus et al. 2020, Ning et al. 2020a, Wang et al. 2020, Yang et al. 2020, Carlhoff et al. 2021, Papac et al. 2021, Librado et al. 2021, Sirak et al. 2021, Wang et al. 2021, Yaka et al. 2021, Zhang et al. 2021, Bergström et al. 2022, Changmai et al. 2022a, Changmai et al. 2022b,

Gnecchi-Ruscone et al. 2022, Lazaridis et al. 2022, Maróti et al. 2022, Oliveira et al. 2022, Patterson et al. 2022, Brielle et al. 2023, Lee et al. 2023, Taylor et al. 2023, Nakatsuka et al. 2023, Zeng et al. 2023, Speidel et al. 2024 for examples). *qpAdm* fits admixture models to a matrix of f_4-statistics of the form f_4("left" group$_i$, "left" group$_j$; "right" group$_i$, "right" group$_j$), which in the absence of missing data at the group level can be reduced to a smaller matrix f_4(target group, "left" group$_j$; "right" group$_1$, "right" group$_j$), considering algebraic relationships between different f_4-statistics (Peter 2016).

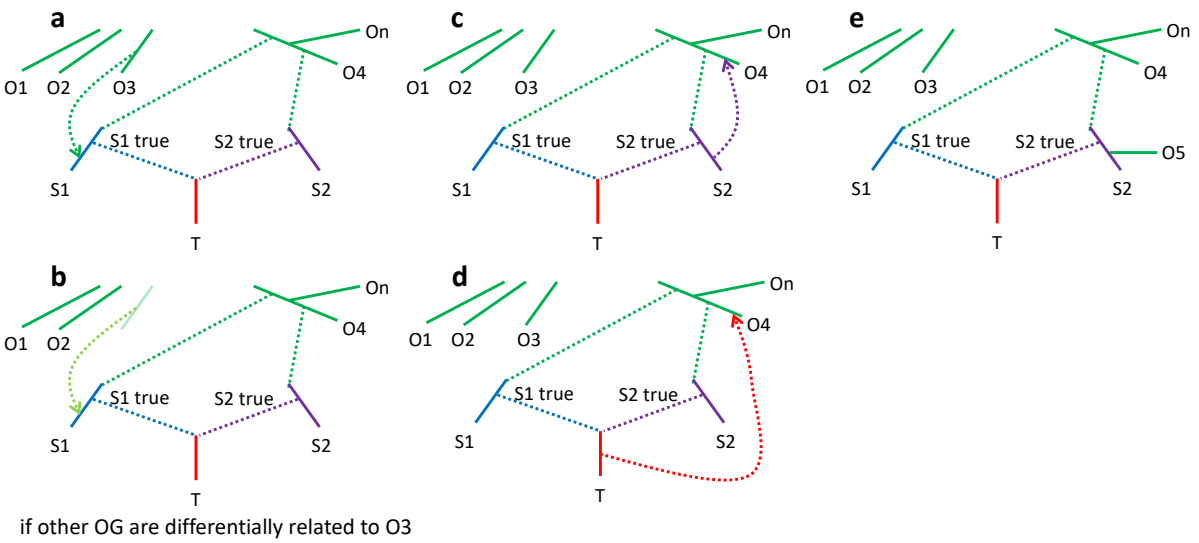


Figure 1. Admixture graphs showing an exhaustive list of assumption violations of the standard *qpAdm* protocol that may lead to rejection of the true simple model, and thus prompt the researcher to test overly complex models. (**a**) A gene flow from an outgroup O* to a proxy source after the divergence of the latter from the true source. (**b**) A gene flow from an unsampled source to a proxy source after the divergence of the latter from the true source. This case is problematic only if the outgroups are differentially related to the unsampled source. (**c**) A gene flow from a proxy source to an outgroup after the divergence of the former from the true source. (**d**) A gene flow from a target to an outgroup. (**e**) An outgroup is cladal with a proxy source.

A *qpAdm* protocol that has become the standard in archaeogenetics (Lazaridis et al. 2016) can be broken down into two parts: estimation of the number of gene flows connecting the "right" and "left" sets (this method was first published as a tool named "*qpWave*", Reich et al. 2012) and inference of admixture proportions in a target group in the "left" set (Haak et al. 2015). *qpWave* tests for the number of distinct gene flows connecting the "right" and "left" population sets, does not infer directionality of gene flows, and does not identify recipients of gene flow in the "left" or "right" population sets. Notably, the standard *qpAdm* protocol relies on the following assumptions in the admixture graph paradigm (Lazaridis et al. 2016,

177 Harney et al. 2021): 1) there is at least one "right" population differentially related to the
178 proxy sources; 2) proxy sources are strictly cladal with the true ancestral admixing sources
179 (**Fig. 1a, b**), 3) there are no gene flows to populations located in the “right” set from the proxy
180 source or target lineages either after the split of the proxy source from the true admixing
181 source population or between the target population and the admixture event that gave rise
182 to it (**Fig. 1c-e**). In the context of our study, true sources are unsampled populations that
183 participated in a simulated admixture event (labelled as “S1 true” and “S2 true” in **Fig. 1**, see
184 also **Box 1**). A more concise definition of these assumptions exists: there should be no genetic
185 drift shared exclusively by proxy sources and “right” populations (but not by the target) and
186 shared exclusively by the target and “right” populations (but not by proxy sources).

187 If the above assumptions are satisfied, it is safe to say that *qpWave/qpAdm* rejections of
188 simpler models, and a failure to reject more complex models, are the result of a genuinely
189 complex admixture history that connects the source and target populations rather than the
190 false rejection of the simple model due to violations of any one of the assumptions described
191 above. Most notably, violations of the second or third assumptions raise the probability of
192 rejecting a simpler (true) model if enough data is available and prompt the researcher to test
193 more complex (false) models (such as in **Fig. 1** rejecting a two-source *qpAdm* model and
194 exploring three-source models).

195 Harney *et al*. (2021) demonstrated on AGS simulated histories that, if the *qpAdm* assumptions
196 are satisfied, it is highly favorable for statistical power of the method (for distinguishing
197 between alternative proxy sources that are unequally distant genetically from the true
198 ancestry source) to move at least some alternative proxy sources between the "left" and
199 "right" sets. In other words, having "on the right" populations that do not violate the
200 topological assumptions of *qpAdm*, but are closely related to proxy sources on the "left",
201 increases the statistical power greatly (see also Ning et al. 2020b and Williams et al. 2024 for
202 demonstrations of this on simple and complex AGS simulated histories, respectively, and
203 Speidel et al. 2024 for another exploration of a *qpAdm*-like algorithm on simple AGS
204 simulations).

205 This new type of *qpAdm* protocols, termed "rotating" protocol, has been adopted in
206 archaeogenetics widely (see, e.g., Skoglund et al. 2017, Harney et al. 2019, Narasimhan et al.
207 2019, Olalde et al. 2019, Carlhoff et al. 2021, Fernandes et al. 2021, Librado et al. 2021,

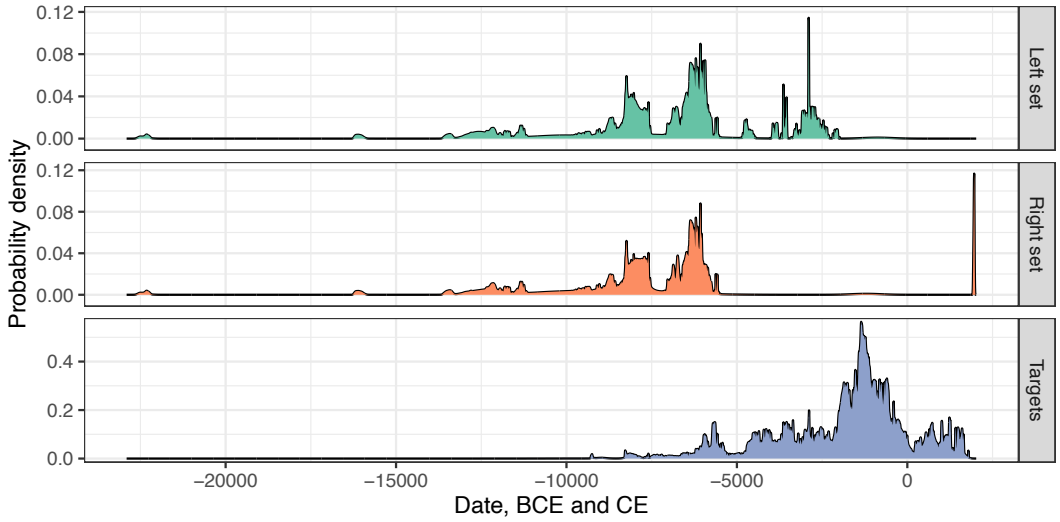
Bergström et al. 2022, Lazaridis et al. 2022, Oliveira et al. 2022, Taylor et al. 2023, Zeng et al. 2023, Speidel et al. 2024). The most extreme version of the "rotating" protocol simply divides a set of reference populations into all possible combinations of "right" and "proxy source" subsets of certain sizes and rotates these combinations through successive *qpAdm* analyses.

Additional constraints can be placed on the rotating combinations such as restricting a set of groups (usually highly divergent from the target) to remain in the “right” set in all models. When evaluating the totality of multiple *qpAdm* tests, the simplest feasible models (e.g., one-way, i.e., unadmixed) are usually favored, and increasingly complex models are explored upon the rejection of simpler models. Model rejection for the simplest models is made according to a chosen *p*-value threshold such that *qpAdm* models are considered feasible or "fitting" the data when the *p*-value is above such a threshold (Skoglund et al. 2017, Harney et al. 2018, Narasimhan et al. 2019, Olalde et al. 2019, Yang et al. 2020, Carlhoff et al. 2021, Fernandes et al. 2021, Librado et al. 2021, Zhang et al. 2021, Bergström et al. 2022, Lazaridis et al. 2022, Oliveira et al. 2022, Nakatsuka et al. 2023, Taylor et al. 2023, Zeng et al. 2023, Kerdoncuff et al. 2024, Speidel et al. 2024). As an additional criterion of a fitting model, all inferred admixture proportions (see, e.g., Harney et al. 2018, Olalde et al. 2019, Yang et al. 2020, Zhang et al. 2021, Lazaridis et al. 2022, Oliveira et al. 2022, Carlhoff et al. 2023, Zeng et al. 2023, Kerdoncuff et al. 2024, Speidel et al. 2024), or proportions $\pm$ 2 standard errors (Narasimhan et al. 2019), may be required to lie between 0 and 1. It is important to remember that the statistical significance of the *qpAdm/qpWave* test is, strictly speaking, a function of model rejection, and thus the failure to reject a model may have underlying causes other than approximating the true situation well enough (such as lack of statistical power or a lack of suitable “right” groups that capture the divergent ancestry sources amongst the “left” set of populations).

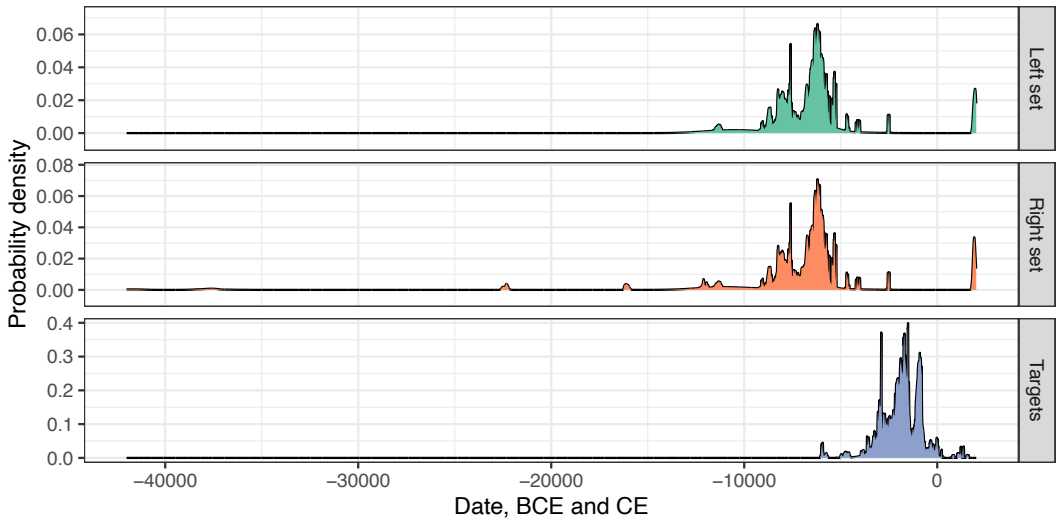
A less exhaustive version of the rotating *qpAdm* protocol, termed "model competition" (e.g., Narasimhan et al. 2019, Fernandes et al. 2020, Carlhoff et al. 2021, Sirak et al. 2021, Zhang et al. 2021, Maróti et al. 2022, Brielle et al. 2023, Lee et al. 2023), is used even more widely than the basic rotating protocol. It involves an initial (standard non-rotating) *qpAdm* analysis on a few source populations (see **Box 1**). Upon identifying a list of plausible sources for a target, the algorithm re-tests feasible models for this target rotating these plausible sources

between the "left" and "right" sets with the expectation of improving the power to reject models including proxy sources that are genetically distant from the true sources.

a, Lazaridis *et al*. 2022, distal rotating protocol



b, Narasimhan *et al*. 2019, distal rotating protocol



c, Narasimhan *et al*. 2019, proximal model competition protocol

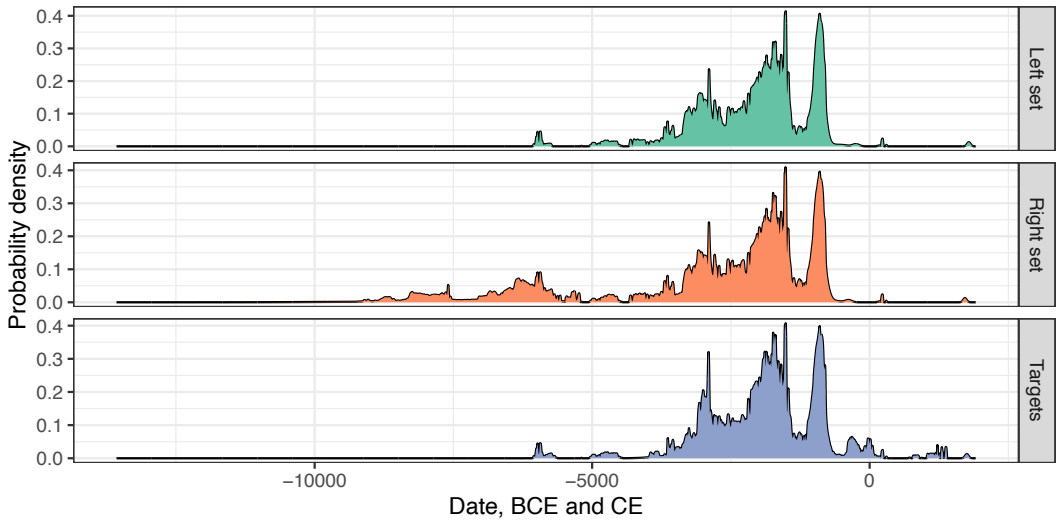


Figure 2. Distributions of radiocarbon and calendar dates for populations sets analyzed with the distal rotating *qpAdm* protocols by Lazaridis *et al*. 2022 (**a**) and Narasimhan *et al*. 2019 (**b**), and with the proximal model competition protocol from Narasimhan *et al*. 2019 (**c**). Probability density curves are shown for three sets of groups: 1) those appearing in the "right" set in at least one *qpAdm* model; 2) those appearing in the "left" set in at least one *qpAdm* model; 3) target groups. Targets in the former study were composed of large clusters of West Eurasian individuals, some of them dating back to the Paleolithic (Lazaridis et al. 2022). For that reason, the date distribution for targets in panel **a** is very wide.

The rotating *qpAdm* protocol and model competition are increasingly used as central methods for testing admixture hypotheses proposed after inspecting distributions of individuals in PC spaces, outcomes of *ADMIXTURE* analyses, and *f/D*-statistics indicative of a gene flow rather than a simple tree relationship. Yet, the first study reporting detailed testing of *qpAdm* on simulated data (Harney et al. 2021) was performed in extremely favorable conditions: the simulated AGS history included just two non-nested admixture events; the sources for the principal target group diverged about 1,200 generations ago (almost 35,000 years ago in the case of humans); the proxy sources were strictly cladal with the actual ancestral groups for the target; several groups differentially related to these ancestry sources were available; the simulated data were free of ascertainment bias since sites were sampled in a random way; one million sites were used for most analyses; and only 50/50% simulated admixture proportions were tested for some analyses. That study confirmed that the method behaves as expected under these ideal conditions and offered some important guidelines on the choice and number of "right" populations for optimal specificity of the method and on model comparison strategies, and also showed that the results are robust to the presence of missing data, imbalanced group sizes, ancient DNA damage, a wide range of genomic block sizes, and to a particular type of SNP ascertainment: selecting sites heterozygous in one individual from a certain population (Patterson et al. 2012). Among challenging historical scenarios, only multifurcation followed by uniform and continuous gene flows in a one-dimensional stepping-stone model composed of six groups was explored (Harney et al. 2021). Meanwhile, false discovery rate (FDR) or false positive rate (FPR) of the method and violations of the topological assumptions of *qpAdm* (**Fig. 1**) remained virtually unexplored. Thus, the method was proven to work in extremely favorable conditions. But what happens in other cases where arguably most of the history of humans and other species fits: a history that is not a nearly perfect tree, and can probably be approximated by complex admixture graphs

280 or stepping-stone landscapes (SSL, see **Box 1**)? This question motivated the current study
281 exploring both simulated AGS and SSL histories and in a companion paper focused on AGS
282 histories (Williams et al. 2024).

283 As detailed above, various versions of *qpAdm* protocols form a core of many recent
284 archaeogenetic studies. These protocols are aimed at finding the simplest feasible *qpAdm*
285 models for target groups, where feasibility is defined by setting a threshold for
286 *qpAdm/qpWave p*-values and by setting plausibility criteria for admixture proportions
287 estimated by *qpAdm*. Finding a feasible two-way or more complex admixture model for a
288 target is often interpreted as solid evidence for gene flow, especially if PCA and *ADMIXTURE*
289 methods confirm the same signal. Thus, *qpAdm* protocols are used in fact as formal tests for
290 admixture, whereas the latter two methods are not formal tests.

291 Relying on general principles, we argue that any high-throughput *qpAdm* protocol on poorly
292 understood complex demographic relationships approximated as admixture graphs and/or
293 stepping-stone landscapes is questionable as a formal test for admixture since the *p*-value
294 threshold allows to *reject*, but not to *accept* models, and it is safer to interpret those models
295 as a certain number of gene flows connecting "left" and "right" sets in any direction, not in
296 the form of proxy sources and admixture proportions for a target. The model feasibility
297 criterion including both *p*-values and admixture proportions estimated with *qpAdm* is a
298 complex construct relying on the topological assumptions outlined in **Fig. 1**. We expect that
299 taking "left" and "right" sets that are not well-separated in time or contemporaneous (**Fig. 2**),
300 and where relationships among groups are poorly understood (which is almost always true
301 for exploratory studies), enriches the system for “left-to-right” gene flows, which in turn leads
302 to frequent rejection of true simple admixture models. Since the behavior of *qpAdm*
303 admixture proportion estimates under diverse demographic scenarios is poorly understood,
304 it is possible that a large fraction of these non-rejected complex models emerges as feasible,
305 resulting in false signals of admixture. Another approach probably leading to the same
306 outcome (enriching the system for “left-to-right” gene flows and other violations of the
307 topological assumptions) is selection of pairs of “right” populations maximizing deviations of
308 statistics $f_4(\text{left}_i, \text{left}_j, \text{right}_i, \text{right}_j)$ from zero since such deviations due to tree-like topologies
309 or “left-to-right” gene flows are impossible to tell apart. This approach is sometimes used for
310 increasing the power of *qpAdm* to discriminate between alternative proxy sources.

However, the extent to which these approaches are problematic remains to be explored, especially in the case of SSL. To address this, we analyze both simulated AGS histories of random topology and fixed complexity and SSL histories with randomized non-uniform gene flow intensities on a landscape of fixed size and test various types of *qpAdm* protocols common in the literature: rotating and non-rotating, with or without temporal stratification of target groups and proxy ancestry sources, and with or without a model competition step. We also reproduced other aspects of a typical archaeogenetic study on AGS simulated data: we combined various *qpAdm* protocols with PCA, f_3-statistics, and an unsupervised *ADMIXTURE* analysis to explore FDR of complex "admixture screening pipelines". In the last section of our study, we analyzed from the SSL perspective a large set of published *qpAdm* models based on real data (Zeng et al. 2023) and interpreted the resulting patterns using insights from the extensive exploration of simulated SSL data.

Results

Since the concepts and analytical approaches are rather different in the cases of AGS and SSL histories, we describe the results on the former class of simulations in the first three sections, and the results on the latter class in the following sections.

AGS histories: approach to simulation and testing of qpAdm protocols

Below we explore on simulated AGS histories performance (mainly FDR) of *qpAdm* protocols representing the spectrum of protocols used in the literature (for an overview of our AGS workflow see **Fig. 3**). The most extreme example is a protocol where all groups are rotated and all are treated alternatively as outgroups, targets, and proxy sources, i.e., there is no temporal stratification between the latter categories. We term this protocol "proximal rotating". Although such an extreme situation is, to our knowledge, rare among published *qpAdm* protocols (see Carlhoff et al. 2021, Oliveira et al. 2022, Zeng et al. 2023; in the latter study both distal and proximal rotating protocols are used in parallel), we use it to illustrate the potential effects of poor temporal stratification of targets and proxy sources in published rotating protocols (**Fig. 2**; see **Suppl. text 1** for a detailed description of the published protocols used for making this figure). Models with targets pre-dating proxy sources are

encountered in high-throughput *qpAdm* screens, but do not constitute a majority of models (Narasimhan et al. 2019, Librado et al. 2021, Bergström et al. 2022, Lazaridis et al. 2022, Taylor et al. 2023, Zeng et al. 2023, Speidel et al. 2024). We also explore FDR of the proximal non-rotating (Harney et al. 2018, van de Loosdrecht et al. 2018, Narasimhan et al. 2019, Prendergast et al. 2019, Sikora et al. 2019, Wang et al. 2020, Carlhoff et al. 2021, Wang et al. 2021, Zhang et al. 2021, Changmai et al. 2022a, Changmai et al. 2022b, Maróti et al. 2022, Brielle et al. 2023, Lee et al. 2023, Nakatsuka et al. 2023), distal rotating (Narasimhan et al. 2019, Librado et al. 2021, Bergström et al. 2022, Lazaridis et al. 2022, Taylor et al. 2023, Zeng et al. 2023, Speidel et al. 2024), and distal non-rotating protocols (Haak et al. 2015, Mathieson et al. 2015, Lazaridis et al. 2016, Antonio et al. 2019, Mathieson et al. 2018, Sikora et al. 2019, Marcus et al. 2020, Yang et al. 2020, Papac et al. 2021, Yaka et al. 2021, Patterson et al. 2022). In the distal protocols, only *qpAdm* models where target group's sampling date is strictly contemporaneous with, or post-dates, sampling of both proxy sources were considered.

We tested performance of these *qpAdm* protocols on relatively complex AGS simulated histories: 13 populations connected with each other by admixture graphs of random topology, including 10 pulse-like admixture events. Ten diploid individuals with no missing data were sampled from each population at "leaves" of the graph. Forty such random topologies were simulated. These simulations generated sets of populations sampled at widely different dates in the past or, in other words, located at various genetic distances from the root.

Further details on the simulated population histories are presented in Methods and illustrated by 14 examples in **Fig. S1**. To explore the influence of data amount and population divergence levels on *qpAdm* performance, we generated three independent sets of ten simulation replicates for each graph topology (referred to as setups no. 1, 2, and 3 below) and single simulation replicates for two further setups (no. 4 and 5; see **Fig. 3**). In these simulations, genome size was 300, 1,000, or 3,000 Mbp; maximal history depth was 800 or 3,000 generations (ca. 23,000 or 87,000 yBP for humans), with all dates scaled up proportionally; populations were sampled at different dates in the past or all terminal branches were extended to the "present" and sampled at that point. Not all combinations of these settings were tested, see **Fig. 3** for details and **Fig. S2** for the number of SNP loci polymorphic in the simulated datasets. The simulations with the 3,000-generation depth

resulted in median F_{ST} per simulated topology at the inter-continental levels of human genetic differentiation (~0.1), and those with the 800-generation depth matched intra-continental levels (median F_{ST} from 0.022 to 0.051, see **Fig. S3**).

A typical archaeogenetic dataset is composed of pseudo-haploid data with high proportions of missing sites and with widely different group sizes, including singleton groups. To generate more realistic "noisy" data, we also performed randomized subsampling of SNP datasets for simulation setups no. 1, 2, and 3 (300-, 1,000-, and 3,000-Mbp-sized genomes, respectively), for one simulation replicate per each setup (**Fig. 3**, see Methods for details). The resulting data were pseudo-haploid, had missing data rates ranging from 5% to 95% across individuals, and had uneven group sizes ranging from 1 to 10 individuals. Ten independent subsampled datasets were generated for each simulated topology (for counts of polymorphic loci see **Fig. S2**).

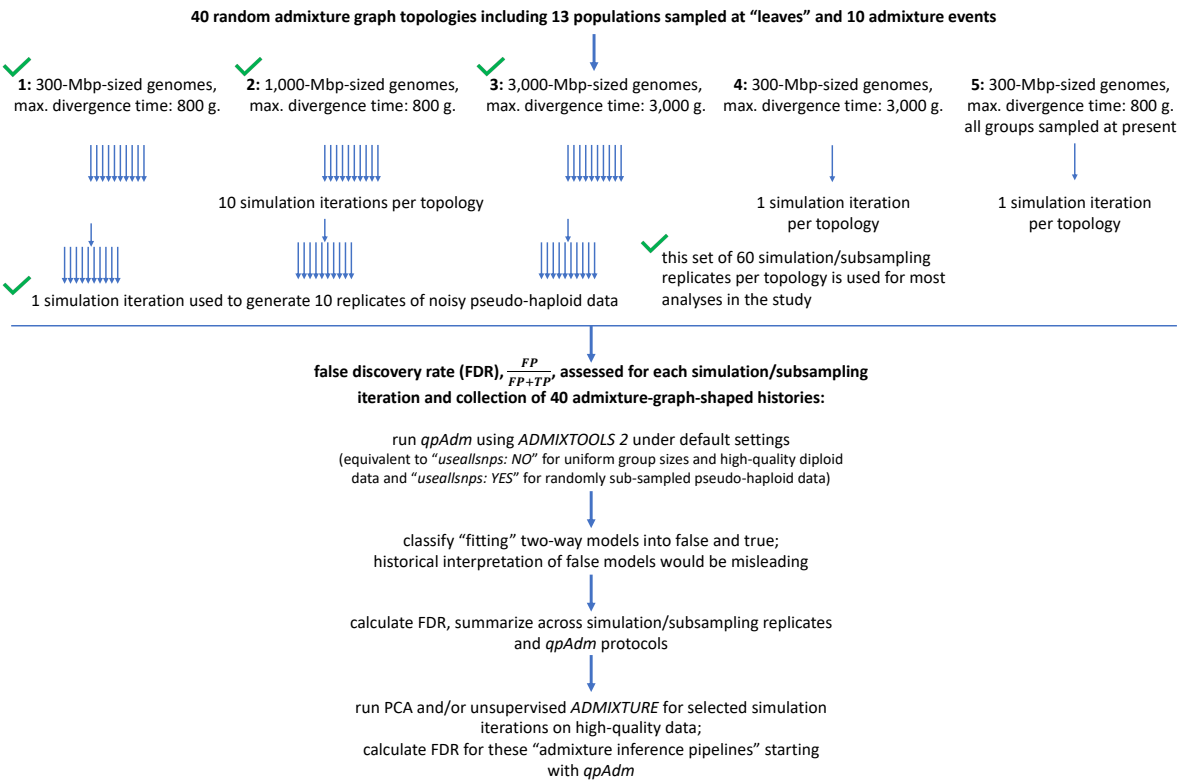


Figure 3. A diagram illustrating our analytical workflow in the case of AGS simulations. The simulation setups are numbered in bold, and these numbers are referred to in the text.

The *qpAdm* protocols we applied to the simulated data were focused on the simplest models: one- and two-way admixture models. The median number of two-source admixture events

390 in a population from our AGS simulations is 1 (average =1.17), suggesting that two-way
391 models are broadly adequate for this type of history, especially considering that by far not all
392 internal branches in AGS histories were sampled (**Fig. S1**). Histories that are more complex
393 than two-way mixtures were also not rare in the data, accounting for 38% of all models tested
394 (**Table S1a**). The model feasibility criterion followed Narasimhan *et al*. (2019), see **Box 1** for a
395 definition. Thus, we tested all possible two-way admixture models for 40 complex population
396 histories (34,320 proximal and distal rotating models per simulation setup and replicate).

397 The non-rotating *qpAdm* approach was implemented as follows: for each simulated graph six
398 most ancient groups were selected as a fixed "right" set (ties were resolved in alphabetical
399 order; these "right" sets remained unchanged for a given graph topology across independent
400 simulations) and for the remaining seven groups all possible one-way and two-way admixture
401 models were tested, applying the same composite feasibility criterion that was used for the
402 rotating protocol (4,200 proximal and distal models per simulation setup and replicate). Thus,
403 models tested with the non-rotating protocol represent a subset of models tested with the
404 rotating protocol.

405

406 ***AGS histories: case studies illustrating false and true feasible qpAdm models***

407 *False positive models*

408 In the context of complex and random admixture graph topologies it is hard to draw a strict
409 boundary between true and false admixture models composed of a target and two proxy
410 sources. However, we classified the feasible *qpAdm* models into false and true ones relying
411 on a set of rules. By far the most common class of false feasible *qpAdm* models (referred to
412 as "false positive" or FP models), comprising 50.9% of all FP models generated by the
413 proximal rotating protocol across all the simulation and subsampling replicates (for setups
414 no. 1 and 2), occurs when the target group is falsely rejected as forming a clade with one or
415 both proxy sources whilst they are simulated as clades. In other words, these are false
416 rejections of single-source admixture models.

417 An example of the situation whereby the target group is falsely rejected as forming a clade is
418 shown in **Fig. 4** where the clade relationship between the target (*J*) and source (*M*) is rejected
419 by the rotating protocol due "left-to-right" gene flows violating the topological assumptions

of *qpAdm*, and more complex models ("*J = A + M*" and "*J = L + M*") are evaluated as true. When a true one-way model "*J = M*" is tested, *M* represents a proxy for the only true source of ancestry in *J*, and outgroups *A*, *C*, *D*, and *L* are partially derived from the target lineage, resulting in rejection of the one-way model with *p*-values from ~10^{-32} to ~10^{-50} (in the case of 1000-Mbp-sized genomes and high-quality data). Rejections of two-way models "*J = A + M*" and "*J = L + M*" according to *p*-values are rare (**Fig. 4c**), and nearly all estimated ancestry proportions ±2 SE are between 0 and 1 even on low-quality pseudo-haploid data (**Fig. 4d**), that is why these models are classified as fitting in nearly all cases (in 34 and 36 of 40 simulation/subsampling replicates, respectively; setups no. 1 and 2 were considered). Groups *A* and *L* are 270 generations younger than *J*, and >50% of their ancestry was simulated as derived from the *J* lineage (**Fig. 4a**). These FP models illustrate not only cladality rejection, but also incorrect gene flow direction suggested by the *qpAdm* results.

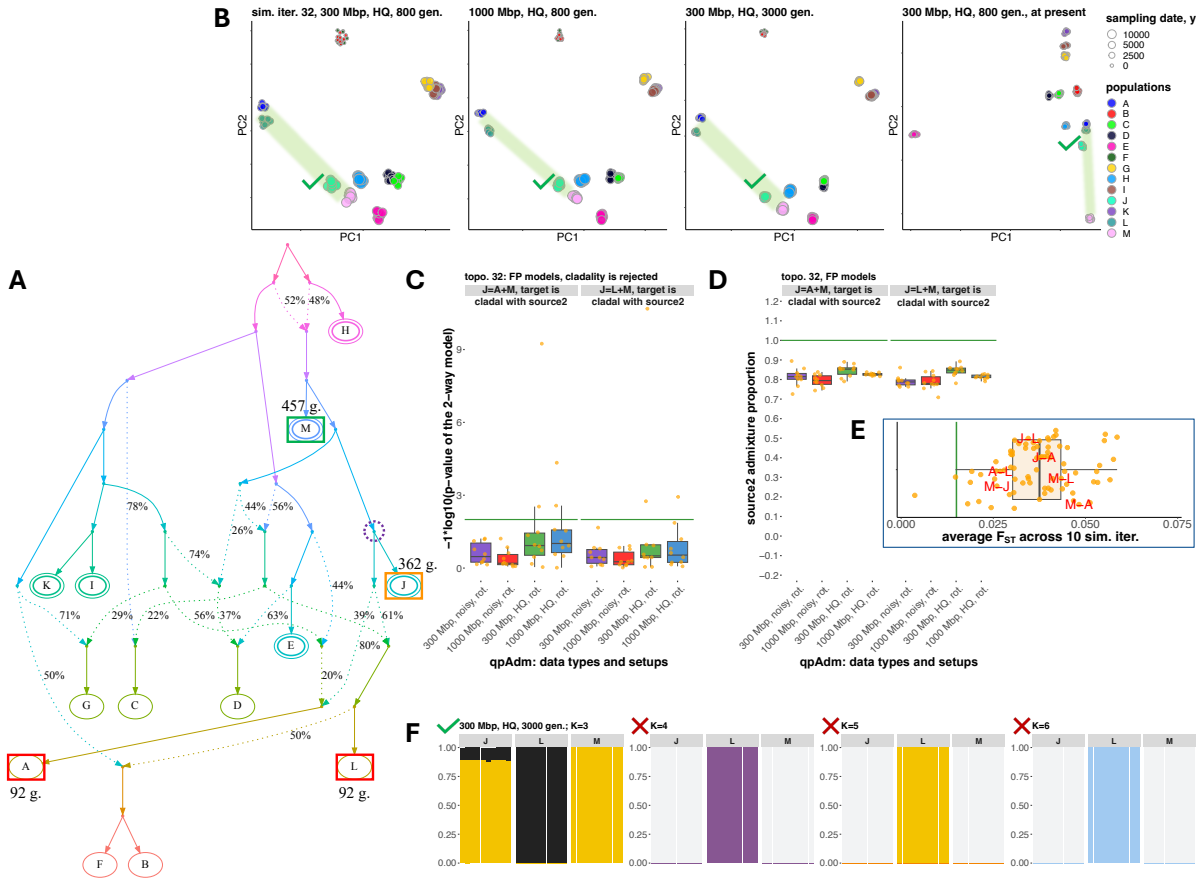


Figure 4. A case study illustrating the most common class of false positive *qpAdm* models supported by the proximal rotating protocol (accounts for 50.9% of all FP models across all the simulation and subsampling replicates in **Fig. 6** and **Suppl. dataset 1**). Models of this type include at least one proxy ancestry source that is simulated as fully cladal with the target. The other proxy source may be simulated as a descendant of the target lineage (as shown here), may belong to an unrelated lineage (**Fig. S4a**), or may be also cladal with the target. Both models shown here, "*J = A + M*" and "*J = L + M*",

are also fully supported by three-dimensional PCA and by an unsupervised *ADMIXTURE* analysis at one or more *K* values. **A.** Simulated AGS history (only topology is shown here, for divergence/admixture dates and effective population sizes see the respective simulated history shown in **Fig. S1**). Sampled populations are marked by letters. The target population from the *qpAdm* models illustrated here is enclosed in an orange rectangle; correct proxy source – in a green rectangle, and inappropriate proxy sources – in red rectangles. Sampling dates for these groups (in generations before present) are shown beside the rectangles (the dates are from the "shallow" simulations up to 800 generations deep). True simulated ancestry source(s) for the target are enclosed in dashed violet circles. Six populations used as outgroups for the non-rotating *qpAdm* protocol are enclosed in double ovals. **B.** Two-dimensional PCA plots for four simulated high-quality datasets indicated in plot titles. Simulated populations are colored according to the legend on the right, and larger points correspond to older sampling dates. The space between the proxy sources is shaded in green. If admixture model(s) are supported by three-dimensional (sic!) PCA (PC1 vs. PC2 vs. PC3) according to the criteria listed in Methods, a green tick mark is placed beside the target group in the plot, and a red cross mark is placed otherwise. **C.** Boxplots summarizing *p*-values of two-way *qpAdm* models (indicated in plot titles) across simulation or subsampling replicates, grouped by simulation setups and *qpAdm* protocols as indicated on the x-axis. Green lines show the *p*-value threshold of 0.01 used in this study for model rejection. *P*-values for one-way models are not shown for brevity. The *qpAdm* models shown here were not tested using the non-rotating protocol since the target falls into the set of six "right" populations chosen for that protocol. **D.** Boxplots summarizing EAF across simulation or subsampling replicates, grouped by simulation setups and *qpAdm* protocols. The proportions are shown either for the first or second proxy source, as indicated on the y-axis; green lines show the simulated admixture proportion. **E.** All F_{ST} values for population pairs from the simulated history. Average F_{ST} values are shown across 10 simulation replicates for 1000-Mbp-sized genomes and high-quality data. Population pairs formed by components of the illustrated two-way *qpAdm* model(s) are labelled in red. The green line shows median F_{ST} (0.015) for the Bronze Age West Asian populations analyzed by Lazaridis et al. (2016) as an example of human genetic diversity at the sub-continental scale. **F.** Ancestry proportions estimated with unsupervised *ADMIXTURE* for the groups constituting the *qpAdm* model(s). For brevity, results are shown for one or two *qpAdm* models, for one simulated dataset indicated in the plot title, and for four selected *K* values. If admixture model(s) are supported by this analysis at a given *K* value, a green tick mark is placed beside the target group in the plot, and a red cross mark is placed otherwise.

These results illustrate risks of applying rotating *qpAdm* protocols to sets including ancient groups "on the left" and much younger groups "on the right" (Carlhoff et al. 2021, Oliveira et al. 2022), since those may be descended partially from the "left" groups (**Fig. 1c, d, e**) and thus lead to *qpAdm* assumption violations and false positive findings of gene flow. PCA on four datasets (**Fig. 4b**) and unsupervised *ADMIXTURE* results at some *K* values (**Fig. 4f**) support these FP models. Thus, a researcher co-analyzing ancient and modern groups due to lack of ancient samples from certain regions may encounter this situation in practice and, given support by all methods in the standard archaeogenetic toolkit (*qpAdm*, PCA, and *ADMIXTURE*), is likely to conclude that FP models are true. We illustrate another similar situation often leading to rejection of cladality and acceptance of two-way or even more

482 complex models (in 33-37 of 40 simulation/subsampling replicates, depending on model and
483 *qpAdm* protocol; setups no. 1 and 2 were considered) in **Fig. S4a**.

484 Other topological classes of FP models can be concisely described as follows (and are more
485 precisely defined in Methods): 1) gene flow goes from the target lineage (after the last
486 admixture event in its history) to a proxy source lineage (**Fig. 4, Fig. S4a,b**); 2) a proxy source
487 included in the model is symmetrically related to all real sources of ancestry in the target (**Fig.**
488 **S4c,d**), or both proxy sources represent the same true source and are symmetrically related
489 to all the other true sources (**Fig. S4e**); 3) both proxy sources represent distinct real sources,
490 however a proxy source is a heavily admixed group sharing less than 40% of ancestry with the
491 real source (**Fig. S4a,d**); 4) the target gets a gene flow from a deep-branching source not
492 represented by any sampled population, and an inappropriate proxy source is included in the
493 fitting model (**Fig. S4f,g**). Violations of the topological assumptions of *qpAdm* encountered in
494 these examples of FP model classes and corresponding *qpAdm*, PCA, and *ADMIXTURE* results
495 are described in **Suppl. text 2**.

496

497 *True positive models*

498 Above and in **Suppl. text 2** we have discussed several examples of FP models that in practice
499 would result in misleading conclusions about population history, and now we turn to feasible
500 models that were interpreted as true in our analysis. We rely on a relaxed definition of true
501 models: two-way admixture models for targets with population history best approximated
502 with three-way and more complex models were considered as true positives if they included
503 at least two source proxies *not* satisfying the topological criteria of false models listed above.
504 We also note that TP models were not restricted to those including most optimal source
505 proxies, if the models do *not* satisfy the false positivity criteria. Stricter definitions of TP
506 models are possible (such as those in the companion paper Williams et al. 2024), but we
507 applied this lenient approach since in the archaeogenetic literature sampling of ancient
508 individuals is often poor, and thus the task is not to find all groups which participated in the
509 history of a given target group, but to identify at least predominant ancestry sources with
510 some degree of precision.

511 On a random sample of 400 two-way admixture models from our 40 simulated histories, the
512 fraction of models that were classified as appropriate (true) according to the rules described
513 above was 17.7%. Since groups that are truly admixed are common in our simulations, we do
514 not expect to encounter a "needle in a haystack" situation where finding true admixture
515 models is exceedingly hard. Examples of simulations yielding TP *qpAdm* models are presented
516 in **Suppl. text 2**.

517

518 ***AGS histories: influence of temporal stratification and data amount/quality on the***
519 ***performance of qpAdm protocols***

520 To make a qualitative overview of all the *qpAdm* results on AGS histories and to gain intuition
521 about factors influencing frequency of false results, we placed models in the coordinates of
522 admixture fractions estimated by *qpAdm* (abbreviated as EAF) and *p*-values (**Figs. 5** and **S5**).
523 Both high- and low-quality data generated by the most realistic simulation setup were
524 considered (3,000-Mbp-sized genomes, maximal population divergence depth at 3,000
525 generations). Since *qpAdm* EAF vary from <<–1 to >>1 and to make this approach scalable to
526 models more complex than two-way, we use one number instead of several EAF per an *n*-
527 way model: maximal absolute difference between EAF in that model and equal fractions $\frac{1}{n}$,
528 abbreviated as max|*EAF–EF*| (**Box 1**). Each two-way model was tested many times: on
529 different simulation setups, data qualities, replicates and *qpAdm* protocols. We explore the
530 max|*EAF–EF*| vs. *p*-value spaces for three model classes separately: models that were
531 consistently non-fitting in all our analyses, models that were fitting at least once and classified
532 as false, and models that were fitting at least once and classified as true (**Figs. 5, S5**). For
533 comparison, we also visualized models populating the space of EAF standard errors (SE)
534 reported by *qpAdm* and *p*-values (**Fig. S6**).

535 Remarkably, the models consistently rejected (not only on the simulations shown in **Fig. 5,**
536 but throughout this study) demonstrated EAF outside the (0, 1) interval in nearly all cases,
537 especially on high-quality data, while models that were fitting at least once and classified as
538 either false or true are characterized by EAF largely in the (0, 1) interval, but their *p*-values
539 vary along the whole scale (**Fig. 5**). Of the consistently rejected models, 40% are models
540 whose target populations have 0 admixture events in their history (**Table S1a**) and as

discussed elsewhere we did not classify consistently negative models with admixed targets into true and false.

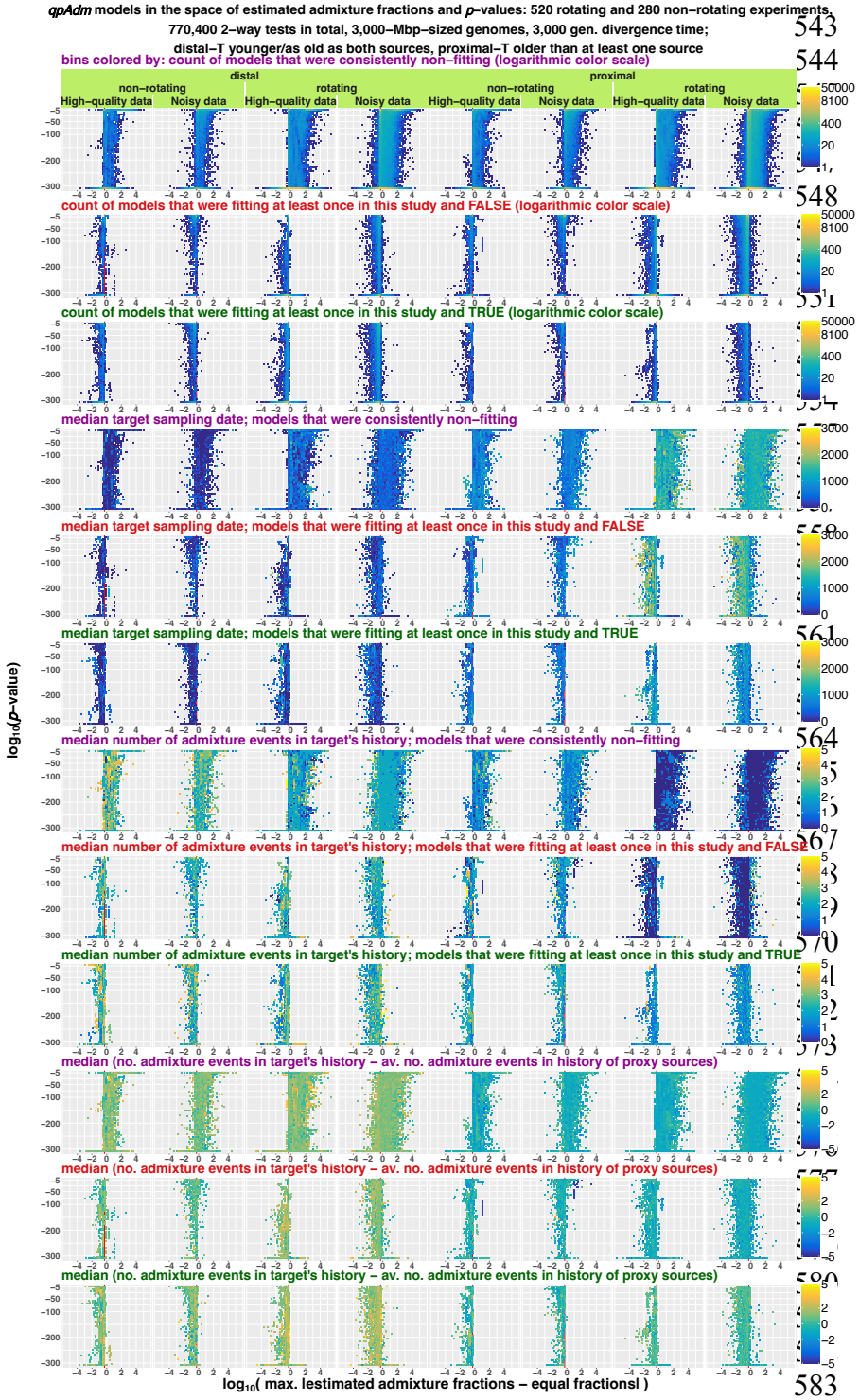


Figure 5. All two-way *qpAdm* models tested on AGS simulations are visualized in the space of max|*EAF*–*EF*| and *p*-values (both axes are logarithmic). Results are shown for the 3,000-Mbp-sized genomes only and stratified in columns by: model subsets (distal or proximal), *qpAdm* protocols, and data quality. Results are also stratified by three model classes: consistently non-fitting (plot rows labelled in purple), fitting at least once in this study and evaluated as false (in red), and fitting at least once in this study and evaluated as true (in green). The whole space is shown: *p*-values from 10^{-306} to 1, and for *p*-values from 3.16×10^{-6} to 1 see **Fig. S5a** (for plots based on fitting models only see **Fig. S5b**). The space is divided into rectangular bins (50×50), and they are colored by density of individual *qpAdm* models populating this space (logarithmic color scale) or by median values of metrics informative for understanding method performance: target sampling date, number of admixture events on all paths from the target to the root of an AGS simulation (number of admixture events in target's history), and difference between number of admixture events in target's history and average number of admixture events in the history of both proxy sources. The red vertical lines (or the red tick marks) mark max|*EAF*–*EF*| = 0.5.

590 This qualitative overview suggests that EAF are much less variable than *p*-values, and we
591 quantify this using standard deviations (SD) of $\log_{10}$(*p*-value) or EAF across 10 (or 20, if a
592 model was evaluated under both rotating and non-rotating protocols)
593 simulation/subsampling replicates (**Fig. S7**). *P*-values of false, true, and consistently negative
594 models vary by many orders of magnitude between replicates, and SD grows with increasing
595 amount of data, up to ca. 30-60 orders of magnitude in the case of noisy data derived from
596 3,000-Mbp-sized genomes (**Fig. S7**). In contrast, median EAF SDs are consistently between
597 0.01 and 0.1 for models that were fitting at least once in this study (with very small differences
598 between false and true models), and between 0.1 and 1 for consistently rejected models (**Fig.**
599 **S7**). This analysis confirms that there are two classes of two-way *qpAdm* models on AGS data
600 that encompass a great majority of all models: those with relatively stable EAF far from (0, 1)
601 and those with stable EAF within (0, 1). The latter models are universally fitting very rarely
602 (**Fig. S8**) due to the high variability of *qpAdm p*-values.

603 It is important to understand if some of these three model classes are enriched for certain
604 features of target's and proxy sources' admixture histories, and to understand how this is
605 shaped by the choice of *qpAdm* protocol and by temporal stratification of targets and proxy
606 sources. According to our definitions, proximal models, and especially proximal models
607 tested with the rotating protocol, are enriched for targets sampled deep in the past (see **Fig.**
608 **5** for a mapping on the max|*EAF*–*EF*| vs. *p*-value space and **Table S1** for a quantitative
609 summary and statistical tests). As admixture events were simulated at random points in time
610 and placed randomly on the graphs, these models are actually enriched for unadmixed
611 targets (**Fig. 5**): the largest class of models are consistently non-fitting models with
612 unadmixed targets: ca. 33% of all models tested belong to this class, and ca. 83% of those are
613 proximal rotating models (**Table S1a**). These models are true negatives (however, this list of
614 true negatives is not exhaustive). In contrast, proximal non-rotating and distal rotating/non-
615 rotating models demonstrate unimodal, not monotonically decaying, distributions of target
616 history complexity, and these distributions are similar in the consistently rejected, false and
617 true model classes, especially in the former two classes (**Table S1a**).

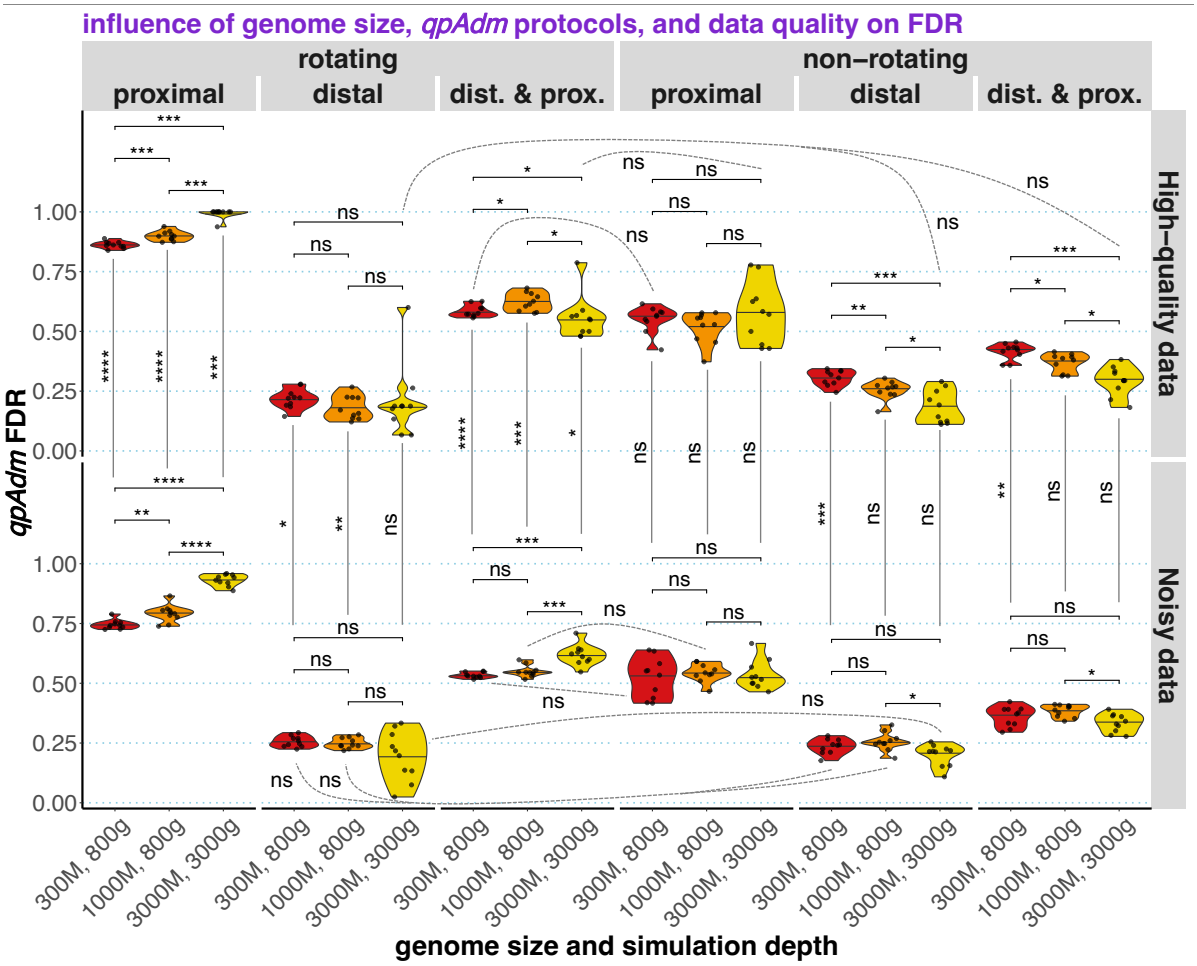
618 For the same fundamental reasons, proximal models that were consistently non-fitting or
619 occasionally fitting and false are characterized by fewer admixture events in target's history
620 as compared to average number of admixture events in sources' histories (median difference

621 is –1 in most cases), but this is not the case for proximal models that were occasionally fitting
622 and true (median difference is around 0, see **Table S1b**). This distinction between true
623 proximal models and the other two model classes is highly significant for both the rotating
624 and non-rotating protocols (**Table S1b**). The much smaller difference in this metric between
625 proximal models that were consistently non-fitting and occasionally fitting and false is
626 significant too (**Table S1b**). Statistically significant trends in the same direction are observed
627 for distal models too, but in this case median differences between number of admixture
628 events in target's history and average number of admixture events in sources' histories
629 (abbreviated as AEH_T – *av.* AEH_S) are positive (**Table S1b**). The chance of encountering
630 violations of the *qpAdm* assumptions (**Fig. 1**) grows when admixture history of proxy sources
631 is more complex than that of a target group, explaining these results.

632 Considering all models on the 3,000-Mbp-sized simulations, AEH_T – *av.* AEH_S is just weakly
633 correlated to max|*EAF*–*EF*| or *p*-values, which is also visible in **Fig. 5**: Spearman's correlation
634 coefficients for various *qpAdm* protocols and data qualities range from -0.49 to 0.3 in five
635 practically relevant sections of the max|*EAF*–*EF*| vs. *p*-value space (**Table S1c**). Thus, although
636 false/consistently non-fitting and true models differ significantly according to the AEH_T – *av.*
637 AEH_S metric, using *p*-values or EAF for predicting which model is true and which is false is not
638 a fruitful approach. For example, if no condition is imposed on EAF and models with just the
639 smallest *p*-values ($<10^{-55}$) are removed, Spearman's ρ is close to 0 for both max|*EAF*–*EF*| and
640 *p*-values, –0.19 at most (**Table S1c**). If *p*-value $\geq 10^{-55}$ and max|*EAF*–*EF*| ≤ 0.5 (which is
641 equivalent to EAF $\in$ [0, 1]), in the case of the distal rotating protocol on both high- and low-
642 quality data, Spearman's $\rho \approx -0.4$ (this value is significant). Thus, higher *p*-values correspond
643 to smaller AEH_T – *av.* AEH_S, thus actually favoring false models, albeit weakly.

644 While FP *qpAdm* models are expected for complex genetic histories, the practical usage of
645 *qpAdm* protocols relies on an assumption that false positives are rare. As discussed above,
646 due a fundamental property of AGS histories where complexity (number of admixture events)
647 grows with time, *qpAdm* false discovery rate (FDR, see **Box 1**) is expected to depend on
648 enrichment of the sets of tested models for those where proxy source history is more
649 complex than that of the target. First, we calculated FDR on the complete set of simulated
650 AGS histories, per each simulation/subsampling replicate (**Fig. 6**). We considered different
651 amounts of data (depending on genome sizes and simulated population divergence),

652 different data quality, rotating and non-rotating protocols, distal or proximal protocols
653 (considering both distal and proximal models, **Box 1**), and protocols considering proximal
654 models only.



655

656 **Figure 6.** Distributions of FDR values over 10 simulation replicates (for high-quality data) or over 10
657 subsampling replicates derived from a single simulation replicate (for low-quality data). The
658 distributions are summarized as violin plots with medians and stratified by *qpAdm* protocols (rotating
659 and non-rotating), by model subsets (proximal, distal, or both model types included), by data quality,
660 and by three combinations of simulated genome sizes (300, 1,000, or 3,000 Mbp) and maximal
661 simulation depths (800 or 3,000 generations). The following statistical comparisons were performed
662 with the two-sided Wilcoxon test: across genome sizes, with protocol and data quality fixed (36 non-
663 paired tests); across data qualities, with protocol and genome size fixed (18 non-paired tests); and
664 across protocols, with genome size and data quality fixed (90 paired tests). Paired tests were used in
665 the latter case since the corresponding sets of models are not independent: non-rotating models form
666 a subset of rotating models, and distal and proximal models can be combined. *P*-values (adjusted for
667 multiple testing using the Holm method) are coded in the following way: ns, >0.05; *, ≤0.05; **, ≤0.01;
668 ***, ≤0.001; ****, ≤0.0001. Most comparisons across protocols are significant at the 0.05 level
669 (omitted for clarity), and only non-significant comparisons of this type are shown with dashed lines.

670

671 Indeed, considering distal models only (tested with both rotating and non-rotating protocols)
672 results in the lowest FDRs down to 2.4%, but even in this case median FDR values across
673 replicates fluctuated from 16.4% to 31.2% (**Fig. 6**). FDRs based on proximal models tested
674 with the rotating protocol varied from 72.5% to 100% and grew significantly with increasing
675 amount of data. A large fraction of false proximal models emerges as fitting due to false
676 rejections of one-way models because of violations of the topological assumptions (**Fig. 1**),
677 for instance, due to "left-to-right" gene flows, and that prompts the investigator to test more
678 complex two-way models, which are often fitting (**Suppl. text 2**). And model rejection is more
679 efficient when more data is available, explaining the effect on FDR observed here.

680 FDRs based on proximal models tested with the non-rotating protocol were lower (medians
681 from 52.1% to 57.7%) and did not depend on the amount of data, suggesting that false model
682 rejections due to assumption violations play a less important role here, which is expected for
683 "right" and "left" population sets which are stratified chronologically. Similarly, the amount
684 of data had no statistically significant influence on FDR when distal models were considered,
685 except for the non-rotating protocol on high-quality data (increasing the amount of data
686 reduced FDR, **Fig. 6**). We also note that random subsampling of SNPs and individuals and a
687 less robust algorithm for calculating f_4-statistics on the low-quality data (with different
688 statistics calculated on different SNP sets) did not lead to dramatic increases/decreases in
689 FDR (**Fig. 6**). The observation that the decrease in FDR with increasing amount of data is small
690 and often non-significant is possibly explained in the following way. Positive models are
691 defined based on both *p*-values and EAF (**Box 1**): while EAF are relatively stable (**Fig. S7**), *p*-
692 values of individual models decrease exponentially with the amount of data (**Fig. S4**).
693 However, model rejection is not always "beneficial" for protocol's performance assessed via
694 FDR since both models with inappropriate proxy sources (false models) and true models with
695 violations of the topological assumptions (**Fig. 1**) are rejected more often when more data is
696 added.

697 In summary, even in optimal conditions (simulation depth of 87000 years on human
698 timescale, 3,000-Mbp-sized genomes, high-quality diploid data) we did not achieve negligible
699 FDR when distal models were tested on a collection of random AGS histories, and high data
700 amount/quality makes FDR based on proximal models grow. In nearly all conditions tested
701 there is very high variability in FDR across simulated AGS topologies, with many topologies

702 demonstrating FDR of either 0% or 100% (**Fig. S9**), and hence it is hard to predict if real genetic
703 histories are prone to generating FP signals of admixture when *qpAdm* protocols are applied.
704 For an illustration of three topologies most problematic for distal *qpAdm* protocols see **Fig.**
705 **S10**. These differences in *qpAdm* performance across topologies are probably driven at least
706 in part by the fact that some random AGS histories have much "denser" violations of the
707 *qpAdm* topological assumptions than others, but this frequency of assumption violations was
708 not quantified in our study since it requires another layer of manual topological analysis.

709 We estimated FDR instead of false positive rate due to a technical limitation: the process of
710 model classification into true and false ones is difficult to automate since it requires careful
711 interpretation of both the simulated topology and admixture proportions. Therefore,
712 classifying all possible 34,320 two-way admixture models (858 per simulated topology) into
713 true and false was hardly feasible. We estimated false positive (FPR) and true positive rates
714 (TPR) only approximately, relying on the fractions of negatives and positives in random
715 samples of two-way admixture models (see below). FPR in our context is the probability that
716 a two-way model with an unadmixed target and/or inappropriate proxy source(s) emerges as
717 fitting in a single *qpAdm* test. TPR in our context is the probability that a two-way model with
718 an admixed target and both proxy sources being appropriate emerges as fitting in a *qpAdm*
719 test.

720 The fraction of two-way admixture models that are inappropriate according to our topological
721 criteria (the fraction of negatives) in a random sample of 400 proximal and distal rotating
722 models from all the simulated AGS histories is 82.3%, which allows us to approximately
723 estimate not only FDR (48%–78.7%), but the FPR of the proximal rotating protocol = number
724 of false positives / number of negatives per simulation replicate [858 models × 40 graphs ×
725 fraction of negatives] = 0.04%–2.6% across simulation parameter combinations and
726 replicates summarized in **Fig. 6**. The TPR of the proximal rotating protocol = number of true
727 positives / number of positives per simulation replicate [858 models × 40 graphs × (1 –
728 fraction of negatives)] = 0.2%–10.1%. Here are the same statistics for the distal non-rotating
729 protocol: the fraction of negatives in a random sample of 400 distal non-rotating models from
730 all simulated graphs, 74.3%; total number of distal non-rotating two-way models across all
731 graph topologies, 1804; FDR across simulation parameter combinations and replicates from
732 **Fig. 6**, 10.9%–34.4%; FPR, 0.2%–3.2%; TPR, 3.5%–24.4%. Thus, although FPR of a single

733 *qpAdm* test applied to a random AGS history is low, due to the high proportion of negatives
734 among all models (**Figs. 5, S8**) and the low TPR, FDR becomes high, compromising historical
735 interpretation of such screens for admixture.

736 As we show in **Suppl. text 3** (**Fig. S11, Tables S2** and **S3**), it is impossible to decrease FDR of
737 the proximal rotating protocol dramatically by combining it with other methods (PCA and
738 unsupervised *ADMIXTURE*) or by adjusting *p*-value thresholds. However, it is possible to
739 decrease FDR of distal *qpAdm* protocols to ca. 0% if positive *qpAdm* results are further
740 validated by PCA and the unsupervised *ADMIXTURE* (Alexander et al. 2009) methods (**Suppl.**
741 **text 4**, **Table S4**). In that supplemental section we also explore the performance of proximal
742 non-rotating *qpAdm* protocols combined with model competition (Narasimhan et al. 2019)
743 (**Suppl. text 4**, **Table S5**).

744

745 ***Introducing SSL histories***

746 Since effects of prolonged gene flows in the context of a simple AGS history were studied by
747 Harney *et al*. (2021), here we explore a more extreme case also investigated in that study
748 briefly, namely stepping-stone landscapes (SSL) where gene flows between the nearest
749 neighbors are continuous and affect the whole landscape (see **Box 1** for a definition). SSL
750 composed of panmictic demes approximate isolation-by-distance landscapes composed of
751 individuals and small exogamous groups. We simulated SSL that are more realistic than
752 uniform gene-flow scenarios and/or one-dimensional strings of demes (Novembre et al.
753 2008, Duforet-Frebourg and Slatkin 2016, Harney et al. 2021, Tournebize and Chikhi 2023),
754 namely two-dimensional landscapes with gene-flow intensities that are non-uniform both in
755 space and time (Petkova et al. 2016, Al-Asadi et al. 2019). Approximately circular landscapes
756 composed of 64 panmictic demes of constant effective size (1,000 diploid individuals) were
757 simulated as arising via multifurcation ("star radiation") and then evolving for ca. 2,500
758 generations with non-uniform bidirectional gene-flows (**Fig. 7**).

759 Four types of SSL were simulated, characterized by varying ranges of gene-flow intensities,
760 with 10 simulation/gene-flow intensity replicates per range. Each intensity range corresponds
761 to gene-flow between two neighbouring demes in one direction per generation. An intensity
762 value means that a given fraction of effective population size in a generation comes from

migrants from a specified neighbouring deme. In all the cases, gene-flow intensities were sampled at the start of each "gene-flow epoch" from uniform distributions, except for the ca. 10^{-5} to 10^{-2} range which were sampled from a normal distribution centred at 0 (**Table S6**). The four gene-flow intensity ranges between two neighbouring demes in one direction are as follows: (ca. 10^{-5} to 10^{-2}), (ca. 10^{-5} to 10^{-4}), (ca. 10^{-4} to 10^{-3}), and (ca. 10^{-3} to 10^{-2}) per generation (**Table S6, Fig. S13b,c**).

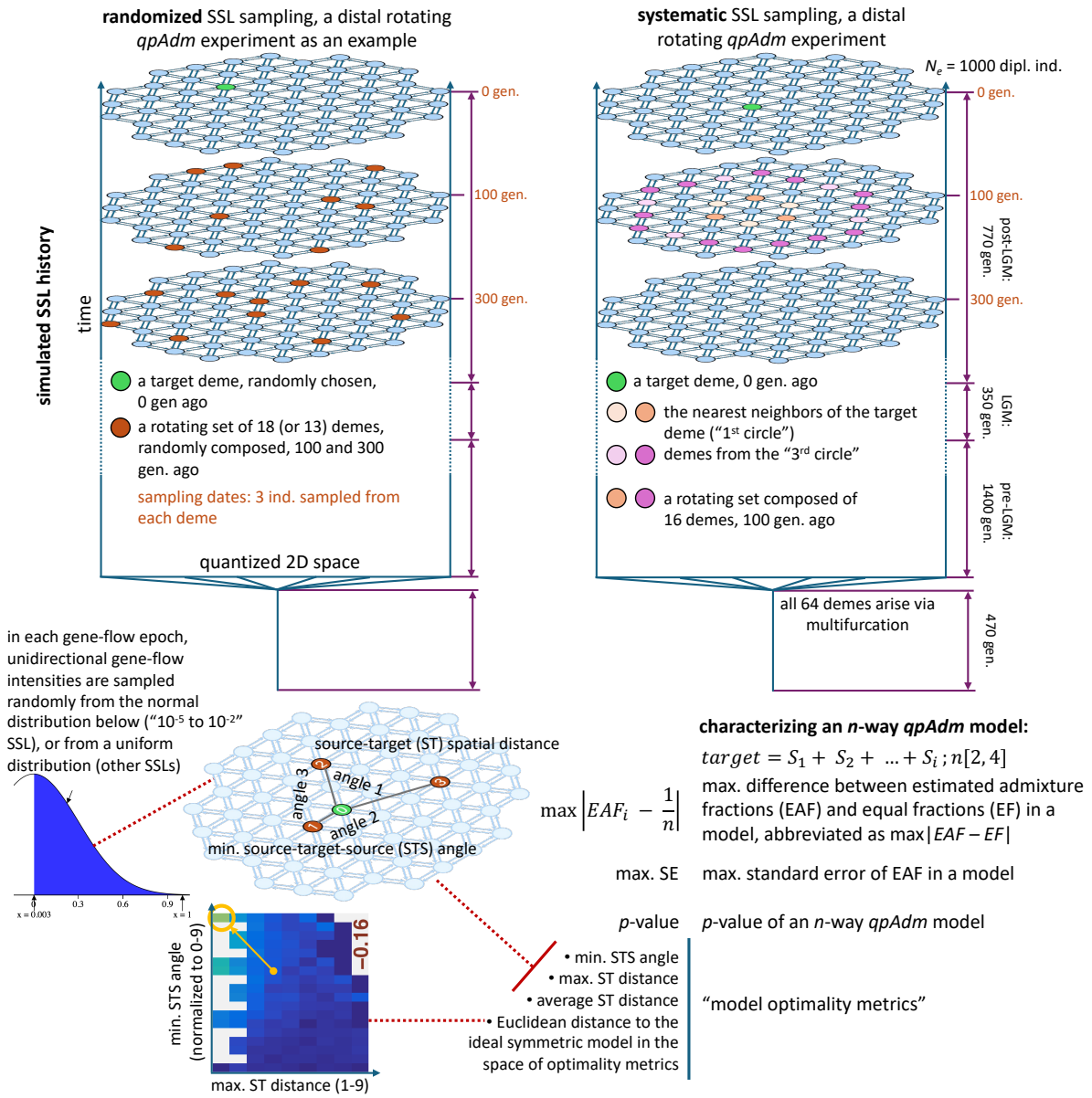


Figure 7. Diagrams summarizing principles of SSL simulations, SSL-based *qpAdm* setups, and approaches we used for interpreting *qpAdm* results in this context. Key features of the simulations are as follows: 1) origin of all the demes via a multifurcation; 2) the 2D space is quantized into 64 demes of small and stable effective population size; 3) gene-flow intensities between neighboring demes are generated by random sampling from normal or uniform distributions (in certain intervals), at the beginning of "gene-flow epochs"; 4) at least three gene-flow epochs were simulated ("pre-LGM", "LGM", and "post-LGM"); 5) all the demes are sampled at three time points ("slices" of the

landscape) in the second half of the "post-LGM" epoch. *qpAdm* experiments (rotating and non-rotating, distal and proximal) were constructed by subsampling this initial set of deme samples randomly or systematically, as illustrated by the examples on the left and the right, respectively. Each *qpAdm* model (two- to four-way) was then characterized by 8 numbers: 3 of those were derived from *qpAdm* outputs, and 5 were properties of the model in the context of the landscape. The former metrics are as follows: max|*EAF*–*EF*|, max. SE, and *p*-value. The latter numbers are termed "model optimality metrics": min. STS angle, max. ST distance, average ST distance, Euclidean distance to the ideal symmetric model (marked by the orange circle) in the space of model optimality metrics (illustrated in the lower left corner of the figure) based on min. STS angle and max. or average ST distance. For simplicity, interpretation of *qpAdm* results at the level of experiments is not covered in these diagrams.

In the case of *f*-statistics, simulating $n \times m$ demes and sampling m of them evenly is equivalent to simulating m demes, but with proportionally lower gene-flow intensities. Since the size of the landscapes (64 demes) was the same across the gene-flow intensity brackets, these landscape types, judging by the F_{ST} distributions (**Fig. S13a**), can be thought of as representing regions of different size: from very sparse global sampling (10^{-5} to 10^{-4}, median F_{ST} = 0.41, which is close to the highest F_{ST} between African and non-African humans) to dense sampling in a sub-continental region (10^{-3} to 10^{-2}, median F_{ST} = 0.02, corresponding, e.g., to the Bronze Age West Asian populations analyzed by Lazaridis *et al*. 2016). Thus, by varying gene-flow intensity (**Fig. S13b,c**) we explored different levels of "'landscape sparsity": a landscape property that, as we show below, is probably the single most important predictor of *qpAdm* performance on SSL. The "10^{-5} to 10^{-2}" type of landscapes, enriched for low-intensity gene flows due to sampling from the non-uniform distribution, but also including high-intensity flows (**Fig. S13b,c**), is probably closer to real-world *qpAdm* protocols (**Suppl. text 1**) where worldwide "right" groups are combined with groups from a small region of interest where intensive admixture is hypothesized. Median F_{ST} for this landscape type is 0.06 (**Fig. S13a**), which is typical for the commonly studied human groups from the Stone- and Bronze Age Europe (F_{ST} between 0.05 and 0.09, Speidel et al. 2024).

To keep the simulations computationally tractable, the genome size was restricted to 500 Mbp with mutation and recombination rates that are realistic for humans (see Methods). See **Fig. S13a** and **Fig. S14** for F_{ST} distributions and counts of polymorphic sites in these simulations, respectively. In the following sections we explore effects on performance of *qpAdm* protocols of the following variables: 1) landscape sparsity; 2) randomized vs. systematic sampling of demes for *qpAdm* experiments (see examples in **Fig. 7**); 3) the four

812 basic types of high-throughput *qpAdm* admixture screens described above (distal and
813 proximal, rotating and non-rotating, summarized in **Fig. S12**); 4) and a wide range of
814 composite *qpAdm* model feasibility criteria (see Methods).

815

816 ***SSL histories: randomized sampling of landscapes***

817 Our motivation for focusing on randomized sampling of landscapes for subsequent *qpAdm*
818 analysis is informed by the archaeological practice. In most regions and periods, a substantial
819 fraction of ancient human population remains archaeologically (and thus archaeogenetically)
820 invisible due to various reasons: 1) burial rites that leave no suitable remains such as
821 cremation (Cerezo-Román et al. 2017), tree burials (Lindig 1964), or burial on or above ground
822 (Pavlinskaya 2009); 2) poor DNA preservation conditions such as acidic soil or hot climate
823 (Baxter 2004); 3) biases in archaeological research that is often driven by salvage excavations
824 in industrial/agricultural landscapes, but leave other landscapes out, or by political agendas;
825 4) lack of exhaustive and systematic sampling of even the existing anthropological collections.
826 Thus, even a completely random sampling of a genetic landscape, as implemented in this
827 study, may be viewed as a reasonable conservative baseline for studying method
828 performance.

829 As stated above, we explored four *qpAdm* protocols (**Fig. S12**). For example, in the case of
830 the distal rotating protocol, for a target deme randomly picked from the "present" sampling
831 point, we sampled 13 or 18 demes randomly from the two older sampling points combined
832 (100 and 300 generations before present), and they formed a rotating set of proxy sources
833 and "right" groups (**Fig. 7**). All possible one- to four-way *qpAdm* models on this subset of 13
834 or 18 demes were tested, and this collection of models for one target we termed "a *qpAdm*
835 experiment" (**Box 1**). Below we describe our analysis of *qpAdm* performance (based on
836 12,800 experiments and 27,511,200 individual *qpAdm* models) and propose guidelines for
837 interpreting *qpAdm* results on randomly sampled SSL, both at the level of models and
838 experiments (the FDR results on the AGS histories above are based on models, not
839 experiments, but experiments were considered in **Suppl. text 2**).

840 Instead of a binary classification of models into TP and FP, which makes no sense in the case
841 of SSL, we ranked them by various "optimality metrics" based on (maximal or average) spatial

842 distances between the target and sources and on angles formed by pairs of sources and the
843 target (minimal source-target-source angle; **Fig. 7, Box 2**). Spatial distance is a noisy
844 approximation for genetic distance, but since gene flows between non-neighboring demes
845 were not allowed, gene-flow intensities on the landscapes (in the last epoch including our
846 sampling points) were relatively uniform for all the SSL types except for "10^{-5} to 10^{-2}" (**Table**
847 **S6, Fig. S13c**), and since our main analyses were randomized and replicated (gene-flow
848 intensities were re-sampled randomly for each simulation replicate and sets of demes for
849 *qpAdm* experiments were drawn randomly), relationships between genetic (F_{ST}) and spatial
850 distances are strong, approximately linear (Pearson's *r* from 0.65 to 0.76), and remain
851 identical across *qpAdm* protocols on the same SSL type and very similar across SSL types,
852 judging by slopes of linear trends (**Fig. S13d,e**). These statements are true for three SSL types,
853 apart from "10^{-5} to 10^{-4}" (**Fig. S13d,e**). Thus, differences in *qpAdm* performance which we are
854 most interested in, namely those across *qpAdm* protocols, gene-flow intensity ranges, and
855 model complexities (established with the help of distance-based model optimality metrics),
856 depend on factors other than the shape of the genetic-spatial distance relationship.
857 Moreover, genetic distance measures such as F_{ST} and haplotype sharing statistics on SSL
858 depend on gene flows in both directions, but *qpAdm* results are traditionally interpreted in
859 terms of source-to-target gene flows only.

860

861 **Box 2.** Definitions relevant for *qpAdm* models on SSL.

maximal (or average) source-target distance (max. ST distance, average ST distance)	The largest (or average) spatial distance on an SSL between source demes and the target deme within one admixture model. In this study, non-contemporaneous samples from the same deme were assigned a distance of 0, the nearest neighbors a distance of 1, and so on.
minimal source-target-source angle (min. STS angle)	The smallest angle formed by any pair of proxy sources and the target in an admixture model in the context of an SSL. If location of a source coincides with the target deme, minimal STS angle is undefined.
symmetrically arranged proxy sources	Demes arranged on an SSL with respect to the target at 180°, 120°, or 90° (for two-, three-, and four-way models, respectively).
ideal symmetric model	In the case of two-way and more complex models on an SSL, these are models where all sources are the nearest neighbors of the target (distance = 1 in this study) and are arranged

	symmetrically. Although models with a source at a distance of 0 are even better, they were not included in this definition since for them min. STS angles were undefined. Due to the topology of the underlying triangular lattice, we lacked ideal four-way models. Although we use the term "ideal", these models are ideal only in the context of the respective model complexity class since full description of target's ancestry on our simulated SSL requires up to 7 sources: the immediate ancestor and up to six neighboring demes.
distance to the ideal symmetric model (in the space of optimality metrics)	Euclidean distance in a two-dimensional space formed by two model metrics: max. (or average) ST distance and min. STS angle. Min. STS angles are normalized: they are divided by a constant such that the maximal angle (180°, 120°, or 90° for two-, three-, and four-way models, respectively) equals 9, which is the maximal possible source-target distance on our landscapes. Distance in the space of optimality metrics is measured between a given admixture model and the ideal symmetric model of the same complexity.
model optimality metrics	The following metrics are collectively termed "optimality metrics": max. (or average) ST distance, min. STS angle, and distance to the ideal symmetric model in the optimality metric space.
landscape sparsity	Landscape sampling density which depends both on the range of gene-flow intensities simulated and on the number of demes selected for an analysis. Landscapes with lower gene-flow intensity and/or deme sampling density for *qpAdm* are sparser, given their equal size in demes and identical topology. We use a term "dense landscape" instead of "landscape with low sparsity".

qpAdm performance at the level of models: model optimality metrics in spaces of estimated admixture fractions and p-values

Our exploration of *qpAdm* performance on AGS histories above and nearly all *qpAdm* results in the literature rely on fixed *p*-value thresholds, usually set at 0.01 or 0.05 (Skoglund et al. 2017, Harney et al. 2018, Narasimhan et al. 2019, Olalde et al. 2019, Yang et al. 2020, Carlhoff et al. 2021, Fernandes et al. 2021, Librado et al. 2021, Zhang et al. 2021, Bergström et al. 2022, Lazaridis et al. 2022, Oliveira et al. 2022, Nakatsuka et al. 2023, Taylor et al. 2023, Zeng et al. 2023, Kerdoncuff et al. 2024). Implicit usage of higher "fuzzy" *p*-value thresholds is also common in the literature when only models with the highest *p*-values are interpreted from historical perspective (Narasimhan et al. 2019, Fernandes et al. 2020, Bergström et al. 2022,

873 Lazaridis et al. 2022, Brielle et al. 2023, Carlhoff et al. 2023, Speidel et al. 2024). Another part
874 of the composite feasibility criteria, namely conditions on estimated admixture fractions
875 (EAF), are also used very often (see the Introduction for a brief overview of the literature).
876 With these considerations in mind, we first inspect results (all *qpAdm* protocols pooled) for
877 the four landscape types in the coordinates of max|*EAF–EF*| and *p*-values (**Fig. 8, Fig. S15**).
878 We also analyzed spaces of EAF standard errors reported by *qpAdm* vs. *p*-values (**Figs. S16,**
879 **S17**). On these two-dimensional spaces we mapped median values of the "optimality metrics"
880 characterizing *qpAdm* models: maximal (**Fig. 8**) or average (**Fig. S15**) source-target (ST)
881 distance in a model and minimal source-target-source (STS) angle (**Figs. 8, S15**).

882 It is immediately obvious that continent-wide landscapes sampled very sparsely
883 (approximated by SSL with gene-flow intensities per generation from ca. 10^{-5} to 10^{-4}) are
884 unsuitable for *qpAdm* analyses since the model bins with low *p*-values and EAF between 0
885 and 1 are not enriched in optimal models (those with small max. ST distances and/or
886 symmetrically arranged sources) (**Fig. 8**) and absolute Spearman's rank correlation coefficient
887 (ρ) for max|*EAF–EF*| or *p*-value vs. any optimality metric does not exceed 0.1 (if max|*EAF–*
888 *EF*| $\leq$0.5 and *p*-value $\geq 10^{-5}$; see **Table S7**). Although this high level of sparsity is probably rare
889 in practice, it may approach the current sampling situation for the Paleolithic humans (Posth
890 et al. 2023, Zeng et al. 2023, Allentoft et al. 2024). In contrast, for the other landscape sparsity
891 levels we simulated there are strong correlations between max|*EAF–EF*| and the model
892 optimality metrics, for some model complexities at least (**Fig. 8, Table S7**; see also correlation
893 coefficients stratified by *qpAdm* protocols in **Table S8**). For instance, in the case of two-way
894 models there is a sharp boundary between models with EAF between 0 and 1 and those
895 outside of this interval (the latter corresponding to max|*EAF–EF*| >0.5): the area to the left
896 of this boundary is enriched in symmetric models, while the area to the right of it is enriched
897 in asymmetric models, and this is true for all three denser SSLs (**Fig. 8**). This corresponds to
898 significant Spearman's ρ between –0.65 and –0.81 for max|*EAF–EF*| vs. min. STS angle for
899 two-way models on the "10^{-4} to 10^{-3}", "10^{-5} to 10^{-2}", and "10^{-3} to 10^{-2}" landscape types (*p*-
900 values $\geq 10^{-5}$ or $\geq 10^{-55}$; **Table S7**).

901 This result is trivial in the context of *qpAdm* and shows that the method usually fails to report
902 EAF in the interval (0, 1) for models with asymmetrically arranged proxy sources. There is a
903 similar, albeit more complex, relationship of max|*EAF–EF*| with maximal (**Fig. 8**) or average

(**Fig. S15**) ST distances. These relationships of max|*EAF*–*EF*| with the angle- and distance-based optimality metrics gradually disappear with growing model complexity and landscape sparsity (**Fig. 8, Fig. S15**) as illustrated by Spearman's ρ (**Table S7**). Thus, there exists a limit to model complexity that makes sense to explore with *qpAdm* on SSL.

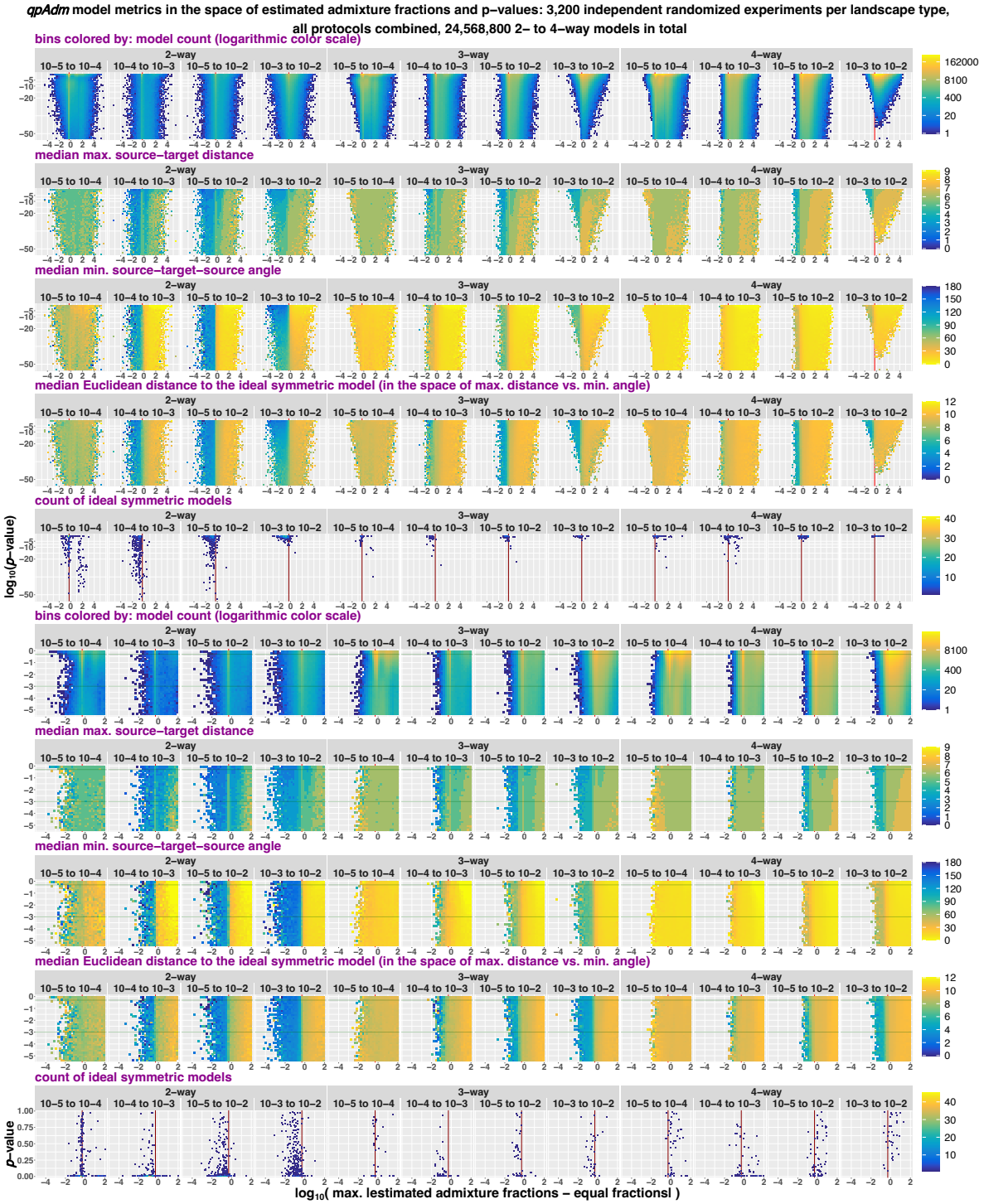


Figure 8. Randomized two to four-way *qpAdm* models in the space of max|*EAF*–*EF*| and *p*-values (both axes are logarithmic). Results for all *qpAdm* protocols are combined but stratified by model complexity and landscape type. Two sections of the space are shown: *p*-values from 10^{-55} to 1 (the

upper five rows of plots) and from 3.16×10^{-6} to 1 (the four rows below them). The space is divided into rectangular bins (50×50 for the larger section and 35×35 for the smaller section), and they are colored by density of individual *qpAdm* models populating this space (logarithmic color scale) or by median values of model optimality metrics in those bins: max. ST distance, min. STS angle, and Euclidean distance to the ideal symmetric model. Density of ideal symmetric models (in the case of four-way models, the most optimal non-ideal models available on our SSL) in this space is also shown (in 75×75 bins). To assess if distributions of ideal symmetric models over *p*-values are uniform, we use non-logarithmic *p*-value scale (the bottom row of plots). The red vertical lines (or the red tick marks) mark max|*EAF*–*EF*| = 0.5, and the green horizontal lines mark the lowest and highest *p*-value thresholds used in this study (0.001 and 0.5).

In the dimension of *p*-values, there also exists a relationship with the model optimality metrics, but it is much weaker (**Fig. 8, Fig. S15**). For example, in the practically relevant range of *p*-values (10^{-5} to 1) and for max|*EAF*–*EF*| ≤0.5, there is no correlation with max. ST distance and with min. STS angle (**Table S7**) on the "10^{-4} to 10^{-3}", "10^{-5} to 10^{-2}", and "10^{-3} to 10^{-2}" landscapes (**Fig. 8**). There is weak correlation with average ST distance in the same part of the space (significant Spearman's ρ up to –0.27, **Table S7**). This relationship also decays with growing model complexity and landscape sparsity (**Table S7**).

We observe that *qpAdm* favors models including target's nearest neighbors that are, moreover, arranged symmetrically around the target, so-called "ideal symmetric models" (**Box 2**), and our analyses in the next section further support that statement (**Fig. S19a**). Considering this observation, it is convenient to merge the angle- and distance-based optimality metrics into one, that is Euclidean distance in the space of max. (or average) ST distances and min. STS angles between a given admixture model and the ideal symmetric model of the same complexity (**Fig. 7, Box 2**). Considering *p*-values in a wide range (10^{-55} to 1) and max|*EAF*–*EF*| ≤0.5, there is no correlation between *p*-values and these two composite optimality metrics in the case of three- and four-way models (the most negative ρ is –0.04, **Table S7**), and weak correlation in the case of two-way models, but on the "10^{-3} to 10^{-2}" landscapes only (significant ρ = –0.39 and –0.48; **Table S7, Fig. 8, Fig. S15**). While conditions on admixture proportion estimates do enrich the set of retained randomized models for symmetric ones, *p*-value thresholds used in practice (0.001, 0.01, 0.05, and even higher implicit thresholds) discard fractions of randomized ideal symmetric models that are much greater than these thresholds (**Table S9**): e.g., the 0.05 threshold discards 44.7% of three-way ideal models on the "10^{-5} to 10^{-2}" landscapes. The only condition where the distributions

946 approach uniformity is the "10^{-3} to 10^{-2}" landscapes and three- and four-way models (**Fig. 8,**
947 **Fig. S15, Table S9**).

948 Thus, the distributions of ideal randomized models on the *p*-value axis become uniform with
949 decreasing landscape sparsity and increasing model complexity, which is an expected
950 behavior since none of the models we tested are truly ideal considering the topology of the
951 landscape with an immediate ancestor and up to six neighboring demes for each target deme
952 (**Fig. 7**, see however an exploration of six-way models in the systematic SSL sampling context
953 below). Another reason for the non-uniformity of the *p*-value distributions are complex
954 networks of gene flows that permeate the landscape and lead to low *p*-values of *qpAdm*
955 models due to violations of the assumptions on the "left" – "right" gene flows (**Fig. 1**) that
956 were derived originally on AGS histories (Haak et al. 2015, Harney et al. 2021), but not on SSL.
957 In summary, in any *qpAdm* screening relying on stringent *p*-value thresholds such as 0.01
958 there exists a large class of models that have non-negative admixture proportion estimates
959 and are relatively close to optimal ones but are rejected according to *p*-values. We explore
960 performance of threshold-based *qpAdm* protocols in the following sections.

961 We note that ideal four-way models (those with the nearest target's neighbors arranged at
962 90°) are lacking on our landscapes based on the triangular lattice: demes there are arranged
963 with a step of 60° around a chosen deme. In **Figs. 8**, **S15-S17**, **S27**, **S28**, and in **Table S9** we do
964 show results for "ideal" four-way models where all sources are targets' nearest neighbors
965 and min. STS angles are above 0, but they show a marked uptick in negativity of admixture
966 proportion estimates as compared to truly ideal two- and three-way models (**Table S9**), which
967 reflects the fact that these models are actually not ideal even in their complexity class.

968

969 *qpAdm performance at the level of models: model optimality metrics in spaces of admixture*
970 *fractions SE and p-values*

971 Maximal standard errors of EAF (abbreviated as max. SE) correlate strongly with max|*EAF–*
972 *EF*| (**Figs. S16, S17**): Spearman's ρ range from 0.81 to 0.94 (when models are stratified by
973 landscape type and model complexity) or from 0.76 to 0.95 (when models are stratified by
974 landscape type, model complexity, *qpAdm* protocol, and sampling density; all the coefficients
975 are statistically significant). Therefore, if we consider max. SE instead of max|*EAF–EF*|, the

same patterns are visible with respect to *p*-values and optimality metrics, but they are less clear-cut in the former case (**Figs. S16, S17, Tables S7, S8**). For instance, Spearman's ρ for min. STS angle vs. max|*EAF*–*EF*| are in all cases higher, albeit slightly, than those vs. max. SE (for two- to four-way models on the "10^{-4} to 10^{-3}", "10^{-5} to 10^{-2}", and "10^{-3} to 10^{-2}" landscape types, and *p*-values $\geq 10^{-55}$ or $\geq 10^{-5}$; **Table S7**). The same is true for correlations between *p*-values and optimality metrics in the most practically relevant parts of the two-dimensional spaces we inspect here (**Table S7**).

Inspecting the distributions of max. (or average) ST distances in the SE dimension, we note "striped" patterns, namely two or more separate SE intervals populated by sources relatively close to targets, probably arising due to the "quantized" nature of our landscape (**Figs. S16, S17**). Interestingly, models having the lowest SE do not demonstrate the smallest median max. ST distances. In contrast, that part of the space is populated by most symmetric models (**Figs. S16, S17**) suggesting that symmetric arrangement of sources is more important for achieving low SE than spatial proximity to the target.

qpAdm performance at the level of models: one-way models

High *p*-values of one-way *qpAdm* models (equivalent to "pairwise *qpWave* tests") are often used as evidence for "cladality" and "genetic continuity" in the archaeogenetic literature (see, e.g., Papac et al. 2021, Changmai et al. 2022a, Skourtanioti et al. 2023), and such *qpWave* tests also became popular as a method for individual clustering that precedes other analyses (Fernandes et al. 2020, Fernandes et al. 2021, Lazaridis et al. 2022, Antonio et al. 2024, Speidel et al. 2024). We show in the context of our randomized *qpAdm* experiments that their interpretation on SSL depends on properties of those landscapes (**Fig. S18**). For instance, of one-way models having *p*-value ≥ 0.05 on the sparsest landscapes ("10^{-5} to 10^{-4}"), just 8.7% include groups sampled from the same deme at different times (spatial distance = 0; **Table S10**), and there is a relatively flat distribution of models over spatial distances in this subset of models: 54.7% of "fitting" one-way models include demes at a distance of 3 or greater (**Fig. S18**, **Table S10**). And frequency of "fitting" one-way models in our equal-sized sets of random experiments is also highly variable across landscape types: e.g., it is ca. 10 times higher for the "10^{-5} to 10^{-4}" landscapes than for the "10^{-4} to 10^{-3}" landscapes (**Fig. S18**, **Table S10**). In

1006 the latter case ("10^{-4} to 10^{-3}"), most "fitting" one-way models (~85%) correspond to samples
1007 from the same deme (**Table S10**). With growing gene-flow intensity the rate of one-way
1008 model rejection grows, as expected, and the mode of this distribution moves from samples
1009 from the same deme to the nearest demes (**Table S10**).

1010 The statements above describe various aspects of the shape of model distributions in the
1011 space of spatial distances between demes and *p*-values. There is a non-linear trend line that
1012 links *p*-value and spatial distance on all the landscapes, except for the sparsest landscape type
1013 where the shape of the distribution is strikingly different and more complex (**Fig. S18**). We
1014 also note that ideal one-way models (defined here as pairs of samples from the same deme)
1015 show very similar distributions over *p*-values for all the four sparsity levels, and with growing
1016 gene-flow intensity on the landscape these distributions become more different from the
1017 uniform one, as expected. Therefore, the differences in model distributions between the
1018 landscapes are mostly due to "non-ideal" models, which are much more common in our
1019 random model samples (**Fig. S18**).

1020 To sum up, on the sparsest landscapes we simulated the high frequency of pairwise *qpWave*
1021 tests reporting high *p*-values reflects the initial multifurcation (**Fig. 7**) and a very low rate of
1022 subsequent gene flows, that is a situation that approaches all vs. all cladality (**Fig. S18**, **Table**
1023 **S10**). However, on landscapes of higher density the method works mainly as a test for genetic
1024 continuity since it identifies samples from the same deme, and, with increasing gene-flow
1025 intensity, clusters of nearby demes become progressively less structured, and the method
1026 identifies members of these clusters as "genetically continuous". On the contrary, samples
1027 from the same deme at different time points become more dissimilar with increasing gene-
1028 flow intensities, and *qpWave* is sensitive to that (**Fig. S18**). Thus, interpretation of pairwise
1029 *qpWave* results in practice is confounded by these two very different scenarios producing
1030 equal outcomes, i.e., high *p*-values. In other words, interpretation of a high *p*-value
1031 encountered in practice could be hindered by a lack of knowledge about the sparsity of the
1032 underlying landscape.

1033

1034 *qpAdm performance at the level of models: threshold-based analyses in spaces of model*
1035 *optimality metrics*

1036 In the previous section we undertook a very general overview of *qpAdm* results on SSL and
1037 reached some preliminary conclusions on performance of threshold-based *qpAdm* protocols
1038 used in practice for finding optimal models for target groups or individuals. Here we apply
1039 the thresholds on *p*-values and EAF explicitly and explore a wide range of composite feasibility
1040 criteria, looking at distributions of models passing these criteria in two-dimensional spaces
1041 formed by the optimality metrics (**Fig. 7**). Models with at least one source occupying the same
1042 position as the target were not included in this analysis as we considered min. STS angles to
1043 be undefined for them.

1044 As explained in **Box 1**, composite feasibility criteria include not only conditions on EAF and *p*-
1045 value thresholds but may also include a condition that all "trailing" simpler models are
1046 rejected according to *p*-values ("trailing" models should not be confused with all simpler
1047 models tested for a given target in an experiment, see **Box 1**). As another condition (rarely
1048 used in practice, to our knowledge), we tested a *p*-value threshold for rejection of "trailing"
1049 models that differs from the "main" *p*-value threshold. In total, we tested 36 composite
1050 feasibility criteria, with *p*-value thresholds ranging from 0.001 to 0.5, which covers the whole
1051 range of thresholds used in the literature explicitly or implicitly, and with three types of
1052 conditions on EAF (see **Table S11** and Methods for a complete list). Results for seven selected
1053 criteria, from the most relaxed to the most stringent, are presented in the figures cited below.

1054 Since this analysis was performed on all possible combinations of source demes drawn
1055 randomly from SSL, the spaces of optimality metrics are populated by models highly unevenly
1056 (**Fig. 9**; see the spaces based on average ST distance in **Fig. S20**). This may be viewed as a
1057 drawback of our analysis but, on the other hand, this randomized analysis may inform us on
1058 real-world performance of *qpAdm* protocols (that test all combinations of many source
1059 groups) since the existing archaeogenetic sampling of humans and other species in many
1060 regions/periods is sparse and nearly random.

1061 First, we explore the positive rate of *qpAdm* (**Box 1**), dividing the spaces of optimality metrics
1062 into two-dimensional bins: the positive rate is simply a fraction of all models tested (within a
1063 bin) passing a feasibility criterion (**Fig. S19a,b,d, S20a,c,e**). Under stringent criteria (such as
1064 EAF $\pm$ 2 SE in the interval (0, 1), *p*-value $\geq$0.05, "trailing" models rejected), positive rates
1065 approach 0 in most parts of the optimality metric space except for its upper left corner where
1066 ideal symmetric models (**Fig. 7, Box 2**) lie. This is true for most levels of landscape sparsity

1067 and for most model complexities we tested (**Figs. S19a, S20a**). These results are based on all
1068 the *qpAdm* protocols combined, and we show results for rotating and non-rotating protocols
1069 separately in **Figs S19b,d** and **S20c,e** (results for each individual protocol are not presented
1070 since they were generally very similar within the rotating and non-rotating classes). However,
1071 enrichment of feasible model sets for ideal symmetric models is not a universal trend: as
1072 expected from our exploration of max|*EAF*–*EF*| vs. *p*-value spaces above, the enrichment
1073 level falls with growing landscape sparsity (see the results for the "10^{-5} to 10^{-4}" landscapes)
1074 and model complexity (see the results for three-way and especially four-way models).
1075 Another trend noticeable is poorer performance of the non-rotating protocols as compared
1076 to rotating protocols: with growing stringency of feasibility criteria, the "hot spot" of *qpAdm*
1077 positive rate moves not to ideal symmetric models, but to symmetric models including at
1078 least one more distant source (compare **Figs. S19b** and **S19d**, **S20c** and **S20e**). A similar trend
1079 is observed when rotating protocols are applied to the "densest" landscape ("10^{-3} to 10^{-2}"),
1080 but in the case of four-way models only (**Figs. S19b** and **S20c**). In other words, sources closest
1081 to a target are usually rejected in the context of complex models on densely sampled
1082 landscapes with high gene-flow intensities.

1083 At the other end of the feasibility criteria stringency range tested (such as EAF ∈ (0, 1), *p*-
1084 value ≥0.01, "trailing" models not checked) in the space of max. ST distance and min. STS
1085 angle, *qpAdm* positive rate is strongly and negatively correlated to density of randomized
1086 models populating the same space (**Table S11**). In other words, *qpAdm* positive rate peaks
1087 not in the upper left corner only, but in an arc on the upper edge of the optimality metric
1088 space (**Fig. S19**). This pattern is due to the fact that large min. STS angles are strongly
1089 associated with non-negative EAF (∈ (0, 1)), which is a part of the composite feasibility
1090 criterion, while max. ST distances are much less strongly associated with both non-negative
1091 EAF and high *p*-values (**Fig. 8, Table S7**). This pattern disappears with growing stringency of
1092 the feasibility criteria, at varying rates depending on landscape sparsity and model
1093 complexity, and it is absent in the optimality metric spaces based on average ST distances,
1094 except for two-way models on the "10^{-5} to 10^{-2}" landscapes (**Fig. S19, Table S11**).

1095 For assessing practical significance of *qpAdm*'s "preference" for nearly ideal symmetric
1096 models it is important to look at the system not only through the lens of positive rates (**Fig.**
1097 **S19**), but also at density of fitting models in the optimality metric space (**Fig. 9**). That density

is influenced both by details of *qpAdm* protocols and by random landscape sampling itself, and we consider this approach as a conservative approximation for high-throughput *qpAdm* protocols on the real-world archaeogenetic sampling. Results for rotating and non-rotating protocols are shown separately in **Figs. S19c,e** and **S20d,f**. Results in the spaces based on average ST distance are shown in **Fig. S20b,d,f**. We use Spearman's ρ based on model counts in bins to compare two-dimensional distributions (in the optimality metric space) of fitting models and of all target-source configurations prior to *qpAdm* testing (**Table S11**).

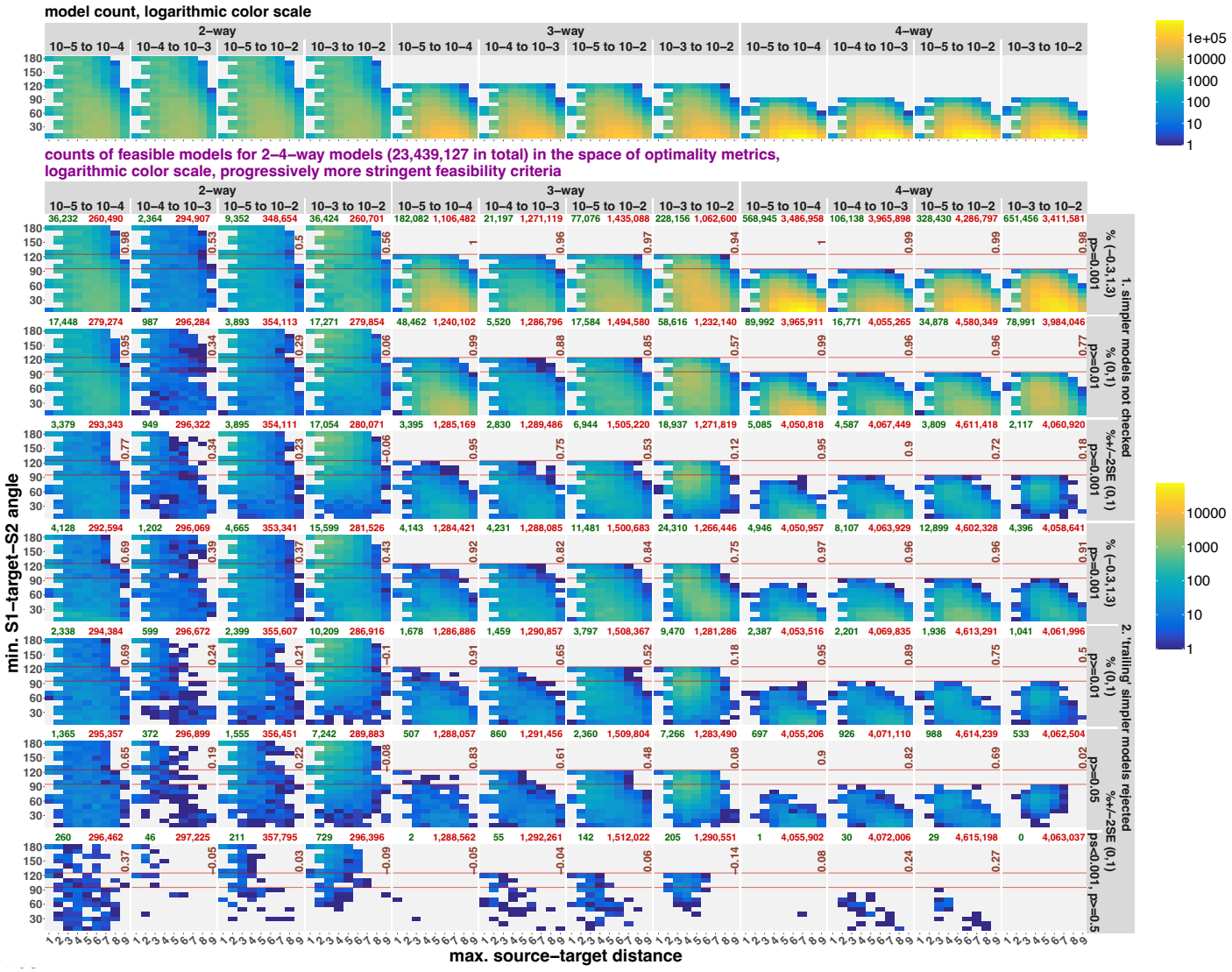


Figure 9. Randomized two to four-way *qpAdm* models in the space of model optimality metrics: max. ST distance and min. STS angle (models with the latter metric undefined were excluded from this analysis). Results for all *qpAdm* protocols are combined but stratified by model complexity and landscape type. The upper row of plots shows density of all randomized models tested, and the other rows show density of fitting models satisfying selected composite feasibility criteria listed in the captions on the right (7 of 36 criteria tested in this study, from the least stringent on top to the most stringent at the bottom). All the color scales are logarithmic. The space was divided into 9 bins on the x-axis (bin width = 1), and 18 bins on the y-axis (intervals [0°, 10°], (10°, 20°], and so forth). Spearman's

correlation coefficients for counts of all models vs. fitting models in these bins are shown in each panel in brown (bins not populated by any models were not considered). The red horizontal lines mark the "ideal" min. STS angles for three- and four-way models (120° and 90°). The red and green numbers above each plot stand for counts of rejected and fitting models in those analyses, respectively.

In the cases of four- and three-way models, only very specific landscape properties and feasibility criteria led to depletion of the sets of fitting models for non-optimal models (those with small min. STS angles and at least one distant source), which is evidenced by Spearman's ρ (for total model counts vs. fitting model counts in two-dimensional bins) not significantly different from 0 under these conditions, but high and significant under nearly all the other conditions tested (**Table S11**). These conditions are as follows: the densest landscape tested ("10^{-3} to 10^{-2}") and EAF $\pm$ 2 SE $\in$ (0, 1). These results are in line with our observations in the preceding section (**Fig. 8**), namely that EAF (and their SE correlated with them) and landscape sparsity, in contrast to model *p*-values (which are highly variable due to gene flows in all directions on the landscapes), are correlated to both angle- and distance-based model optimality metrics strongly. We do not highlight other acceptable combinations of landscape sparsity and feasibility criteria such as "10^{-3} to 10^{-2}" and EAF $\in$ (0, 1) (**Table S11**) since they are applicable to a narrower range of model complexities (do not perform well for four-way models, for example), or yield very low overall positive rates.

Introducing very stringent feasibility criteria (such as *p*-value $\geq$0.5, "trailing" models rejected with p <0.001, EAF $\pm$ 2 SE $\in$ (0, 1)) is probably impractical since it drives the overall *qpAdm* positivity rate to extremely low values, especially in the case of complex models: on the order of 10^{-4} to 10^{-6} for three-way models and 10^{-6} to 10^{-7} and even lower for four-way models (**Fig. 9, Table S11**). As we discuss below in the section "Systematic sampling: exploring truly ideal six-way models", despite *p*-values growing with increasing model complexity, fraction of models with EAF $\in$ (0, 1) falls sharply, probably due to increasing uncertainty in EAF estimates, and that is responsible for the low positive rates based on composite feasibility criteria. For practical purposes, to keep overall *qpAdm* positive rate reasonably high (ca. 10^{-2} to 10^{-4} depending on model complexity, see **Table S11**), we recommend using the least stringent feasibility criteria satisfying the conditions described above: restricting *qpAdm* sampling (in both "right" and "left" sets) to sub-continental regions such as West Asia in the case of humans (which corresponds approximately to the "10^{-3} to 10^{-2}" landscapes); stringent

1146 conditions on EAF such as EAF $\pm$ 2 SE $\in$ (0, 1); a low *p*-value threshold for "fitting" models
1147 (such as 0.001) and no *p*-value thresholds for rejecting simpler "trailing" models. We note
1148 that, following this approach, it is hardly possible to find a single best model complexity level
1149 for a target and a single best model at that complexity level, but that is hardly achievable in
1150 principle on SSL as we discuss in the next section.

1151 We analyzed this most optimal combination of landscape sparsity and feasibility criteria in
1152 greater detail, looking at results stratified by *qpAdm* protocols in **Figs. S21** (optimality metric
1153 spaces based on max. ST distance) and **S22** (optimality metric spaces based on average ST
1154 distance). On these spaces we map density of models tested (**Figs. S21a, S22a**), *qpAdm*
1155 positive rates (**Figs. S21b, S22b**), and counts of feasible models (**Figs. S21c, S22c**). No striking
1156 differences in enrichment for nearly optimal models are observed across all the four
1157 protocols and two sampling densities tested.

1158 Inspecting the spaces formed by average ST distance and min. STS angle, we observe the
1159 same trends (**Fig. S20b,d,f**). Considering the radius and shape of the SSL studied here (**Fig. 7**),
1160 the triangular region in the lower right corner of the optimality metric space (**Fig. S20b,d,f**) is
1161 populated by models where a target is located close to (or on) one edge of the landscape and
1162 at least some proxy sources are located close to the opposite edge of the landscape (average
1163 ST distance >4.5, which is the radius of the landscape). Complete elimination of those highly
1164 asymmetric models from sets of fitting models is more achievable for two-way models, but is
1165 achievable for three- and especially four-way models only on the densest landscape "10^{-3} to
1166 10^{-2}" and/or when the most stringent conditions on EAF are applied (**Fig. S20b, Table S11**).

1167 In summary, all the analyses discussed in this section point to a simple conclusion: when
1168 *qpAdm* models and sets of "right" groups or rotating sets are constructed randomly,
1169 depleting the sets of fitting models for highly non-ideal models is probably practical only for
1170 two-way models, but for three-way and more complex models (which are, naturally, more
1171 interesting in the SSL context) it requires stringent conditions on EAF and access to densely-
1172 sampled landscapes. Importantly, high *p*-value thresholds often used in the literature
1173 explicitly or implicitly ($\geq$0.05) are counter-productive in the SSL context since they do not
1174 improve recovery of optimal models in relative terms but result in very low positive rates.

1175

1176 *qpAdm performance at the level of threshold-based experiments*

1177 In this section *qpAdm* results are interpreted from the perspective of experiments (**Box 1**)
1178 which progress from simple to complex models (from one-way to two-, three-, and four-way
1179 models in our study) and are stopped (produce a positive result) as soon as at least one fitting
1180 model is found for a chosen target. Experiments are defined not only by a target, but also by
1181 a set of "right" demes or a rotating set. Approaches of this sort are often used in high-
1182 throughput screens for admixture that rely on *qpAdm* (see, e.g., Narasimhan et al. 2019,
1183 Bergström et al. 2022, Lazaridis et al. 2022, Zeng et al. 2023, Speidel et al. 2024).

1184 When interpreting *qpAdm* results at the level of experiments, we paid attention to fractions
1185 of experiments producing at least one fitting model at various model complexity levels and
1186 to fractions of those "positive" experiments with outcomes that would be misleading in
1187 practice (see Methods for a definition). For declaring a positive result, it was required that all
1188 simpler models for the target (all possible source combinations) are rejected according to a
1189 chosen composite feasibility criterion. Models at a complexity level where the first fitting
1190 model was found were binned into rejected (non-fitting) and non-rejected (fitting) and
1191 ranked by Euclidean distances to ideal symmetric models in the space of optimality metrics
1192 (**Fig. 9**).

1193 For a complete overview of our results at the level of experiments see **Figs. S23** (optimality
1194 metric spaces based on max. ST distance) and **S24** (optimality metric spaces based on average
1195 ST distance). A key observation is that the most optimal *qpAdm* setup (the densest landscape,
1196 the stringent condition on EAF, and the *p*-value threshold = 0.001) found in the previous
1197 section indeed results in the lowest fractions of positive experiments with "bad" outcomes,
1198 as compared to other conditions tested (compare **Fig. 10** and **Fig. S23**). Here we focus on the
1199 results for all the *qpAdm* protocols tested on the "10^{-3} to 10^{-2}" landscapes under this
1200 condition (EAF $\pm$ 2 SE $\in$ (0, 1), *p*-value $\geq$0.001, "trailing" models not checked), and they share
1201 a common pattern: fitting models were found for 92% of targets; 55% of experiments ended
1202 at one-way models, and 36.5% at two-way models (on average across all simulation replicates
1203 and protocols; **Fig. 10**). Nearly all experiments of both types were not classified as producing
1204 misleading results, namely models rejected in those experiments were not much closer to
1205 the ideal symmetric models (see Methods for a precise definition) as compared to fitting ones

(**Fig. 10**). Nevertheless, the outcome can hardly be viewed as satisfactory since one-way and two-way models are too simple for these landscapes with high-intensity gene flows.

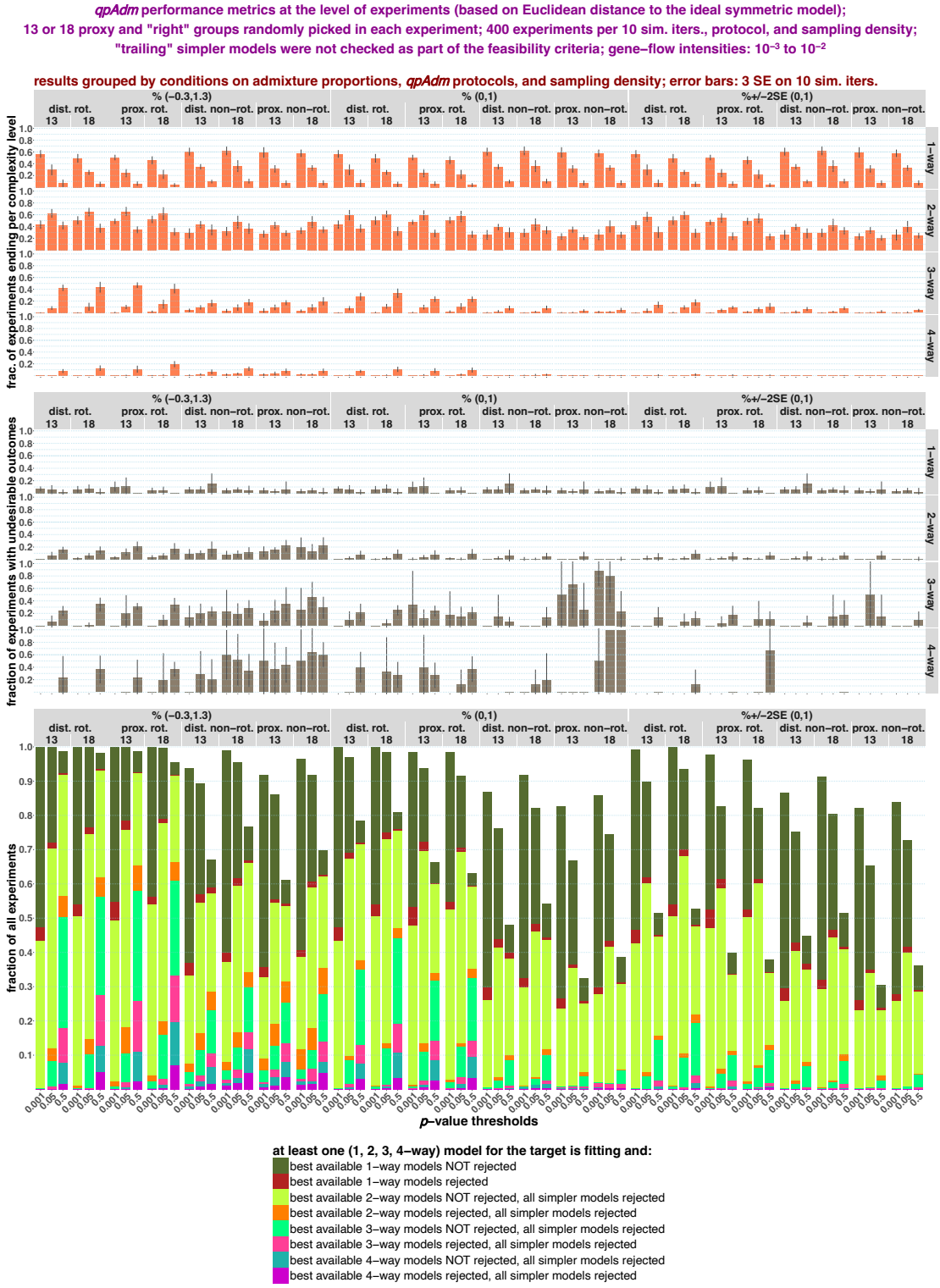


Figure 10. Randomized *qpAdm* results at the level of experiments visualized for "dense" landscapes (10^{-3} to 10^{-2}). Landscapes of this type only demonstrated robust rejection of highly asymmetric sub-optimal models. The results are stratified by *qpAdm* protocols and landscape sampling density (13 or 18 demes in rotating sets and 10 or 15 demes in non-rotating "right" sets, respectively), and also by

model complexity level at which experiments generate positive results, by conditions on EAF, and by *p*-value thresholds (results for three thresholds only are shown: 0.001, 0.05, and 0.5). We show as bar plots fractions of experiments producing positive results (at least one fitting model per target) at each complexity level and fractions of those experiments with positive outcomes classified as potentially misleading or "undesirable". See Methods for a definition based on Euclidean distances in the space of optimality metrics: max. ST distance and min. STS angle. In the case of "bad" experiments, the most optimal model (according to ST distances and angles) available in the chosen deme set was rejected, but a much less optimal model emerged as fitting. The error bars show 3 SE intervals calculated on 10 simulation replicates. Fractions of experiments falling into these different classes (producing positive results at complexity levels from one to four and classified as either misleading or not misleading according to rejection of most optimal models available) are visualized by the stacked bar plots. Error bars are not shown for visual clarity in the latter plot.

Raising the *p*-value threshold to very high values (0.5) and keeping conditions on EAF more relaxed (EAF ∈ (0, 1) or (-0.3,1.3)) resulted in significantly more experiments producing three- (23.6% vs. 0.5% on average) and four-way models (7.7% vs. 0% on average) than under the feasibility criterion applied above, but many experiments in these categories had potentially misleading outcomes (25% on average for three- and 34% for four-way models), and the fraction of "bad" experiments also grew significantly (12% vs. 0.5% on average) for experiments ending at the level of two-way models (**Fig. 10**). This agrees with the results in the previous section since the relaxed conditions on EAF do not allow to reject efficiently the highly asymmetric models in the lower right corner of the optimality metric space (which constitute the bulk of our randomized *qpAdm* model samples) (**Figs. 9, S20b**). Introducing the condition on EAF that was found to be optimal in the previous section (EAF $\pm$ 2 SE ∈ (0, 1)) and keeping the *p*-value threshold at 0.5 allowed to achieve low rates of "bad" experiments ending at two- and three-way models (5% and 12% on average across the protocols and replicates), however most experiments following this setup did not end (57% on average), and those that did end yielded two-way models in most cases (62% on average; **Fig. 10**), and we consider those over-simplified in the context of SSL we simulated (see Discussion).

Similar trends were observed in a much wider range of landscape sparsity and feasibility criteria (**Fig. S23**). Let us consider the most realistic landscape type simulated ("10^{-5} to 10^{-2}") where gene-flow intensities are the least uniform across the landscape (**Fig. S13**). As compared to the "10^{-3} to 10^{-2}" landscapes discussed above, experiments yielded three- and four-way models significantly more often, however if the *p*-value threshold was set at 0.001 and EAF $\pm$ 2 SE ∈ (0, 1), 22% of experiments yielded one-way models, and 31% yielded two-

1248 way models (**Fig. S23c**). As discussed above, increasing the *p*-value threshold shifts the
1249 balance in favor of more complex models, but at the cost of significantly higher rates of
1250 positive experiments with potentially misleading outcomes (**Fig. S23c**). However, introducing
1251 the stringent condition on EAF (EAF $\pm$ 2 SE $\in$ (0, 1)) allowed to achieve low rates of "bad"
1252 positive experiments (up to 10-20%) even for high *p*-value thresholds (**Fig. S23c**).

1253 Another major trend is that a substantial fraction of non-rotating experiments did not end
1254 even at the stage of four-way models (from ca. 90% to 30%, depending mostly on the *p*-value
1255 threshold), in contrast to rotating experiments, and the distributions of experiments by
1256 model complexity at which they end are even more skewed towards simpler models in this
1257 case (**Fig. S23c**). The same is true for landscapes sparser than "10^{-5} to 10^{-2}" (**Fig. S23a,b**). We
1258 supposed that a fundamental difference between these two protocol types may underly this
1259 observation: sub-optimal sets of "right" groups (e.g., those where "right" demes are closer to
1260 the target than most source demes) are unchangeable in each non-rotating experiment, and
1261 thus may have larger influence on their outcomes than more "fluid" sets of "right" demes in
1262 rotating protocols. We explored this issue further in the systematic sampling context (see the
1263 next section): for two target demes in the center of the "10^{-5} to 10^{-2}" landscapes, we tested
1264 all possible combinations of target-"right" and target-proxy distances in the distal non-
1265 rotating protocol (**Fig. S12**). In each experiment, all proxy sources were equidistant from the
1266 target, and the same was true or approximately true for "right" demes (**Fig. S25**). Indeed,
1267 there are combinations of target-"right" and target-proxy distances which cause nearly all
1268 experiments (of 20 performed in each case) to progress beyond four-way models (**Fig. S25**).
1269 But some "inverted" proxy and "right" deme arrangements (the latter closer to the target in
1270 space than the former) demonstrate relatively high experiment ending rates, such as all
1271 proxies at a distance of 4 and "right" demes at distances 1 and 2 (**Fig. S25**). Although our
1272 landscapes are small for a conclusive analysis of this problem, it appears that when "right"
1273 and proxy source demes are close to each other (in space, but not in time), experiment ending
1274 rate becomes low.

1275 As expected from all the analyses above, the sparsest landscapes "10^{-5} to 10^{-4}" demonstrated
1276 by far the worst results at the level of experiments (**Fig. S23a**): under nearly all settings, most
1277 experiments which ended yielded one- and two-way models, with high rates of potentially
1278 misleading experiments for all model complexities (even for experiments ending at one-way

models). A general conclusion from this section is that it is very difficult to find setups for *qpAdm* experiments that give both complex and accurate results: at least three-way models that are close to the ideal symmetric ones.

To illustrate this trend, we show outcomes of two experiments (**Fig. S26**) relying on one of the most optimal feasibility criteria (EAF $\pm$ 2 SE $\in$ (0, 1), p-value $\geq$0.001, "trailing" models not checked), where only one fitting two-way model was found in each (the "10^{-5} to 10^{-2}" landscapes, distal rotating protocol, rotating sets composed of 18 demes sampled 100 and/or 300 generations ago). Attempting to get more complex models (three-way) resulted in non-optimal models being accepted and much more optimal models being rejected (**Fig. S26**). For instance, when deme 20 was set as a target, the most optimal fitting three-way model included a source at a distance of 5 from the target, but the most optimal rejected three-way model included three nearest neighbors of the target (distance = 1), however not all of them sampled at the same time. This alternative model was rejected not because of a low *p*-value, but because of high SE of EAF. This particular case (drawn randomly from a pool of similar "bad" outcomes of experiments) can be interpreted as follows: due to the lack of a sample close in time to the target from one of the nearby demes, a very distant deme emerges as a source due to conditions on EAF and its SE. Considering patchy sampling simulated here (**Fig. S26**), this situation may be wrongly interpreted as a signal of a long-distance migration, especially in a high-throughput setting when each experiment is not carefully assessed.

SSL histories: systematic sampling of landscapes

We aimed to explore how the availability of the close ancestry sources (the nearest neighbors of a target deme, also termed "1st circle") influences outcomes of *qpAdm* experiments on the most realistic simulated landscapes ("10^{-5} to 10^{-2}", see F_{ST} distributions in **Fig. S13a**). In contrast to the randomized deme sets tested above, we used "right" sets composed of demes equidistant from the target and distributed uniformly on the landscape, forming circles around the target (**Figs. 7, S12**). Two kinds of *qpAdm* protocols were tested: distal rotating and distal non-rotating (**Fig. S12**).

1308 *Systematic sampling: model optimality metrics in the spaces of estimated admixture fractions*
1309 *and p-values*

1310 We first explored rotating sets composed of two deme circles around the target: at distances
1311 of 3 and 1 (for a description of the corresponding non-rotating setup see **Fig. S12**). We again
1312 inspect median values of the model optimality metrics in the spaces of max|*EAF*–*EF*| and *p*-
1313 values (**Figs. S27** and **S28**) and show the corresponding Spearman's ρ for the optimality
1314 metrics vs. max|*EAF*–*EF*| or *p*-values in **Table S7**, and correlation coefficients stratified by
1315 *qpAdm* protocols in **Table S8**. In these tables, the results relying on randomized and
1316 systematic deme sampling are juxtaposed. In **Figs. S27** and **S28** (based on max. and average
1317 ST distance, respectively), we also juxtapose results relying on randomized and systematic
1318 sampling of the "10^{-5} to 10^{-2}" landscapes. These two sampling approaches demonstrate
1319 similar results (**Fig. S28**), however there are generally stronger correlations between
1320 max|*EAF*–*EF*| or *p*-values and the optimality metrics in the case of systematic sampling (**Table**
1321 **S7**). This is true, for example for correlations between max|*EAF*–*EF*| and the following
1322 optimality metrics in probably the most practically relevant part of the space (max|*EAF*–*EF*|
1323 ≤0.5 and *p*-value ≥10^{-5}): min. STS angle (for three- and four-way models), max. and average
1324 ST distances (for two- and three-way models), and Euclidean distances to the ideal models in
1325 the space of optimality metrics (for three- and four-way models) (**Table S7**). Even more
1326 interesting are correlations between *p*-values and some optimality metrics (most notably
1327 average ST distance and the corresponding Euclidean distance to the ideal models) that are
1328 strengthened by systematic sampling, however this effect is noticeable only when
1329 considering a wider range of *p*-values (max|*EAF*–*EF*| ≤0.5 and *p*-value ≥10^{-55}) (**Fig. S28, Table**
1330 **S7**).

1331 These results suggest that a particular arrangement of "right" groups (evenly distributed on
1332 a circle around the target) improves performance of the method (that is, correlation between
1333 *p*-values and model optimality metrics), however the improvement is not striking. This
1334 arrangement of "right" demes also leads to rejection of a larger fraction of ideal two- or three-
1335 way models as compared to that under randomized deme sampling (**Table S9**). We note that
1336 for nearly all ideal two- and three-way models under systematic sampling, EAF are non-
1337 negative (**Table S9**).

1338 Next, we broadened the sets of proxy sources tested to encompass all possible one- to four-
1339 deme sets at a particular time (this is feasible in the non-rotating context only) and inspected
1340 the influence of target-"right" distance on *qpAdm* performance: minimal target-"right"
1341 distance varied from 1 to 4 for two suitable target demes at the center of the landscape, while
1342 keeping the size of the "right" set constant at 16 demes. Since the 1st and 2nd circles include
1343 fewer than 16 demes, the "right" sets included demes from the 1st and 2nd circles, or from 2nd
1344 and 3rd, or from 3rd only, or from 4th only. We present the results in the usual way: in the
1345 spaces of max|*EAF–EF*| and *p*-values (**Figs. S29** and **S30**) and as Spearman correlation
1346 coefficients between the model optimality metrics and max|*EAF–EF*| or *p*-values in various
1347 parts of these spaces (**Table S12**). These spaces are in general similar to those based on
1348 randomized *qpAdm* tests explored above (**Figs. 8** and **S15**), with simpler models (two-way)
1349 demonstrating better performance than three- or four-way models. The best performance of
1350 the method (strongest correlation between the model optimality metrics and max|*EAF–EF*|
1351 or *p*-values) is observed when "right" demes are in the 3rd or 4th circles around the target
1352 (**Table S12**). Here we focus on correlations in the following sections of the space: max|*EAF–*
1353 *EF*| vs. the optimality metrics for *p*-values $\geq 10^{-5}$ or 10^{-55}; *p*-value vs. the optimality metrics
1354 for max|*EAF–EF*| $\leq$0.5 (**Table S12**). The distributions of ideal symmetric two- to four-way
1355 models also become more uniform on the *p*-value axis with increasing target-"right" distance
1356 (**Figs. S29** and **S30**). Thus, in the case of the distal non-rotating protocol at least, usage of
1357 demes sampled in the past and close to the target as "right" groups is not a good approach.

1358

1359 *Systematic sampling: exploring truly ideal six-way models*

1360 Due to computational limitations, in all our randomized and systematic *qpAdm* experiments
1361 described above the most complex models tested were four-way. However, on the SSL
1362 simulated in this study (**Fig. 7**), truly ideal models are six-way. Here we do not consider cases
1363 when the direct ancestor is available since this makes modelling population history trivial in
1364 our systematic setups based on the "10^{-5} to 10^{-2}" landscapes: instead of seven-way models,
1365 one-way models fit the data in nearly all cases. Thus, when we referred to "ideal symmetric
1366 models" above, we meant ideal symmetric models in the respective complexity class, but not
1367 truly ideal models.

Here we explore one- to four-way and six-way models for our standard systematic setup: distal rotating and distal non-rotating protocols based on demes from the 1st and 3rd circles around the target (all demes with complete 3rd circles were included as targets). More precisely, demes from the 1st and 3rd circles comprised 16-deme rotating or proxy sets (**Fig. S12**). Results based on this setup are also presented in **Figs. S27** and **S28** and **Tables S7-S9**, but here we show results in ideal conditions only, where all six demes from the 1st circle were available in *qpAdm* experiments (not to be confused with individual models). We first look at the results in the max|*EAF*–*EF*| vs. *p*-value spaces (**Figs. 11a** and **S31a,b**).

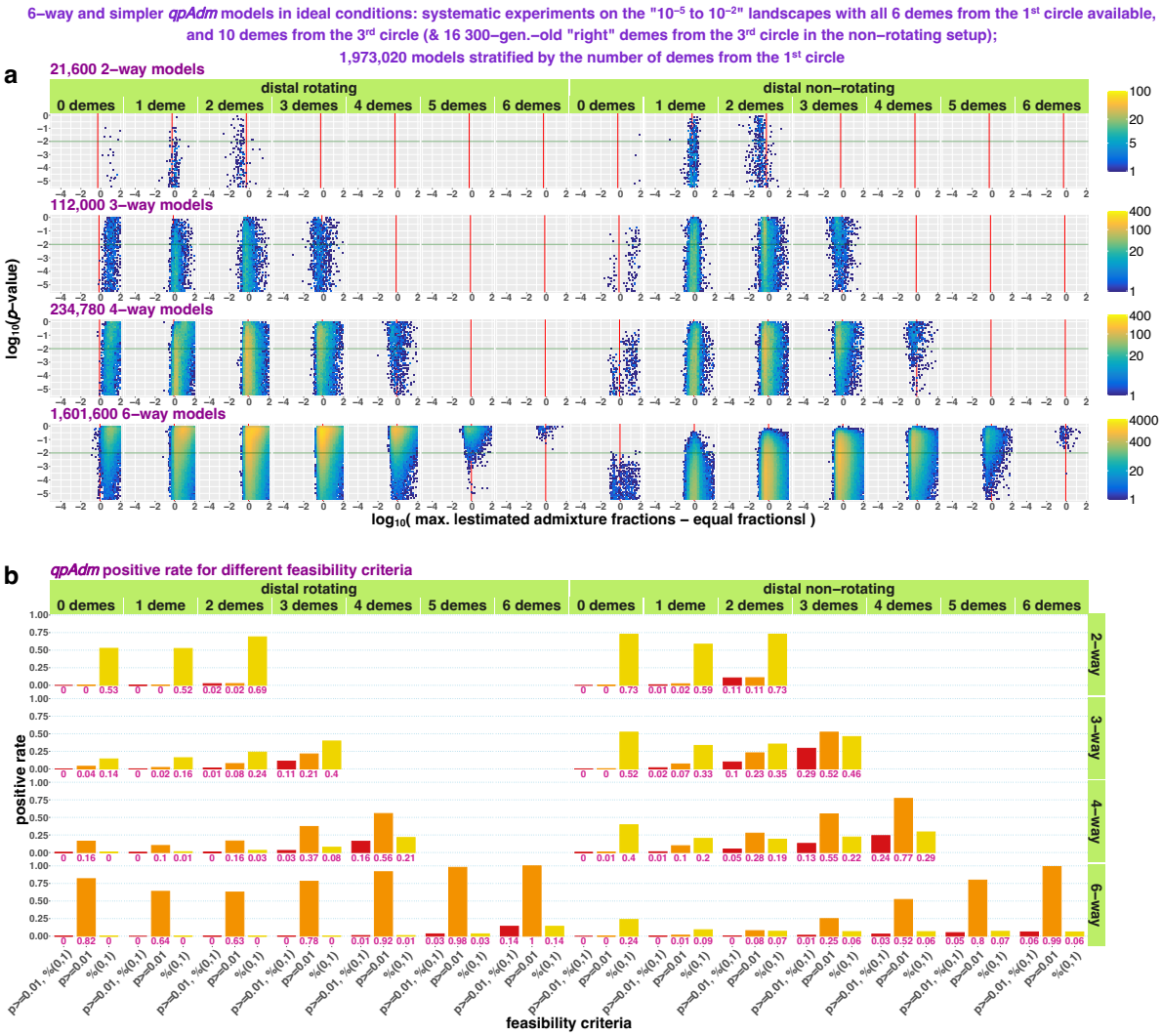


Figure 11. Systematic two to four- and six-way *qpAdm* models on the "10^{-5} to 10^{-2}" landscapes in the space of max|*EAF*–*EF*| and *p*-values (**a**). Results are stratified by model complexity, *qpAdm* protocol, and number of target's nearest neighbors in a model, which is the only optimality metric in this analysis. A part of the whole space is shown (*p*-values from 3.16×10^{-6} to 1), and for a wider section of the space and for one-way models see **Fig. S31a**. The space is divided into rectangular bins (50×50), and they are colored by density of individual *qpAdm* models populating this space (see the logarithmic color scales on the right). The red vertical lines (or the red tick marks) mark max|*EAF*–*EF*| = 1 – *EF*; for three-way and more complex models this boundary is similar but not identical to the boundary

between EAF within (0, 1) and outside (0, 1). The green horizontal lines mark the *p*-value threshold used in all our systematic experiments (0.01). In panel **b**, we show as bar plots corresponding fractions of models that satisfy three kinds of feasibility criteria marked on the x-axis. The fractions are also indicated below the bars.

We noted above that *p*-value is not an optimal criterion for ranking randomized or systematic *qpAdm* models on the "10^{-5} to 10^{-2}" landscapes (and most other landscape types tested) and that max|*EAF*–*EF*| correlates better with the model optimality metrics (**Fig. 8, Tables S7** and **S9**). Indeed, large fractions of "ideal" symmetric 3-way models are rejected according to *p*-value thresholds such as 0.01 or 0.05 (**Table S9**). However, truly ideal six-way models were rejected according to the 0.01 threshold we use for systematic experiments in exceptional cases only (**Fig. 11a**). Non-optimal sources come from the 3rd circle, the same circle all the "right" groups (200 generations deeper in time) in the non-rotating setup and most of the rotating set come from (**Fig. S12**). Thus, removing all demes from the 3rd circle from a *qpAdm* model eliminates detectable "left" to "right" gene flows and results in a high *p*-value. In contrast, 86% or 94% of ideal six-way models were rejected according to the condition EAF ∈ (0, 1) in the rotating and non-rotating setups, respectively (**Fig. 11b**), demonstrating that AF estimates become highly uncertain for complex models.

To look at feasibility criteria systematically, we show *qpAdm* positive rates for models stratified by complexity and the number of demes from the 1st circle included in each model (**Figs. 11b, S31c**). For simplicity, just three criteria are tested: *p*-value ≥ 0.01 and EAF ∈ (0, 1), *p*-value ≥ 0.01 only, and EAF ∈ (0, 1) only. For the systematic distal non-rotating protocol, relying on *p*-values only is the best way of distinguishing highly non-optimal six-way models (comprising demes from the 3rd circle only) from ideal six-way models comprising all the nearest neighbors of the target: in the former case all models fail, but in the latter case 99% of models pass (**Figs. 11b**). We note that *p*-value thresholds do not allow to separate models less different in optimality so clearly (**Fig. 11**). In contrast, such a "clean" result was not achieved with the distal rotating protocol: model feasibility rate based on *p*-value grows with increasing model complexity, but at the same time its correlation with model optimality disappears (**Fig. 11**). Thus, both *p*-values and EAF should be taken into consideration, although that results in rejection of 86% of ideal six-way models due to highly uncertain AF estimates (**Figs. 11b**). Thus, in this particular systematic setup (**Fig. S12**), the non-rotating

1417 protocol with model feasibility criteria based solely on *p*-values is performing substantially
1418 better than the rotating one, but it is hard to say if this observation can be generalized to
1419 other regions of the huge parameter space we try to explore.

1420

1421 *Systematic sampling: qpAdm performance at the level of experiments*

1422 In this section we show that the improvements of *qpAdm* performance under systematic
1423 sampling, reflected in the max|*EAF–EF*| vs. *p*-value spaces (**Figs. S27** and **S28**), do not make
1424 outcomes of *qpAdm* experiments relying on *p*-value thresholds more interpretable. For both
1425 the rotating and non-rotating protocols based on demes from the 1st and 3rd circles around
1426 the target (**Fig. S12**), we observe the following trend: model complexity at which an
1427 experiment ends (the first feasible model is found) decreases with the number of close
1428 ancestry sources (demes from the 1st circle) available in that experiment (**Fig. 12a,b**). This
1429 trend is especially striking in the case of the distal non-rotating protocol: when 0 close sources
1430 were available, although we relied on a relatively relaxed composite feasibility criterion (*p*-
1431 value threshold at 0.01, EAF between 0 and 1) nearly all experiments did not end at the stage
1432 of four-way models, suggesting that five-way or even more complex models are needed (**Fig.**
1433 **12b**). If one to four close sources were available, experiments most often ended at three-way
1434 models, and if all six nearest neighbors of a target were available, a great majority of
1435 experiments ended at two-way models. The trend is less pronounced in the case of the
1436 rotating protocol: it prefers simpler models (four-way) when no close ancestry sources are
1437 available, and prefers three-way models in all other cases, with a slowly growing fraction of
1438 two-way models (**Fig. 12a**).

1439 We also expanded the sets of available proxy sources in the distal non-rotating protocol to all
1440 possible combinations of demes at 100 generations before "present" and varied target-
1441 "right" distance (see above). In line with the results above, most experiments with the widest
1442 selection of sources available ended at two-way models, and a majority of those fitting (EAF
1443 $\in$ (0, 1); *p*-value $\geq$0.01) two-way models included one nearest neighbor of a target (**Fig. S33**).
1444 The fraction of experiments ending at three-way models grew with decreasing target-"right"
1445 distance, and no experiment ended at four-way models (**Fig. S33**). We do not assess statistical
1446 significance of these trends since just 20 experiments were performed per each parameter

combination (two target demes in the center of SSL were suitable for this analysis and 10 simulation replicates were available).

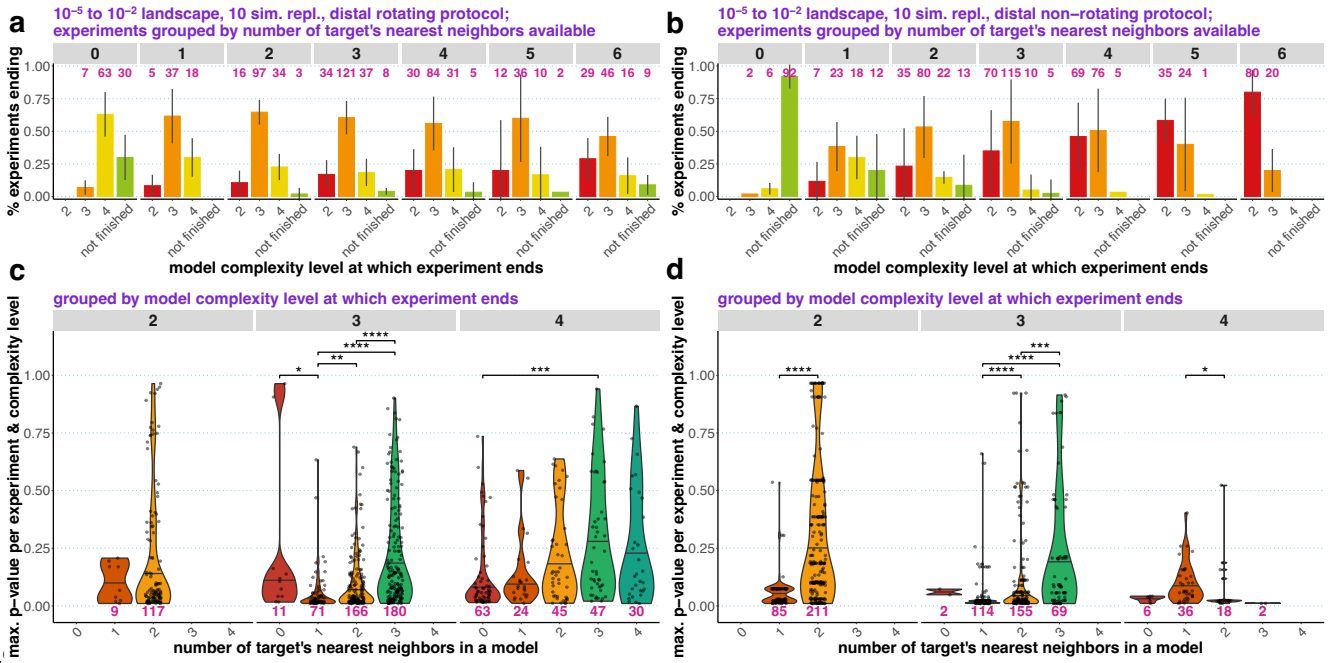


Figure 12. *qpAdm* performance in the case of systematic and symmetric landscape sampling, interpreted at the level of experiments. Results for the distal rotating (**a, c**) and distal non-rotating (**b, d**) protocols are shown. The protocols relied on the following composite feasibility criterion: *p*-value threshold at 0.01, EAF between 0 and 1. In panels **a** and **b**, distributions of experiments over model complexity levels at which they end are shown, and experiments are grouped by the number of target's nearest neighbors (demes from the 1st circle) available. The other proxy sources and "right" groups were taken from the 3rd circle around the target. The error bars show standard deviation calculated on simulation replicates. In panels **c** and **d**, each experiment is represented by only one fitting model with the highest *p*-value. We show distributions of *p*-values for these models (the scale on the left) as violin plots with medians, stratified by two variables: model complexity level at which an experiment ends (the scale on top), and number of target's nearest neighbors included in a model that was chosen to represent an experiment (the scale at the bottom). For each model complexity level, all pairwise comparisons of the distributions were performed using the two-sided non-paired Wilcoxon test, and only significant *p*-values (adjusted for multiple comparisons with the Holm method) are shown. The asterisks stand for the following significance levels: *, ≤0.05; **, ≤0.01; ***, ≤0.001; ****, ≤0.0001. The pink numbers in panels **a-d** stand for the number of experiments in the respective categories in all simulation replicates combined.

Naively, we would expect to see an opposite trend: *qpAdm* stops at more complex models when more true ancestry sources are available in a panel of populations tested. But in our experiments *qpAdm* results are misleading in all cases. When no or few close ancestry sources are available, the method tends to pick distant sources in the context of complex models. When (nearly) all closest ancestry sources are available, the method oversimplifies the history

of the target by picking just two or three sources. In other words, the results are not very informative about real history since sub-optimal sources are picked very often when the vicinity of the target is poorly sampled, and some optimal sources are very often dropped when the vicinity of the target and even the whole landscape is well-sampled.

We also ranked fitting models in each experiment by *p*-value and picked a model with the highest *p*-value to represent an experiment, as is often done in the literature (Narasimhan et al. 2019, Fernandes et al. 2020, Bergström et al. 2022, Lazaridis et al. 2022, Brielle et al. 2023, Carlhoff et al. 2023, Speidel et al. 2024). We expected *p*-values to grow with the number of target's nearest neighbors included in a model, and indeed we observe this effect, i.e., median *p*-values grow with the number of target's nearest neighbors (**Fig. 12c,d**). However, the distributions are all skewed towards low *p*-values and are largely overlapping, therefore only for experiments ending at three-way models (the largest group of experiments in our analysis) we observe significant differences between the *p*-value distributions for representative models including one, two, or three nearest neighbors (**Fig. 12c,d**). In the case of the rotating protocol and experiments ending at four-way models, just one statistically significant difference was observed (**Fig. 12c**). Similar results were obtained when inspecting all fitting models per experiment, not only those with the highest *p*-values (**Figs. S32, S33**).

Do estimated admixture fractions reflect gene flow intensities on SSL?

Previously we did not assess the accuracy of admixture fraction estimates. On SSL with non-uniform gene flows (**Fig. S13b,c**) in all directions it is not trivial to derive the most accurate predictor of admixture fractions, and the concept of admixture fractions itself becomes questionable. It is more intuitive and useful to conceptualize ancestry on SSL in the form of clouds of ancestor density (Grundler et al. 2024). With these caveats in mind, we assessed correlation between simulated source-to-target gene-flow intensities and EAF. For simplicity, we considered the systematic sampling scenario and models with the nearest neighbors only as proxy sources; direct ancestors were not considered as sources (**Fig. 13**). More precisely, we used the following simple admixture fraction predictor:

predicted $AF_i = \frac{S_{i\xrightarrow[gf]{}T}}{\sum_{i=1}^{n} S_{i\xrightarrow[gf]{}T}}$, where the numerator is a source-to-target gene-flow intensity in the last gene-flow epoch (lasting 153 generations, while proxy sources were sampled 100

generations in the past and targets at "present"; **Figs. 7** and **S12**), and the denominator is a sum of such intensities over all sources. Another analogous predictor can be constructed based on sums of source-to-target and target-to-source gene-flow intensities in the same epoch: $\frac{S_i \underset{gf}{\rightarrow} T + T \underset{gf}{\rightarrow} S_i}{\sum_{i=1}^{n}(S_i \underset{gf}{\rightarrow} T + T \underset{gf}{\rightarrow} S_i)}$.

The most notable pattern we observe when comparing EAF and AF calculated using the former simple predictor (**Fig. 13a**) comes from the linear algebra nature of *qpAdm* (Haak et al. 2015): symmetric and non-symmetric models produce EAF distributions of very different widths. In other words, only in the case of symmetric models nearly all EAF ∈ (0, 1). Thus, we conclude that EAF depend more strongly on STS angles than on gene-flow rates (**Fig. 13a**).

For highly non-symmetric two- or three-way models (on our SSL, all four-way models with nearest neighbors as sources are non-symmetric, as discussed above), namely those with min. STS angles = 60°, Pearson's correlation coefficients for predicted AF and EAF were non-significantly different from 0 over the whole range of EAF, and non-significant or low (up to 0.27) over EAF in the range (0, 1) (**Fig. 13a**). The linear correlation becomes much stronger in the case of ideally symmetric two- and three-way models: Pearson's correlation coefficients vary from 0.5 to 0.58, and all are significantly different from 0 with Bonferroni correction (**Fig. 13a**).

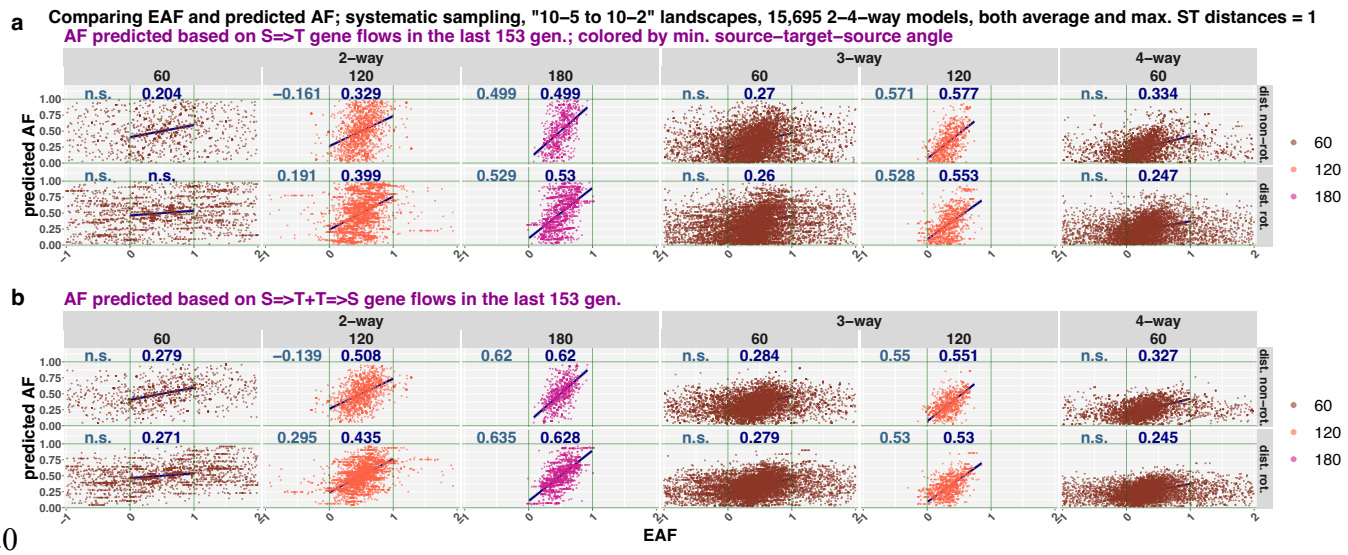


Figure 13. Scatterplots illustrating correlations between EAF and predicted AF on systematically sampled "10^{-5} to 10^{-2}" landscapes. Models composed of target's nearest neighbors only were considered, and AF were predicted based on simulated source-to-target gene-flow intensities or on sums of source-to-target and target-to-source gene-flow intensities (see the text for details). Each *n*-way model in these scatterplots is represented by *n* points, one per source. The results are stratified

by model complexity, by min. STS angle, and *qpAdm* protocol. Pearson's correlation coefficients were calculated for EAF vs. predicted AF: for EAF between 0 and 1 (the numbers in navy blue above the plots) or across the whole range of EAF (the numbers in slate gray; however, only models with EAF between -1 and 2 are shown). Non-significant correlation coefficients are marked as "n.s." Linear trend lines and 95% confidence intervals are also shown for EAF between 0 and 1.

However, even in the case of ideal symmetric models, the correlation is far from perfect (**Fig. 13a**). This may be a consequence of poor AF prediction since only direct source-to-target gene flows were considered. In other words, proxy sources in ideal symmetric models stand not only for themselves, but for equally sized sections of the landscape. Considering sums of reverse and forward ST gene-flow intensities (the second predictor above) allows to improve linear correlation between EAF and predicted AF, but for two-way models only (**Fig. 13b**). Very similar results were obtained using both predictors on randomized *qpAdm* models (**Fig. S34a,b**)

Geographical analysis of a large set of qpAdm models on real data

In this section we use the intuition and approaches from our analyses on SSL to interpret results of high-throughput *qpAdm* protocols applied to human archaeogenetic data from the Paleolithic to Bronze Age Siberia (Zeng et al. 2023). In short, we interpret each model (source-target configuration) tested with *qpAdm* on real data in a geographical sense by calculating great circle distances between centroids of source and target groups and min. STS angles on a sphere. Thus, we calculated all the model optimality metrics defined in the SSL context (**Box 2**). The analysis is based on ca. 422,000 two- to four-way models collected across 725 distal and proximal rotating experiments from Zeng *et al*. (some experiments in the cited study progressed far beyond four-way models, but to match our randomized protocol on SSL, these complexity classes were not considered). All the model metrics (EAF, *p*-values, and geographical coordinates of individuals) were taken from Zeng *et al*. (2023).

Placing the real *qpAdm* models in the spaces of max|*EAF–EF*| and *p*-values and juxtaposing them with the most realistic SSL simulated in this study ("10^{-5} to 10^{-2}", **Fig. S13**), we observe that distributions of models in these spaces are strikingly similar (**Fig. S35**). Model optimality metrics such as min. STS angles and maximal/average ST distances are correlated with max|*EAF–EF*| and (weakly) with *p*-values on both the real and simulated data (**Fig. S35, Table**

S13). It is not unexpected that on the simulated SSL data the correlations are stronger (**Table S13**) since, as mentioned above, on the real data we used just simple great circle distances, and not realistic travel distances on the landscape of Eurasia. Moreover, some groupings on the real data cover large areas of irregular shape (Zeng et al. 2023), adding to uncertainties about their centroids.

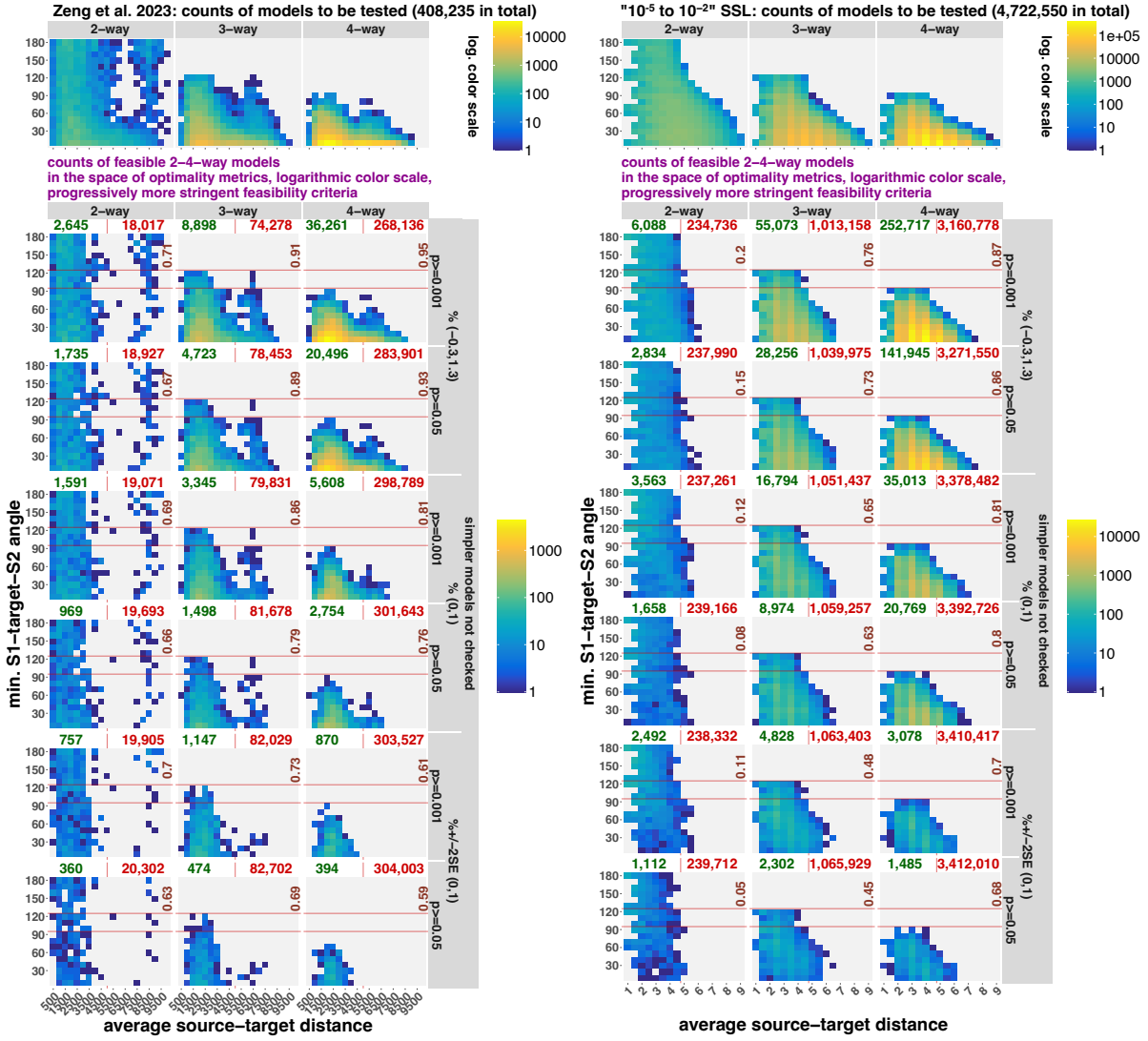


Figure 14. Comparing *qpAdm* models analyzed by the high-throughput protocol from Zeng *et al.* (2023) and results of the randomized *qpAdm* protocol on the "10^{-5} to 10^{-2}" landscapes by placing them in the spaces of optimality metrics (average ST distance vs. min. STS angle). Results are stratified by model complexity. The upper row of plots shows density of all models tested in this space, and the other rows show density of fitting models satisfying selected composite feasibility criteria listed in the captions on the right (6 of 36 criteria tested in this study, from the least stringent on top to the most stringent at the bottom). All the color scales are logarithmic. The space was divided into 20 or 18 bins on the x-axis (on the real and simulated data, respectively), and 18 bins on the y-axis (intervals [0°, 10°], (10°, 20°], and so forth). Spearman's correlation coefficients for counts of all models vs. fitting models in these bins are shown in each panel in brown (bins not populated by any models were not considered). The red horizontal lines mark the "ideal" min. STS angles for three- and four-way models (120° and 90°). The red and green numbers above each plot stand for counts of rejected and fitting

models in those analyses, respectively, and the red vertical lines mark the average ST distance (4.5) that equals the radius of the landscape on simulated data and the midpoint (5,000 km) between minimal and maximal average ST distances on the real landscape. The distances and angles on the real data (great circle distances in km and angles between two bearings) are based on centroids of groups calculated using an *R* implementation of the *Mean Center* tool from *ArcGIS Pro*. We considered ST distances ≤50 km to be negligible, and for such models min. STS angles were not defined (following the approach we applied to zero-length ST distances on the simulated data), excluding them from this analysis

In other words, there is just very minor enrichment for optimal models achieved on both real data and simulated "10^{-5} to 10^{-2}" SSL, in contrast to the enrichment level achieved for denser "10^{-3} to 10^{-2}" landscapes (**Fig. 9, Table S11**). To be precise, on the "10^{-5} to 10^{-2}" SSL the pre-*qpAdm* and post-*qpAdm* distributions of models in the optimality metric space are uncorrelated for the simplest, two-way, models (**Figs. 9, 14, Table S11**), but even this is not true in the case of the real data (**Fig. 14**).

At the next step, we explored how both real and simulated models populate the spaces of model optimality metrics (average ST distances vs. min. STS angles, **Fig. 14**). Since in both cases all possible models constructed from dozens of alternative sources were tested for each target, the distributions of models tested have the hallmarks of random samples: both are skewed towards highly asymmetric models (those with min. STS angles close to 0 and average ST distances approaching the landscape radius). And on both the real and simulated data, *qpAdm*, equipped with stringent composite feasibility criteria, does reject efficiently such models with average ST distances above a certain threshold (3,500 km for the real landscape). However, in the remaining densely populated area of the optimality metric space, the distributions of models passing the *qpAdm* "filter" and the starting distributions are similar as evidenced by Spearman's correlation coefficients (**Fig. 14**).

Discussion

Limitations of the AGS-based analyses in this study

We explored performance of various *qpAdm* protocols on a collection of simulated AGS histories of random topology and uniform complexity, where admixture history of target groups may vary from the simplest (no admixture) to complex (up to five admixture events in group's history). It is because of this research design and other limitations discussed below

1609 that our analysis is focused mostly on one performance metric: false discovery rate or FDR.
1610 We started from a collection of random simulated AGS histories of uniform complexity, tested
1611 all possible models of chosen complexity classes only (one- and two-way), considered two-
1612 way models supported by a *qpAdm* protocol (feasible or fitting models; as detailed in **Box 1**,
1613 our feasibility criteria included rejection of "trailing" one-way models), classified them
1614 manually into false and true positives according to a set of topological rules, and subjected
1615 them to further screening by the PCA and/or *ADMIXTURE* methods (**Suppl. text 4**). We did
1616 not attempt to classify rejections of two-way models by *qpAdm* or other methods into false
1617 rejections due to violations of the topological assumptions of *qpAdm* (**Fig. 1**) and true
1618 rejections when the true admixture history of the target does not fit a two-way model. This
1619 problem was deliberately left out since in the literature more attention is paid to
1620 interpretation of "fitting" than of rejected *qpAdm* models. Since archaeogenetic sampling is
1621 often very sparse, we also did not interpret fitting, but overly simple, models including correct
1622 sources as false in the AGS context (see the section "True positive models" above), but we
1623 highlight this problem in the SSL context where the simulated history is even more complex.
1624 In other words, we were interested in false detections of gene flow in the AGS context, but
1625 not in false omissions of detection.

1626 Another important caveat is that complexity of genetic history in the region and period of
1627 interest often remains unknown, and it is difficult to judge if a particular admixture graph
1628 complexity is adequate for simulating the real history (and if admixture graphs are
1629 appropriate models at all, see below). However, we have not explored *qpAdm* performance
1630 over a range of simulated admixture graph complexities, over a range of model feasibility
1631 criteria (except for those in **Table S3**), for models more complex than two-way, and have
1632 estimated FDR instead of FPR due to an important technical limitation: the process of model
1633 classification into true and false ones is difficult to automate fully since it requires careful
1634 interpretation of the simulated topology and simulated admixture proportions (this is
1635 illustrated by the case studies in **Fig. S4**).

1636

1637 ***Limitations of the SSL-based analyses in this study***

1638 While complex admixture models and a wide range of feasibility criteria were not explored
1639 on AGS histories above, the following issues were, conversely, not explored on SSL histories:

effects of differing amounts and quality of data (all the simulations were based on high-quality diploid data from 500-Mbp-sized genomes with mutation and recombination rates typical for humans), combining *qpAdm* with other methods (PCA, *ADMIXTURE*, f_3-statistics), and model competition *qpAdm* protocols. Some of these differences in research coverage stem from the fact that it is impossible to strictly delineate false and true admixture models on SSL since at the timescales we explore (up to 300 generations) all demes are connected by gene flows, directly or indirectly. Instead, it is possible to rank proxy sources by genetic or spatial proximity to the target and symmetry with respect to the target. Since hard-to-automate classification of *qpAdm* models into FP and TP can be avoided, SSL histories are more tractable analytically than complex AGS histories.

Since SSL simulations are computationally demanding, our analysis was restricted to relatively small landscapes composed of 64 demes, small samples per deme (three individuals), and just three sampling points in time. For the same reason, sampling was not randomized over >>3 points in time, and different time intervals between sampling points were not tested. More complex landscapes with rapid long-distance migrations, extinctions, repopulations, and changes in effective population size were also not explored for reasons of computational complexity. Thus, our SSL are still simple as compared to the real population history of humans and other species as we understand it. Ultimately, true isolation-by-distance landscapes should be explored instead of quantized deme-based landscapes, and the former can also be simulated by the *slendr* package (Petr et al. 2023) we used.

Another notable limitation discussed in the section "SSL histories: randomized sampling of landscapes" is our reliance on spatial source-target distances for measuring model optimality. F_{ST} (**Fig. S13a**) and haplotype-sharing statistics depend on both source-to-target and target-to-source gene-flows, and for that reason they are also not ideal as metrics for ranking proxy sources and were not used as such. Ideally, density clouds of ancestors (Grundler et al. 2024) should be inferred for each target and then used for ranking proxy sources or models or for prediction of EAF. However, this analysis is also computationally challenging in our setup aiming to explore a wide range of *qpAdm* models, protocols and landscape properties, and most prominent features of *qpAdm* performance on SSL such as effects of landscape sparsity and sampling density in target's vicinity can be explored even through the lens of our noisy spatial approximation for true ancestry density (see **Fig. S13d,e** and its discussion above).

1671 Moreover, our spatial analytical framework based on distances and angles can be directly
1672 applied to rich sets of *qpAdm* models from the literature (**Figs. 14, S35**) without inference of
1673 clouds of ancestry density which is even more complicated for real ancient DNA data.

1674

1675 ***AGS histories: general conclusions***

1676 Temporal stratification tested in this study and practiced in the literature is of two sorts: 1)
1677 most or all populations in the "right" set are sampled deeper in the past than those in the
1678 "left" set (non-rotating protocols); 2) a target group post-dates (or is as old as) all its proxy
1679 sources (distal protocols). Since complexity of AGS histories (number of admixture events)
1680 grows over time, temporal stratification of targets and proxy sources is a very efficient way
1681 of reducing FDR of *qpAdm* protocols (**Fig. 6, Table S1b**) since it allows to avoid situations
1682 where admixture history of proxy sources is more complex than that of the target, which
1683 increases the probability of encountering violations of the topological assumptions of *qpAdm*
1684 (**Fig. 1**). The distal rotating and non-rotating protocols invariably demonstrated FDR
1685 significantly lower than those of the rotating non-rotating and protocols without temporal
1686 stratification (**Fig. 6**). Both distal protocols demonstrated FDRs around 20% and not different
1687 significantly from each other, at least in the case of 3,000-Mbp-sized genomes (**Fig. 6**). In
1688 contrast, application of the proximal rotating protocol (**Box 1**) that can be summarized as
1689 "whatever is not on the right is on the left" without any attempt at temporal stratification of
1690 the "right" and "left" sets and of proxy sources and targets carries a risk of an FDR above 50%.
1691 Adding further levels of support (considering only models supported by PCA and/or an
1692 *ADMIXTURE* analysis) does not help to decrease FDR drastically in this case (**Table S2, Suppl.
1693 text 3**). Although the proximal model competition protocol (Narasimhan et al. 2019, Carlhoff
1694 et al. 2021, Zhang et al. 2021, Maróti et al. 2022, Brielle et al. 2023, Lee et al. 2023) was not
1695 tested on multiple simulation or subsampling replicates (**Table S5**), we note that it
1696 demonstrated FDR values higher than those of the distal non-rotating protocol (**Table S4**).

1697 The proximal rotating protocol is an extreme example of *qpAdm* protocols that is rarely
1698 encountered in the archaeogenetic literature (Carlhoff et al. 2021, Oliveira et al. 2022, Zeng
1699 et al. 2023) but serves as a reference point in our analysis. Other protocols such as distal
1700 rotating (e.g., Narasimhan et al. 2019, Librado et al. 2021, Bergström et al. 2022, Lazaridis et
1701 al. 2022, Taylor et al. 2023, Zeng et al. 2023, Speidel et al. 2024), distal non-rotating (e.g.,

1702 Haak et al. 2015, Mathieson et al. 2015, Lazaridis et al. 2016, Antonio et al. 2019, Sikora et al.
1703 2019, Marcus et al. 2020, Yang et al. 2020, Papac et al. 2021, Yaka et al. 2021, Patterson et
1704 al. 2022), and proximal model competition (e.g., Narasimhan et al. 2019, Carlhoff et al. 2021,
1705 Zhang et al. 2021, Maróti et al. 2022, Brielle et al. 2023, Lee et al. 2023) are often used in
1706 practice, and median FDRs of these three classes of protocols on our simulated data ranged
1707 from 16.4% to 45.8% across simulation parameter combinations (**Fig. 6, Table S5**). These FDRs
1708 for best-performing *qpAdm* protocols are high but should not be over-interpreted since they
1709 are expected to depend on the complexity of simulated histories, but only one graph
1710 complexity level was tested, that is 13 groups and 10 admixture events. Thus, it is hard to
1711 extrapolate this aspect of our results to real analyses and predict FDR levels on real data.

1712 Another way of radically improving FDR of *qpAdm* protocols applied to AGS histories is
1713 combining *qpAdm* with an unsupervised *ADMIXTURE* analysis (**Suppl. text 4**), however we
1714 have not tested *ADMIXTURE* on low-quality data, and supervised *ADMIXTURE* protocols were
1715 not tested either. Unsupervised *ADMIXTURE* has known problems even on high-quality data
1716 and AGS histories (Lawson et al. 2018). That is why this recommendation is still tentative. f_3-
1717 statistic is a simple method for proving that a population is admixed, and it demonstrated
1718 FDR values much lower (6%, see **Suppl. text 3**) than those of standalone *qpAdm* protocols,
1719 but f_3-statistics are applicable only to recent admixture events and/or populations of large
1720 effective size since post-admixture drift on the target lineage obscures the signal (Patterson
1721 et al. 2012, Peter 2022). Moreover, calculating f_3-statistics for a target composed of a single
1722 pseudo-haploid individual is impossible since a heterozygosity estimate is required (Maier et
1723 al. 2023), and such singleton groups are common in archaeogenetic studies. Researchers
1724 should also be aware that f_3-statistics are defined on unrooted trees, and that may lead to
1725 rare but strong false signals of admixture (**Fig. S4f**).

1726

1727 ***If population history conforms to SSL, qpAdm results are probably misleading in practice***

1728 But there exist problems much more profound than those described above for random AGS
1729 histories: our results (considering the caveats outlined above) suggest strongly that the
1730 interpretability and usefulness of *qpAdm* protocols are radically different for AGS (see also
1731 Williams et al. 2024) and SSL histories and for varying levels of SSL sampling density
1732 ("sparsity"). Although it is very often assumed that the genetic history of humans and many

other species is reasonably approximated by complex admixture graphs (see a short review in Maier et al. 2023), this assumption remains unproven. Strictly speaking, it remains unknown if admixture history of humans, wolves (Bergström et al. 2022), and horses (Librado et al. 2021, Taylor et al. 2023, Librado et al. 2024), to cite some recent examples from the archaeogenetic literature, is approximated reasonably well by admixture graphs composed of isolated populations and episodes of admixture, or rather by two-dimensional stepping-stone models, or a combination thereof.

In the case of AGS histories, as suggested by our results, temporal stratification of targets and proxy sources is a simple approach for obtaining relatively accurate admixture models for a randomly picked target (at least for simple, two-way, models that are broadly adequate for the complexity of histories we simulated). In contrast, in the context of SSL we found results of such *qpAdm* protocols to be generally poorly informative about topology of gene flows on SSL and misleading. The following aspects of *qpAdm* protocols are problematic for SSL demographies: 1) composite feasibility criteria (**Box 1**) combining conditions on EAF and fixed relatively stringent *p*-value thresholds such as 0.01 or 0.05, or ranking prioritizing highest-*p*-value models for interpretation (Narasimhan et al. 2019, Fernandes et al. 2020, Bergström et al. 2022, Lazaridis et al. 2022, Brielle et al. 2023, Carlhoff et al. 2023, Speidel et al. 2024); 2) the approach we term "*qpAdm* experiments" when a researcher explores progressively more complex models for a target and stops at the simplest model satisfying a composite feasibility criterion. First, due to gene flows in all directions, correlation between *p*-values and distance- and angle-based model optimality metrics on SSL is very poor, but EAF correlate with both types of optimality metrics better (**Fig. 8, Table S8**; the situation is different for models of maximal complexity which makes sense on our SSL under systematic landscape sampling, see **Fig. 11**). In addition, our analysis on AGS histories showed that EAF are stable across simulation/subsampling replicates having the same simulated AF, while *qpAdm p*-values vary by many orders of magnitude (**Fig. S7**). Thus, considering both AGS- and SSL-based results we conclude that reliance on *qpAdm p*-value thresholds, and especially *p*-value-based model sorting where slight *p*-value differences may affect if a model receives interpretation or is ignored (Narasimhan et al. 2019, Fernandes et al. 2020, Bergström et al. 2022, Lazaridis et al. 2022, Brielle et al. 2023, Carlhoff et al. 2023, Speidel et al. 2024), are problematic.

1763 Another problematic aspect of the composite feasibility criteria common in the literature is
1764 the fact that EAF lie largely within (0, 1) and correlate with simulated gene-flow intensities
1765 only for sources located symmetrically around the target, but for asymmetric sources they
1766 vary widely both outside and, complicating the matters even further, inside (0, 1) (**Fig. 13**).
1767 Although this property of AF estimation is trivial from the mathematical perspective, it is not
1768 recognized in practice and thus may often confound interpretation of *qpAdm* results on SSL
1769 (if the interpretation of rejected and fitting source configurations is not explicitly spatial). For
1770 instance, a very distant proxy source may be chosen instead of a much closer one because
1771 the latter configuration of sources happens to be slightly less symmetric and happens to yield
1772 a several orders of magnitude lower *p*-value due to complex patterns of gene flow on SSL (see
1773 two examples in **Fig. S26**). And highlighting a distant source and rejecting a close source is, in
1774 principle, equivalent to a claim of long-distance migration to the target deme.

1775 Due to these problems with composite feasibility criteria, the accuracy of fitting two-way
1776 *qpAdm* models on SSL is much higher than that of more complex models such as three- and
1777 four-way (**Figs. 8, 9**, and **S20b, Tables S7, S8, S11**), but the latter could have been more
1778 informative about complex gene-flow topologies of SSL and more suggestive of complex
1779 genetic history than the former (see two case studies in **Fig. S26**). Among even more complex
1780 six-way models we explored under systematic landscape sampling, efficient discrimination of
1781 the worst models from ideal ones is possible with a *p*-value threshold and the distal non-
1782 rotating (but not rotating!) protocol, but cleanly separating models less different in optimality
1783 is probably impossible (**Fig. 11**). Although the largest number of admixture events in the
1784 history of a population drawn from our AGS simulations is 5, the average is 1.17 and median
1785 is 1 (**Table S1a**), which corresponds to ≈2 ancestry sources distinguishable under given
1786 population sampling. In contrast, node degree (number of nearest neighbors exchanging
1787 migrants with a given deme) on our SSL varies from 3 to 6, average degree is 5.1, and median
1788 degree is 6 (**Fig. 7**). For this reason, we consider two-way models to be an acceptable result
1789 for AGS histories but a simplistic result for SSL histories. Although we have not explored FDRs
1790 of three-way and more complex models in the AGS context, considering our results in the SSL
1791 context, it is highly likely that complex fitting models in that context are also less accurate
1792 than two-way ones.

1793 For many species, geographical regions and historical periods, archaeogenetic sampling is
1794 very sparse and may be random with respect to population density of ancient individuals in
1795 those regions/periods. Moreover, a general quantitative algorithm for selection of optimal
1796 "right" groups for *qpAdm* analyses has never been published, and "right" groups are usually
1797 chosen based on qualitative criteria (Harney et al. 2021) or borrowed from study to study.
1798 Motivated by these considerations and by the occurrence in the literature of *qpAdm*
1799 protocols exploring thousands of alternative models for dozens of target groups (e.g.,
1800 Narasimhan et al. 2019, Lazaridis et al. 2022, Zeng et al. 2023; the latter study was re-visited
1801 here), we aimed to find expectations for *qpAdm* performance under completely random
1802 sampling of "right", target, and source demes from SSL, and our central analysis relies on
1803 thousands of such random sets of "right" demes or random rotating sets of two sizes (**Figs.**
1804 **8-10**). We showed that in this situation robust rejection of highly asymmetric complex models
1805 with distant sources (with average ST distance greater or approximately equal to the
1806 landscape radius) is possible only for very specific combinations of landscape sparsity and
1807 model feasibility criteria, but models of that sort are by far the most common in random sets
1808 of demes drawn from both simulated landscapes (**Figs. 9** and **S20b, Table S11**) and real
1809 archaeogenetic data (**Fig. 14**). Furthermore, this optimal combination of sparsity and
1810 feasibility criteria we found (EAF $\pm$ 2 SE $\in$ (0, 1) and the "10^{-3} to 10^{-2}" landscapes) is very
1811 uncommon in practice: restricting both "right" and "left" deme sets to a small sub-continental
1812 region and requiring that EAF $\pm$ 2 SE $\in$ (0, 1) are very rare approaches, to our knowledge. The
1813 "10^{-5} to 10^{-2}" landscapes resemble real *qpAdm* setups much more in distributions of F_{ST} (**Fig.**
1814 **S13**) and in distributions of models in the max|*EAF*–*EF*| vs. *p*-value coordinates (Fig. S35), and
1815 they demonstrated much poorer *qpAdm* performance in this respect (**Table S11**).

1816 Since explicitly spatial treatment of *qpAdm* results is rare in the literature, inefficient rejection
1817 of highly non-optimal models may represent a serious problem and a source of many claims
1818 of rapid long-distance migrations that abound in the literature. In other words, random sets
1819 of demes drawn from SSL are so enriched in highly non-optimal models (see the model
1820 density hotspots in figures like **Fig. 14** and note the logarithmic color scale) that even slightly
1821 imperfect rejection of this model class by a *qpAdm* protocol may flood small samples of ideal
1822 symmetric models (strictly speaking, ideal within their complexity class) that exist in random
1823 deme sets, and that is what we observe when classifying *qpAdm* experiments ending at the

levels of three- or four-way models into those with “desirable” and “undesirable” outcomes (**Figs. 10, S23, S24**).

Moving from interpreting individual models to this level of experiments (all models of varying complexity for the same target and given the same “right” or rotating set) which are usually stopped as soon as a simplest fitting model is found (see, e.g., Narasimhan et al. 2019, Lazaridis et al. 2022, Zeng et al. 2023), we see that *qpAdm* protocols generally find either accurate, but simplistic models (one- to three-way models with sources close to the target and approximately opposite each other), or relatively inaccurate complex models (**Figs. 10**, **S23**, and **S26**). Systematic sampling of proxy source and “right” demes in circles around the target deme, although marginally improving *qpAdm* performance at the level of models (**Figs. S27, S28**) if an optimal target-“right” distance is found (**Figs. S25, S29, S30, Table S12**), demonstrates similar problems at the level of experiments: when the vicinity of the target is poorly sampled *qpAdm* tends to report very complex inaccurate models, but when the vicinity of the target or even the whole landscape is well-sampled, the *qpAdm* experiment approach tends to find overly simple two- or three-way models (**Figs. 12, S33**).

There is also an important difference between *qpAdm* performance in the context of AGS and SSL histories, and that is the very subtle influence of temporal stratification of targets and sources on *qpAdm* results on SSL (**Fig. S21, Table S8**). This difference is probably because complexity of AGS histories grows in time, but complexity of SSL histories we simulated is at equilibrium in the chronological range of our analysis (from 0 to 300 generations into the past). However, on SSL histories there is a noticeable influence of the temporal stratification of “right” and “left” groups (**Figs. 10-12**).

As mentioned above, failure to reject a model with at least one distant source (**Figs. 9** and **S20b**) and simultaneous rejection of a much more optimal model for the same target (as is common for *qpAdm* experiments aimed at recovery of complex models, see **Figs. 10** and **S23**, and two specific examples in **Fig. S26**) may be often equivalent in practice to supporting a long-distance migration in favor of a local gene-flow landscape. And stopping the analysis at simple models may (**Figs. 10, 12, S23,** and **S33**) favor thinking about history in terms of simple graphs instead of SSL. Thus, in the context of SSL it is probably counterproductive to search for a single best model complexity level and a single best *qpAdm* model for a target. Instead, considering sets of many fitting models of varying complexity for a target (see **Fig. 11**),

preferably with relaxed *p*-value cutoffs <<0.01, is probably a much safer and informative approach, and that approach may be considered a course approximation for the inference of ancestor density clouds. This is especially true since SSL sparsity is inherently hard to estimate, and choice of optimal feasibility criteria and optimal model complexity range on SSL depends on sparsity in a profound way (**Figs. 8** and **9, Table S11**). For example, assuming that population history conforms largely to SSL, "distal *qpAdm* modelling" as used in the literature (Haak et al. 2015, Mathieson et al. 2015, Mathieson et al. 2018, Lazaridis et al. 2016, Antonio et al. 2019, Narasimhan et al. 2019, Sikora et al. 2019, Marcus et al. 2020, Yang et al. 2020, Librado et al. 2021, Papac et al. 2021, Yaka et al. 2021, Bergström et al. 2022, Lazaridis et al. 2022, Patterson et al. 2022, Taylor et al. 2023, Zeng et al. 2023; here the word "distal" is applied more broadly than elsewhere in our study) likely falls in the category of sparsely sampled continent-wide landscapes (**Fig. S13**) since worldwide "right" groups are combined with potential sources distant in time and/or space from the target, and all our analyses concurred that at a certain level of sparsity exploring any models above two-way generates very noisy results, and at even higher levels of sparsity the method stops working completely (at least for our simulations where all demes arise via a multifurcation). Our geographical analysis of a large set of *qpAdm* models tested on real human archaeogenetic data (Zeng et al. 2023) revealed patterns strikingly similar to those observed on sparse SSL (**Figs. 14, S35**), namely poor enrichment of sets of feasible *qpAdm* models for models optimal in the SSL context. This lack of enrichment for optimal models has two potential explanations. First, it is possible that the genetic history of humans in the Paleolithic to Bronze Age Siberia (Zeng et al. 2023) was dominated by long-distance migrations and not by an isolation-by-distance landscape. However, our exploration of simulated SSL suggests another alternative: we are dealing with a sparsely sampled landscape which is not very suitable for high-throughput *qpAdm* protocols (**Table S11**).

Considering all the problems highlighted above, and the inherent uncertainty about types of genetic history we are dealing with, we recommend that it is safest to use the *qpAdm* method in controlled conditions, when relationships among populations are understood well enough to exclude violations of the topological assumptions (**Fig. 1**), when radiocarbon or context dates of ancient populations are reliable and allow accurate temporal stratification, or when sets of potential proxy sources are well-constrained based on archaeological or historical

scholarship: see, for instance, the earliest publications where the *qpAdm* method was employed (Haak et al. 2015, Mathieson et al. 2015) and some recent studies (e.g., Marcus et al. 2020, Papac et al. 2021, Yaka et al. 2021, Changmai et al. 2022a, Changmai et al. 2022b, Patterson et al. 2022). Obviously, the amount of new information that the *qpAdm* method provides in these conditions is unique for each study and may be low. In contrast, high-throughput *qpAdm* screens aiming to find optimal models for a wide variety of targets (e.g., Narasimhan et al. 2019, Bergström et al. 2022, Lazaridis et al. 2022, Zeng et al. 2023, Speidel et al. 2024) are not recommended, especially in extremely sparsely sampled regions/periods, such as the Paleolithic in the case of human population history (e.g., Posth et al. 2023, Zeng et al. 2023, Villalba-Mouco et al. 2023, Allentoft et al. 2024), Paleolithic to Neolithic Siberia (Sikora et al. 2019, Zeng et al. 2023, Gill et al. 2024), ancient South (Narasimhan et al. 2019, Kerdoncuff et al. 2024) and Southeast Asia (Carlhoff et al. 2021, Oliveira et al. 2022, Carlhoff et al. 2023; *qpAdm* played a central role in some of those studies).

In conclusion, a crucial problem is as follows: *qpAdm* on SSL (with more or less non-uniform gene-flow intensities, **Fig. S13b,c**) generates results that very often do not look suspicious but are either overly simple or inaccurate (favoring very distant sources over available spatially close sources). For instance, finding a two-way fitting model and rejecting all alternative models of the same complexity can be interpreted in several ways: 1) as a true or (much less likely for distal protocols) a false model in the context of a relatively simple AGS history; 2) as a suboptimal model (with large ST distances and/or small STS angles) in the context of a relatively poorly sampled SSL and/or unsuitable feasibility criteria; 3) as an optimal but overly simple model in the context of a densely sampled SSL with well-chosen feasibility criteria. We contend that this variety of potential interpretations makes the method (strictly speaking, the protocols aiming to find and interpret one or few “best-fitting” models for a target in large sets of alternative models) hardly suitable for one of its chief use cases in the literature, namely as a method supplying main or supporting evidence for long-distance migrations in exploratory studies.

1916 ## Methods

1917 ***AGS histories: simulations***

1918 For simulating genetic data, we used *msprime v.1.1.1* which allows accurate simulation of
1919 recombination and of multi-chromosome diploid genomes relying on the Wright-Fisher
1920 model (Nelson et al. 2020, Baumdicker et al. 2022). We simulated 3, 10, or 30 diploid
1921 chromosomes (each 100 Mbp long) by specifying a flat recombination rate (2×10^{-8} per nt per
1922 generation) along the chromosome and a much higher rate at the chromosome boundaries
1923 ($\log_e 2$ or ~0.693 per nt per generation, see
1924 https://tskit.dev/msprime/docs/stable/ancestry.html#multiple-chromosomes). A flat
1925 mutation rate, 1.25×10^{-8} per nt per generation (Scally & Durbin 2012), and the binary
1926 mutation model were used. To maintain the correct correlation between chromosomes, the
1927 discrete time Wright-Fischer model was used for 25 generations into the past, and deeper in
1928 the past the standard coalescent simulation algorithm was used (as recommended by Nelson
1929 et al. 2020).

1930 AGS histories of random topology including 13 populations and 10 pulse-like admixture
1931 events were generated using the *random_admixturegraph* and *random_sim* functions from
1932 the *ADMIXTOOLS 2* package
1933 (https://uqrmaie1.github.io/admixtools/reference/random_sim.html), which produced
1934 scripts for running the *msprime v.1.1.1* simulator. Demographic events were separated by
1935 date intervals ranging randomly between 20 and 120 generations, with an upper bound on
1936 the graph depth at 800 generations (or ca. 23,000 years in the case of humans). In other sets
1937 of simulations, all the dates were scaled up 3.75 times, with an upper bound on the graph
1938 depth at 3,000 generations (or 87,000 years in the case of humans). To be more precise,
1939 demographic events were not placed in time entirely randomly, but were tied to one or few
1940 other events of the same "topological depth" within the graph, as illustrated by ten examples
1941 of simulated topologies in **Fig. S1**. The same principle was applied to sampling dates, which
1942 were tied to other demographic events such as divergence and admixture of other
1943 populations. These restrictions were used to ensure topological consistency of random
1944 graphs.

1945 Ten diploid individuals with no missing data were sampled from each population at "leaves"
1946 of the graph. Effective population sizes were constant along each edge and were picked
1947 randomly from the range of 1,000 – 10,000 diploid individuals. Admixture proportions for all
1948 admixture events varied randomly between 10% and 50%. This setup generates groups
1949 sampled at widely different dates in the past or, in other words, located at various genetic
1950 distances from the root. Alternatively, all terminal branches were extended to the "present"
1951 of the simulation and sampled at "present", keeping their respective effective population
1952 sizes and topological relationships unchanged. Thus, another set of simulations was
1953 generated for the same topologies, where groups were more drifted with respect to each
1954 other (see F_{ST} distributions in **Fig. S3**).

1955 In summary, five sets of independent simulations ("simulation setups") differing by the
1956 amount of data generated and by population divergence metrics were performed for a set of
1957 40 random admixture graph topologies (see an overview in **Fig. 3**):

1958 1) three 100-Mbp-sized chromosomes; groups sampled at different points in time; maximal
1959 simulated history depth at 800 generations (10 simulation replicates, for median number of
1960 polymorphic sites see **Fig. S2**);

1961 2) ten 100-Mbp-sized chromosomes; groups sampled at different points in time; maximal
1962 simulated history depth at 800 generations (10 simulation replicates);

1963 3) thirty 100-Mbp-sized chromosomes; groups sampled at different points in time; maximal
1964 simulated history depth at 3,000 generations (10 simulation replicates);

1965 4) three 100-Mbp-sized chromosomes; groups sampled at different points in time; maximal
1966 simulated history depth at 3,000 generations (one simulation replicate);

1967 5) three 100-Mbp-sized chromosomes; all terminal branches extended to the "present" of the
1968 simulation and sampled at that point; maximal simulated history depth at 800 generations
1969 (one simulation replicate).

1970 To create more realistic datasets, we performed randomized subsampling of polymorphic
1971 sites and individuals (replicates no. 1 of the first, second, and third simulation setups were
1972 used for this, see the list above). First, we randomly sampled alleles at heterozygous sites,
1973 creating pseudo-haploid data. Then we introduced missing data by randomly selecting a
1974 missing rate between 5% and 95%, followed by randomly selecting sites according to the

1975 missing rate. This site subsampling was repeated for each individual independently. Lastly,
1976 we randomly sampled *n* (from 1 to 10) individuals from each population independently. The
1977 subsampling procedure described above was conditioned on the number of sites
1978 polymorphic in the set of 13 simulated populations and was repeated until a subsampling
1979 replicate with more than 20,000 (for 300-Mbp-sized genomes) or 66,000 such sites (for 1,000-
1980 and 3,000-Mbp-sized genomes) was obtained (**Fig. S2**). We generated 10 independent
1981 subsampled replicates for each topology and simulation setup (1,200 replicates in total).
1982 Polymorphism data in the *EIGENSTRAT* format were generated from the tree sequences using
1983 the *TreeSequence.genotype_matrix* function (https://tskit.dev/tskit/docs/stable/python-
1984 api.html#tskit.TreeSequence.genotype_matrix) and used for all subsequent analyses (*f*-
1985 statistics and *qpAdm*, PCA, *ADMIXTURE*).

1986

1987 ***AGS histories: calculating f-statistics***

1988 For all the work on *f*-statistics and *qpAdm*, the *ADMIXTOOLS 2* software package (Maier et al.
1989 2023) was used. For diploid SNP sets without missing data, we first calculated all possible f_2-
1990 statistics for 4-Mbp-sized genome blocks (with the "*maxmiss=0*",
1991 "*adjust_pseudohaploid=TRUE*", and "*minac2=FALSE*" settings) and then used them for
1992 calculating f_3- and f_4-statistics as linear combinations of f_2-statistics and for testing *qpAdm*
1993 models using the *qpadm* function in *ADMIXTOOLS 2*
1994 (https://uqrmaie1.github.io/admixtools/) under default settings. Harney *et al*. (2021) showed
1995 that *qpAdm* is robust to changes in genome block size except for extreme values: 0.0001 and
1996 1 Morgans (0.05 Morgans is a recommended default value, which corresponds to ~5 Mbp in
1997 the case of humans). Inferred admixture proportions in our analysis were not constrained
1998 between 0 and 1. For pseudo-haploid SNP sets with missing data and uneven group sizes, the
1999 *qpadm* function was applied to genotype files directly, with the "*allsnps=TRUE*" setting. In
2000 other words, f_4-statistics utilized by *qpAdm* and f_3-statistics were calculated off the genotype
2001 files without intermediate f_2-statistics, and removal of missing data was done for each
2002 population quadruplet or triplet separately. This setup is often used in the literature in the
2003 presence of missing data (e.g., Harney et al. 2018, Harney et al. 2019, Narasimhan et al. 2019,
2004 Lazaridis et al. 2022).

2005

2006 ***AGS histories: qpAdm protocols***

2007 *QpWave* tests were performed on sets of 13 groups divided randomly into 2 "left" and 11
2008 "right" groups, testing all possible bisections of this form. *QpAdm* was applied to the same
2009 sets of 13 groups divided randomly into 3 "left" and 10 "right" groups, testing all possible
2010 bisections of this form for all possible target groups in "left" sets (858 such models per
2011 simulated history). This proximal rotating protocol was applied to all simulation setups.
2012 Subsequent work was focused only on feasible *qpAdm* models defined as follows: 1) *p*-values
2013 calculated by *qpWave* for one-way models "target = proxy source$_1$", "target = proxy source$_2$",
2014 and "proxy source$_1$ = proxy source$_2$" are all below 0.01; 2) in the case of the two-way model
2015 "target = proxy source$_1$ + proxy source$_2$", EAF $\pm$ 2 standard errors are between 0 and 1; 3) the
2016 *p*-value calculated by *qpAdm* for the two-way model $\geq$0.01. For exploring performance of the
2017 distal rotating protocol, feasible two-way *qpAdm* models were simply filtered according to
2018 sampling dates of target groups and proxy sources. If target group's sampling date was equal
2019 to or smaller than sampling dates of both proxy sources, such a model was considered distal.

2020 In the non-rotating protocol, for each simulated admixture graph six oldest sampled groups
2021 were selected as a fixed "right" set (ties in sampling dates were resolved in alphabetical order;
2022 these "right" sets remained unchanged for a given topology across all independent
2023 simulations), and for the remaining seven groups all possible one-way and two-way
2024 admixture models were tested (105 models), applying the same composite feasibility
2025 criterion that was used above for the rotating protocol. This is the proximal non-rotating
2026 protocol, and alternatively we focused on distal admixture models only (distal non-rotating
2027 protocol).

2028 In the proximal model competition protocol, subsequent analysis was focused on targets for
2029 whom two or more alternative *qpAdm* models emerged as feasible at the first step. For each
2030 target, alternative proxy sources were pooled and rotated between the "left" and "right" sets,
2031 testing only the models that emerged as feasible at the first step and applying the composite
2032 feasibility criterion (*p*-value $\geq$0.01, EAF ± 2 SE are between 0 and 1). Rotation of alternative
2033 proxy sources was performed in two alternative ways: "whatever is not on the left is on the
2034 right", or placement of alternative sources in the "right" set one by one. In the latter case
2035 several "right" sets were tested for each model, and the model was considered supported by
2036 the model competition protocol only if it was not rejected under any of these "right" sets (the

2037 latter protocol follows Maróti et al. 2022). If only one model was feasible for a target, such a
2038 model was evaluated as passing the model competition procedure. A distal model
2039 competition protocol was not tested in this study.

2040 For testing statistical significance of differences in FDR between *qpAdm* protocols, the
2041 following approach was used. FDR was calculated either on low-quality data for 10 random
2042 site/individual subsampling replicates derived from simulation replicate no. 1 (simulation
2043 setups no. 1 to 3) or on high-quality data for 10 independent simulation replicates (simulation
2044 setups no. 1 to 3). Comparisons of six *qpAdm* protocols (rotating and non-rotating;
2045 considering exclusively proximal, exclusively distal models, or both) were performed on these
2046 six sets of replicates using the two-sided paired Wilcoxon test (implemented as the
2047 *pairwise_wilcox_test* function in the *rstatix* R package v. 0.7.2). Comparisons of the same
2048 *qpAdm* protocols on simulations with different amounts or quality of data were performed
2049 using the two-sided non-paired Wilcoxon test since simulation replicates were independent
2050 unlike alternative *qpAdm* protocols applied to the same data (**Fig. 6**). In all applications of the
2051 Wilcoxon test for various purposes (**Figs. 6, 12, S32, Table S1b**), *p*-values adjusted for multiple
2052 testing with the Holm method are presented.

2053 For comparing numerical results of *ADMIXTOOLS 2* v. 2.0.6 (released in 2024) with those of
2054 classic *ADMIXTOOLS* v. 7.0 (*qpAdm* v. 1201, released in 2020), we used the most realistic AGS
2055 simulations we generated, that is low-quality pseudohaploid data for 3,000-Mbp-sized
2056 genomes: 10 AGS topologies, with one simulation and one data subsampling replicate per
2057 topology. All possible two-way models were tested with the rotating protocol. *ADMIXTOOLS*
2058 *2* was run with the "*fudge_twice=TRUE*" argument which makes the results more similar
2059 numerically to those of classic *ADMIXTOOLS*. Since the data was of low quality
2060 (pseudohaploid, with high missing rates, and with some populations composed on one
2061 individual), we used the "*allsnps=TRUE*" setting in *ADMIXTOOLS 2* and the identical
2062 "*useallsnps: YES*" setting in classic *ADMIXTOOLS* (all the other settings were default ones). If
2063 "*useallsnps*" is set to "*YES*", the latest *ADMIXTOOLS* v. 8.0.1 (*qpAdm* v. 2050, released in 2024)
2064 automatically calls *qpfstats*
2065 (https://github.com/DReichLab/AdmixTools/blob/master/README.qpfstats), an algorithm
2066 for handling low-quality data that is substantially different from that of the "*allsnps=TRUE*"
2067 setting in *ADMIXTOOLS 2* (https://uqrmaie1.github.io/admixtools/reference/qpfstats.html).

2068 To perform a fair comparison, we used *ADMIXTOOLS* v. 7.0 instead of the latest version v.
2069 8.0.1.

2070 The results produced by these two packages are not identical, but highly correlated across
2071 8,580 unique models we tested on low-quality data (**Table S14**). Pearson's correlation
2072 coefficient is 0.99 for *p*-values (this analysis prioritizes high *p*-values), 0.91 for $\log_{10}(|EAF_1|)$,
2073 0.73 for $\log_{10}$(*p*-values), and 0.79 for $\log_{10}$(SE). Spearman's coefficients are slightly higher for
2074 the logarithmic values (all coefficients are statistically significant with two-sided *p*-values =
2075 0). The weaker correlation seen for logarithmic *p*-values is in line with the high variability in
2076 logarithmic *p*-values calculated by *ADMIXTOOLS 2* on simulation replicates (**Fig. S7**).

2077

2078 ***AGS histories: classifying two-way admixture models into false and true positives***

2079 Since the simulated admixture graph topologies were complex and random, target groups
2080 modelled with *qpAdm* had complex admixture history in some cases, being a part of gene
2081 flow networks. In this context it is hard to draw a strict boundary between true and false
2082 admixture models composed of a target and only two proxy sources. Two-way admixture
2083 models were considered false only if at least one of the following criteria was satisfied
2084 (considering only graph topologies and admixture proportions):

2085 1. The target and at least one of the proxy sources are simulated as strictly cladal (**Fig. 4, Fig.**
2086 **S4a**). In this case the target may either be unadmixed, or it may have experienced gene
2087 flows earlier in its history that do not break its cladality with one of the proxy sources.

2088 2. A proxy source lineage is a recipient of gene flow from the target lineage (after the last
2089 admixture event in target's history), possibly mediated by other lineages (**Fig. 4, Fig. S4a,**
2090 **b**). In other words, the incorrect proxy source is a descendant of the target lineage, i.e.,
2091 the expected gene flow direction is reversed.

2092 3. A proxy source does not represent any true source. In other words, it is symmetrically
2093 related to all true sources of ancestry in the target (**Fig. S4c, d**). Alternatively, both proxy
2094 sources represent the same true source, and are symmetrically related to all the other
2095 true sources (**Fig. S4e**).

2096 4. A proxy source shares genetic drift with the corresponding true source that is not shared
2097 by the second proxy source (and the same is true for the other proxy source and another
2098 true source, i.e., condition no. 3 above is not satisfied), however less than 40% of its
2099 ancestry is derived from the true source (**Fig. S4a, d**).

2100 5. The target gets a gene flow from a deep-branching source not represented by any
2101 sampled population, and an inappropriate proxy source is included in the fitting model
2102 (**Fig. S4f, g**).

2103 We illustrate these topological rules with eight case studies of FP and feasible *qpAdm* models
2104 in **Fig. 4** and **Fig. S4a-g**. Two-way models for targets whose population history is best
2105 approximated with three-way and more complex models were considered as TP if they
2106 included source proxies (that do *not* satisfy the criteria above) for at least two of three or
2107 more true ancestry sources (see three case studies in **Fig. S4k-m**).

2108

2109 ***AGS histories: principal component analysis***

2110 PCA was performed for one simulation replicate per simulation setup. Only high-quality data
2111 was used, and all individuals sampled from a simulation were co-analyzed. Prior to the
2112 analysis, linked sites were pruned with *PLINK v.2.00a3LM* (Chang et al. 2015) using the
2113 following settings: window size, 2000 SNPs; window step, 100 SNPs; r^2 threshold = 0.5
2114 (argument "--indep-pairwise 2000 100 0.5"). PCA was also performed using *PLINK v.2.00a3LM*
2115 under default settings, calculating 10 PCs. Interactive three-dimensional plots visualizing PC1,
2116 PC2, and PC3 were made using the *plotly* R package. A two-way admixture model was
2117 considered supported by PCA if:

2118 1. the target group (the center of the cluster of target individuals, to be precise) lay between
2119 the clusters of proxy source individuals on a straight line in the three-dimensional PC
2120 space;

2121 2. or if it was located at a distance of no more than three target cluster diameters from that
2122 straight line connecting the proxy source clusters.

2123 The second pattern was more common among both TP and FP two-way admixture models:
2124 1.5 and 1.3 times, respectively (across all non-subsampled simulated datasets). This situation

2125 is expected since many targets represent three-way and more complex mixtures, and due to
2126 post-admixture genetic drift (Peter 2022).

2127

2128 ***AGS histories: ADMIXTURE analysis***

2129 *ADMIXTURE* analysis was performed for one simulation replicate per simulation setup. Only
2130 high-quality data was used, and all individuals sampled from a simulation were co-analyzed.
2131 Prior to the analysis, linked sites were pruned with *PLINK v.2.00a3LM* (Chang et al. 2015)
2132 using the following settings: window size, 2000 SNPs; window step, 100 SNPs; r^2 threshold =
2133 0.5 (argument "--indep-pairwise 2000 100 0.5"). *ADMIXTURE v.1.3* (Alexander et al. 2009)
2134 was used in the unsupervised mode under the default settings. The algorithm was run on
2135 each SNP dataset only once, with the number of hypothetical ancestral populations (*K*)
2136 ranging from 3 to 10. This range was selected since the total number of populations in each
2137 simulated history was 13. A two-way admixture model was considered supported by
2138 *ADMIXTURE* analysis if:

2139 1. for at least one *K*, at least 5 of 10 target individuals were modelled as a mixture of at least
2140 two ancestry components, with a minor ancestry component exceeding 2%;

2141 2. typically, ancestry component *A* in the target group was shared with at least 5 individuals
2142 in proxy source 1, but not in proxy source 2, and ancestry component *B* was shared with
2143 at least 5 individuals in proxy source 2, but not in proxy source 1 (see examples in **Fig. 4**
2144 and **Fig. S4**); in some cases, both components *A* and *B* were found in the proxy sources,
2145 but in different proportions;

2146 3. if only one ancestry component in the target was shared with the two proxy sources, the
2147 model was considered unsupported;

2148 4. ancestry components in the target that are absent in any of the sources were ignored
2149 since three-way and more complex admixture histories are common in the set of random
2150 admixture graphs explored here (see examples of such topologies in **Fig. S4k-m**);

2151 5. ancestry components in a proxy source that are absent in the target were also ignored
2152 since a proxy source may not be fully cladal with the real source (see examples of such
2153 topologies in **Fig. S4i, j, m**).

2154 These rules were designed to reproduce typical reasoning of an archaeogeneticist
2155 interpreting *ADMIXTURE* results. Observing a pattern of ancestry components in the target
2156 group and proxy sources compatible with the admixture model "target = proxy source$_1$ +
2157 proxy source$_2$" for one *K* value was enough for declaring that the model is supported by the
2158 *ADMIXTURE* analysis. This condition was motivated by an observation that models supported
2159 at one *K* value only were equally common among FP and TP *qpAdm* models (10% and 13%,
2160 respectively, across simulation setups no. 1, 2, 4, and 5). Models supported at four or more *K*
2161 values were more common among TP *qpAdm* models (3.3% of FP and 12.6% of TP models
2162 across the four simulation setups).

2163

2164 ***SSL histories: simulations***

2165 The *slendr* v. 0.7.2.9000 R package (Petr et al. 2023) with the *msprime* v. 1.2.0 backend was
2166 used for the stepping-stone simulations. Sixty-four panmictic demes (with stable effective
2167 sizes of 1,000 diploid individuals) arose via multifurcation at generation 470 (counted from
2168 the past), and, starting at generation 471, three gene-flow eras followed each other: "pre-
2169 LGM" lasting 1,400 generations, "LGM" lasting 350 generations, and "post-LGM" lasting 770
2170 generations (LGM stands for the last glacial maximum). Three diploid individuals were
2171 sampled from each deme at the end of the simulation ("present"), at 100, and 300
2172 generations before "present". The demes were arranged on a triangular lattice, forming an
2173 approximately circular landscape, where each deme was connected by bidirectional gene
2174 flows to 3 to 6 neighboring demes (**Fig. 7**). Gene-flow intensities were sampled randomly (in
2175 each direction independently) from either uniform or normal distributions, generating four
2176 types of landscapes with the following approximate per-generation gene-flow intensity
2177 ranges: 10^{-5} to 10^{-4}, 10^{-4} to 10^{-3}, 10^{-3} to 10^{-2}, 10^{-5} to 10^{-2} (the values stand for percentage of
2178 individuals in a generation coming from a given neighboring deme; see **Table S6** for details).
2179 In the case of the "10^{-5} to 10^{-2}" landscape, normal distributions centered at 0 were used (see
2180 **Table S6** for their parameters). Random sampling of the gene-flow intensity distributions was
2181 done at the beginning of each era, except for the "10^{-3} to 10^{-2}" landscape, where sampling
2182 was repeated five times per era, at equal intervals of time. Ten independent simulation
2183 replicates were generated for each landscape type, with gene-flow intensities re-sampled for

2184 each replicate. Genome size was 500 Mbp, recombination rate was simulated at 1×10^{-8} per
2185 nt per generation, and mutation rate was 1.25×10^{-8} per nt per generation.

2186

2187 ***SSL histories: qpAdm protocols with randomized landscape sampling***

2188 Four *qpAdm* protocols were applied to SSL: distal and proximal non-rotating, distal and
2189 proximal rotating. For constructing these *qpAdm* experiments, demes were sampled
2190 randomly from the landscapes, keeping their original sample size of three diploid individuals.
2191 For each protocol, 36 composite feasibility criteria and two deme sampling densities were
2192 tested: 13 or 18 demes and a target per a rotating *qpAdm* experiment; 10 or 15 proxy sources,
2193 10 or 15 "right" demes, and a target per a non-rotating experiment. For brevity, we refer to
2194 these sampling densities as "13 demes" or "18 demes" for all types of protocols. Here we
2195 describe random deme sampling rules for the four protocols, and they are also illustrated in
2196 **Fig. S12**:

2197 1. Distal non-rotating. Target demes were sampled at "present", and for each target two
2198 alternative sets of "right" groups and proxy sources were generated: 10 (or 15, depending
2199 on the sampling density) "right" demes were from 300 generations before present (ca.
2200 8700 years in the case of humans), and 10 or 15 proxy source demes were from 100
2201 generations before present (ca. 2900 years in the case of humans).

2202 2. Proximal non-rotating. Ten or fifteen "right" demes were sampled at 300 generations
2203 before present, 10 or 15 proxy source demes were sampled at 100 generations and at
2204 "present" (demes from the two dates were combined before sampling), and a target
2205 deme was sampled from either the 100-generations or 300-generations sampling point.

2206 3. Distal rotating. Target demes were sampled at "present", and for each target two
2207 alternative rotating sets were generated: 13 or 18 "right"/source demes were sampled
2208 from 300 and 100 generations ago (demes from the two dates were combined before
2209 sampling).

2210 4. Proximal rotating. A rotating set of 14 or 19 demes was sampled from all the three time
2211 points combined, and a single target was sampled from that subset, but 300-generations-
2212 old targets were not allowed.

2213 The *qpAdm* experiments described above consisted of all possible one-, two-, three-, and
2214 four-way models for one target (one-, two-, and three-way models for proxy sources were
2215 also counted towards some feasibility criteria as described below). Before each *qpAdm* test,
2216 per-block f_2-statistics were calculated with *ADMIXTOOLS* v.2.0.0 (Maier et al. 2023) for the
2217 selected set of demes under the following settings: "*maxmiss=0"* (no missing data allowed at
2218 the group level), "*blgsize=4000000"* (genome block size in nt), "*minac2=FALSE"*,
2219 "*adjust_pseudohaploid=FALSE"* (the latter two settings have their default values intended for
2220 high-quality diploid data). *QpAdm* tests were performed in *ADMIXTOOLS* v.2.0.0 under the
2221 default settings, except for the "*fudge_twice=TRUE*" setting designed to make numerical
2222 results more similar to those of classic *ADMIXTOOLS* (Patterson et al. 2012). Here we describe
2223 36 composite feasibility criteria (arranged from more to less stringent) tested on SSL which
2224 are variations on the criterion used for the work on AGS histories (**Box 1**):

2225 1. Admixture fraction estimates (EAF) $\pm$ 2 SE $\in$ (0, 1); *p*-value threshold for *n*-way models is
2226 either 0.05 or 0.5; all "trailing" simpler models (**Box 1**) are rejected according to the 0.001
2227 *p*-value threshold;

2228 2. EAF $\pm$ 2 SE $\in$ (0, 1); *p*-value threshold for *n*-way models is either 0.001, 0.01 (this version
2229 of the criterion matches the main version applied to AGS histories), 0.05, 0.1, or 0.5; all
2230 "trailing" simpler models are rejected according to the same *p*-value thresholds;

2231 3. EAF $\in$ (0, 1); *p*-value threshold for *n*-way models is either 0.05 or 0.5; all "trailing" simpler
2232 models are rejected according to the 0.001 *p*-value threshold;

2233 4. EAF $\in$ (0, 1); *p*-value threshold for *n*-way models is either 0.001, 0.01, 0.05, 0.1, or 0.5; all
2234 "trailing" simpler models are rejected according to the same *p*-value thresholds;

2235 5. EAF $\in$ (-0.3, 1.3); *p*-value threshold for *n*-way models is either 0.05 or 0.5; all "trailing"
2236 simpler models are rejected according to the 0.001 *p*-value threshold;

2237 6. EAF $\in$ (-0.3, 1.3); *p*-value threshold for *n*-way models is either 0.001, 0.01, 0.05, 0.1, or
2238 0.5; all "trailing" simpler models are rejected according to the same *p*-value thresholds;

2239 7. EAF $\pm$ 2 SE $\in$ (0, 1); *p*-value threshold is either 0.001, 0.01, 0.05, 0.1, or 0.5; "trailing"
2240 simpler models are not checked;

8. EAF ∈ (0, 1); *p*-value threshold is either 0.001, 0.01, 0.05, 0.1, or 0.5; "trailing" simpler models are not checked;
9. EAFs ∈ (-0.3, 1.3); *p*-value threshold is either 0.001, 0.01, 0.05, 0.1, or 0.5; "trailing" simpler models are not checked.

If *qpAdm* results were interpreted at the level of experiments, but not models, for declaring a positive result it was required that all simpler models for the same target (with all possible source combinations) are rejected according to one of the feasibility criteria described above. Models at a complexity level where the first fitting model was found were binned into rejected (non-fitting) and non-rejected (fitting) and ranked by Euclidean distances to ideal symmetric models (of the same complexity) in the space of optimality metrics (**Fig. 7**). We defined "outcomes of experiments that would be misleading in practice" as satisfying the following condition at the model complexity level where the first fitting model is found for the target: min. Euclidean distance$_{non\text{-}rejected}$ – min. Euclidean distance$_{rejected}$ > $\sqrt{8}$, which is a diagonal of a right triangle with legs = 2 (corresponding to spatial distance of 2 or to angular distance of 40°). The reasoning behind this approach is as follows: we hope to accept models that are not much worse in spaces of optimality metrics than the models we reject. The spatial distance of 2 and the angular distance of 40° are substantial in the context of the simulated landscapes with maximal distance = 9 and maximal angle between sources in a two-way model = 180°.

This criterion based on Euclidean distance to the ideal symmetric model was applied to any experiment producing two-way and more complex fitting models, and in the case of experiments yielding one-way fitting models the two-dimensional space of model optimality metrics does not exist, and a simpler condition defined "bad" outcomes of experiments: *min. source-target distance in the non-rejected model group > min. source-target distance in the rejected model group*. For two-way and more complex models with at least one source located at the same position as the target (distance = 0), min. STS angles were undefined, and thus distances in the space of min. STS angles and max. or average ST distance were undefined for them too. Fitting models of this sort were counted towards positive results at a certain complexity level, but not towards experiments with misleading outcomes.

2270 Forty randomized independent experiments progressing to four-way admixture models
2271 (irrespective of model feasibility assessments) were performed per one simulation replicate,
2272 *qpAdm* protocol (distal and proximal, rotating and non-rotating), and landscape sampling
2273 density (13 or 18 demes) (12,800 independent experiments in total). For the most realistic
2274 landscape type, that with per-generation gene-flow intensities ranging from ca. 10^{-5} to 10^{-2},
2275 160 independent experiments progressing to three-way models were performed per one
2276 simulation replicate, *qpAdm* protocol, and landscape sampling density (12,800 experiments
2277 in total). For the three other landscape types, 40 independent experiments progressing to
2278 three-way models were performed per one simulation replicate, *qpAdm* protocol, and
2279 landscape sampling density (9,600 experiments in total). The former set of experiments
2280 progressing to four-way models was used for all the analyses in the study, and the latter two
2281 sets were employed as additional data for comparing estimated and predicted admixture
2282 fractions only (**Figs. 13, S34**).

2283 Spearman's or Pearson's correlation coefficients and statistical significance of their difference
2284 from 0 were calculated in the spaces of EAF or SE vs. model *p*-values (**Tables S1c, S7, S8, S12**),
2285 or in the spaces of optimality metrics (**Table S11**), using the *cor.test* function from the R
2286 package *stats* v. 4.3.0. Two-sided significance tests were used: for Spearman's test, *p*-values
2287 were computed via the asymptotic *t* approximation and subjected to the Bonferroni
2288 correction. The same approaches were used for all applications of Spearman's (**Tables S1c,**
2289 **S7, S8, S11-S13**) or Pearson's (**Figs. 13, S13, S34, Table S14**) correlation coefficients in this
2290 study.

2291

2292 ***SSL histories: qpAdm protocols with systematic landscape sampling***

2293 For *qpAdm* tests based on systematic landscape sampling, we used the landscapes (10
2294 simulation iterations) with per-generation gene-flow intensities ranging between
2295 approximately 10^{-5} and 10^{-2}. On this landscape, we tested two protocols (**Fig. S12**):

2296 1. distal rotating (in each experiment a target was selected from demes sampled at
2297 "present", and 16 demes sampled 100 generations before "present" formed a rotating
2298 set);

2. distal non-rotating (in each experiment a target was selected from demes sampled at "present", 16 demes sampled 100 generations before "present" were used as proxy sources, and 16 demes sampled 300 generations before "present" were used as "right" groups).

Deme no. 32, occupying a central position on the landscape, were selected as a principal target. In the case of the rotating protocol, 16 demes forming a rotating set were selected exclusively from the 1st and 3rd circles of demes surrounding the target (**Fig. S12**). We tested all rotating sets of this class including from 0 to 6 demes from the 1st circle (the nearest neighbors of the target deme). In total, 64 experiments were performed for each simulation replicate. Since there was only one experiment per simulation replicate with 0 or 6 demes from the 1st circle, we repeated experiments of this type (with 0 and 6 demes from the 1st circle) for nine other target demes located in the center of the landscape (if it was possible to find a full 3rd circle for them). The same experiments were reproduced in the non-rotating version: all the demes from the 16-deme rotating sets were used as proxy sources, and the "right" set was identical in all the non-rotating experiments and included 16 demes from the 3rd circle (there were from 17 to 18 demes in that circle in total, depending on the target).

Above we described a "standard systematic setup" (presented or used for generating the results visualized in **Figs. 7, 11-13, S12, S27, S28, S31, S32,** and **Tables S7-S9**), however other configurations of demes were tested too: 1) 16 "right" groups coming from circles at different distances from the target and proxy sources in a distal non-rotating protocol coming alternatively from the 1st, 2nd, 3rd, or 4th circles around the target (**Fig. S25**); 2) 16 "right" groups coming from circles at different distances from the target and all 64 demes sampled at 100 generations ago included as alternative proxy sources in a distal non-rotating protocol (**Figs. S29, S30, S33, Table S12**).

Before each *qpAdm* test, per-block f_2-statistics were calculated with *ADMIXTOOLS* v.2.0.0 (Maier et al. 2023) for the selected set of demes under the following settings: "*maxmiss=0*", "*blgsize=4000000*", "*minac2=FALSE*", and "*adjust_pseudohaploid=FALSE*". *QpAdm* tests were performed under the default settings, except for the "*fudge_twice=TRUE*" setting designed to make numerical results more similar to those of classic *ADMIXTOOLS* (Patterson et al. 2012).

Systematic experiments progressed from simpler to more complex models (from one- to four-way models) until at least one feasible model was encountered. Unlike the randomized experiments which always included tests of all possible models from one- to four-way, systematic experiments were stopped and did not progress all the way to four-way models if a simpler fitting model was encountered. The following composite feasibility criterion was used: EAF ∈ (0, 1); *p*-value threshold is 0.01; *p*-values of "trailing" models were not assessed since all simpler models for the target had to be rejected for an experiment to progress.

Probability density curves for radiocarbon and calendar dates

Probability density curves for published radiocarbon and calendar dates were constructed in *OxCal v.4.4*. For calendar dates, we used the *C-Simulate* function in *OxCal v.4.4* for simulating normally distributed dating methods, taking the average calendar date as a median and the length of the timespan as a 95% confidence interval. For radiocarbon dates, we used calibration based on the IntCal20 calibration curve. Probability densities were summarized using the *Sum* function in *OxCal v.4.4* for each of the three groups of individuals, those included in the "left", "right", and "target" population sets in at least one of the published *qpAdm* models (Narasimhan et al. 2019, Lazaridis et al. 2022), and then plotted together.

Data Availability Statement

The study is based on simulated data exclusively. Principal software packages used in the study are *ADMIXTOOLS 2* (Maier et al. 2023, https://github.com/uqrmaie1/admixtools), *msprime v. 1.1.1* and *v. 1.2.0* (Baumdicker et al. 2022, https://github.com/tskit-dev/msprime), and *slendr v. 0.7.2.9000* (Petr et al. 2023, https://github.com/bodkan/slendr). The supplementary materials for this paper are located at gsajournals.figshare.com.

Acknowledgements

O.F, P.C., and P.F. were supported by the Czech Science Foundation (project no. 21-27624S). O.F., P.F., U.I., J.K., L.A.V., and P.C. were supported by the Czech Ministry of Education, Youth and Sports (program ERC CZ, project no. LL2103). P.F. was supported by the Czech Ministry

of Education, Youth and Sports ERD Funds (project OPVVV 16_013/0001775). P.F. and L.A.V. were also supported by the John Templeton Foundation (grant no. 61220 to David Reich) and by a gift from Jean-Francois Clin. C.D.H. and M.P.W were funded by the National Institutes of Health under award number R35GM146886. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic, and by the ELIXIR-CZ project (ID:90255), part of the international ELIXIR infrastructure.

References

1. Al-Asadi H, Petkova D, Stephens M, Novembre J. Estimating recent migration and population-size surfaces. *PLoS Genet*. **15** (2019) e1007908. doi: 10.1371/journal.pgen.1007908.
2. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals, *Genome Res*. **19** (2009) 1655–1664. doi: 10.1101/gr.094052.109.
3. Allentoft ME, Sikora M, Refoyo-Martínez A, *et al*. Population genomics of post-glacial western Eurasia. *Nature*. **625** (2024) 301–311. doi: 10.1038/s41586-023-06865-0.
4. Antonio ML, Gao Z, Moots HM, Lucci M, Candilio F, Sawyer S, Oberreiter V, Calderon D, Devitofranceschi K, Aikens RC, Aneli S, Bartoli F, Bedini A, Cheronet O, Cotter DJ, Fernandes DM, Gasperetti G, Grifoni R, Guidi A, La Pastina F, Loreti E, Manacorda D, Matullo G, Morretta S, Nava A, Fiocchi Nicolai V, Nomi F, Pavolini C, Pentiricci M, Pergola P, Piranomonte M, Schmidt R, Spinola G, Sperduti A, Rubini M, Bondioli L, Coppa A, Pinhasi R, Pritchard JK. Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science*. **366** (2019) 708–714. doi: 10.1126/science.aay6826.
5. Antonio ML, Weiß CL, Gao Z, Sawyer S, Oberreiter V, Moots HM, Spence JP, Cheronet O, Zagorc B, Praxmarer E, Özdoğan KT, Demetz L, Gelabert P, Fernandes D, Lucci M, Alihodžić T, Amrani S, Avetisyan P, Baillif-Ducros C, Bedić Ž, Bertrand A, Bilić M, Bondioli L, Borówka P, Botte E, Burmaz J, Bužanić D, Candilio F, Cvetko M, De Angelis D, Drnić I, Elschek K, Fantar M, Gaspari A, Gasperetti G, Genchi F, Golubović S, Hukeľová Z, Jankauskas R, Vučković KJ, Jeremić G, Kaić I, Kazek K, Khachatryan H, Khudaverdyan A, Kirchengast S, Korać M, Kozlowski V, Krošláková M, Kušan Špalj D, La Pastina F, Laguardia M, Legrand S, Leleković T, Leskovar T, Lorkiewicz W, Los D, Silva AM, Masaryk R, Matijević V, Cherifi YMS, Meyer N, Mikić I, Miladinović-Radmilović N, Milošević Zakić B, Nacouzi L, Natuniewicz-Sekuła M, Nava A, Neugebauer-Maresch C, Nováček J, Osterholtz A, Paige J, Paraman L, Pieri D, Pieta K, Pop-Lazić S, Ruttkay M, Sanader M, Sołtysiak A, Sperduti A, Stankovic Pesterac T, Teschler-Nicola M, Teul I, Tončinić D, Trapp J, Vulović D, Waliszewski T, Walter D, Živanović M, Filah MEM, Čaušević-Bully M, Šlaus M, Borić D, Novak M, Coppa A, Pinhasi R, Pritchard JK. Stable population structure in Europe since the Iron Age, despite high mobility. *eLife*. **13** (2024) e79714. doi: 10.7554/eLife.79714.
6. Armstrong G, Rahman G, Martino C, McDonald D, Gonzalez A, Mishne G, Knight R. Applications and comparison of dimensionality reduction methods for microbiome data. *Front Bioinform*. **2** (2022) 821861. doi: 10.3389/fbinf.2022.821861.
7. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, Gladstein AL, Gorjanc G, Guo B, Jeffery B, Kretzschumar WW, Lohse K, Matschiner M, Nelson D, Pope NS, Quinto-Cortés CD, Rodrigues MF, Saunack K, Sellinger T, Thornton K, van Kemenade H, Wohns AW, Wong Y, Gravel S, Kern AD, Koskela J, Ralph PL, Kelleher J. Efficient ancestry and mutation simulation with msprime 1.0, *Genetics*. **220** (2022) iyab229. doi: 10.1093/genetics/iyab229.

8. Battey CJ, Coffing GC, Kern AD. Visualizing population structure with variational autoencoders. *G3*. **11** (2021) jkaa036. doi: 10.1093/g3journal/jkaa036.
9. Baxter K. Extrinsic factors that effect the preservation of bone. *The Nebraska Anthropologist*. 19 (2004) 38–45.
10. Bergström A, Stanton DWG, Taron UH, Frantz L, Sinding MS, Ersmark E, Pfrengle S, Cassatt-Johnstone M, Lebrasseur O, Girdland-Flink L, Fernandes DM, Ollivier M, Speidel L, Gopalakrishnan S, Westbury MV, Ramos-Madrigal J, Feuerborn TR, Reiter E, Gretzinger J, Münzel SC, Swali P, Conard NJ, Carøe C, Haile J, Linderholm A, Androsov S, Barnes I, Baumann C, Benecke N, Bocherens H, Brace S, Carden RF, Drucker DG, Fedorov S, Gasparik M, Germonpré M, Grigoriev S, Groves P, Hertwig ST, Ivanova VV, Janssens L, Jennings RP, Kasparov AK, Kirillova IV, Kurmaniyazov I, Kuzmin YV, Kosintsev PA, Lázničková-Galetová M, Leduc C, Nikolskiy P, Nussbaumer M, O'Drisceoil C, Orlando L, Outram A, Pavlova EY, Perri AR, Pilot M, Pitulko VV, Plotnikov VV, Protopopov AV, Rehazek A, Sablin M, Seguin-Orlando A, Storå J, Verjux C, Zaibert VF, Zazula G, Crombé P, Hansen AJ, Willerslev E, Leonard JA, Götherström A, Pinhasi R, Schuenemann VJ, Hofreiter M, Gilbert MTP, Shapiro B, Larson G, Krause J, Dalén L, Skoglund P. Grey wolf genomic history reveals a dual ancestry of dogs, *Nature*. **607** (2022) 313–320. doi: 10.1038/s41586-022-04824-9.
11. Brielle ES, Fleisher J, Wynne-Jones S, Sirak K, Broomandkhoshbacht N, Callan K, Curtis E, Iliev L, Lawson AM, Oppenheimer J, Qiu L, Stewardson K, Workman JN, Zalzala F, Ayodo G, Gidna AO, Kabiru A, Kwekason A, Mabulla AZP, Manthi FK, Ndiema E, Ogola C, Sawchuk E, Al-Gazali L, Ali BR, Ben-Salem S, Letellier T, Pierron D, Radimilahy C, Rakotoarisoa JA, Raaum RL, Culleton BJ, Mallick S, Rohland N, Patterson N, Mwenje MA, Ahmed KB, Mohamed MM, Williams SR, Monge J, Kusimba S, Prendergast ME, Reich D, Kusimba CM. Entwined African and Asian genetic roots of medieval peoples of the Swahili coast. *Nature*. **615** (2023) 866–873. doi: 10.1038/s41586-023-05754-w.
12. Carlhoff S, Duli A, Nägele K, Nur M, Skov L, Sumantri I, Oktaviana AA, Hakim B, Burhan B, Syahdar FA, McGahan DP, Bulbeck D, Perston YL, Newman K, Saiful AM, Ririmasse M, Chia S, Hasanuddin, Pulubuhu DAT, Suryatman, Supriadi, Jeong C, Peter BM, Prüfer K, Powell A, Krause J, Posth C, Brumm A. Genome of a middle Holocene hunter-gatherer from Wallacea. *Nature*. **596** (2021) 543–547. doi: 10.1038/s41586-021-03823-6.
13. Carlhoff S, Kutanan W, Rohrlach AB, Posth C, Stoneking M, Nägele K, Shoocongdej R, Krause J. Genomic portrait and relatedness patterns of the Iron Age Log Coffin culture in northwestern Thailand. *Nat Commun*. **14** (2023) 8527. doi: 10.1038/s41467-023-44328-2.
14. Cerezo-Román JI, Wessman A, and Williams H. (Eds.), 2017 *Cremation And The Archaeology Of Death*. Oxford University Press (pages XVIII and 364, fig. 41).
15. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets, *Gigascience*. **4** (2015) 7. doi: 10.1186/s13742-015-0047-8.
16. Changmai P, Jaisamut K, Kampuansai J, Kutanan W, Altınışık NE, Flegontova O, Inta A, Yüncü E, Boonthai W, Pamjav H, Reich D, Flegontov P. Indian genetic heritage in Southeast Asian populations, *PLoS Genet*. **18** (2022a) e1010036. doi: 10.1371/journal.pgen.1010036.
17. Changmai P, Pinhasi R, Pietrusewsky M, Stark MT, Ikehara-Quebral RM, Reich D, Flegontov P. Ancient DNA from Protohistoric Period Cambodia indicates that South Asians admixed with local populations as early as 1st-3rd centuries CE, *Sci. Rep*. **12** (2022b) 22507. doi: 10.1038/s41598-022-26799-3.
18. Diaconis P, Goel S, Holmes S. Horseshoes in multidimensional scaling and local kernel methods. *Ann. Appl. Stat*. **2** (2008) 777–807. doi: 10.1214/08-AOAS165.
19. Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet*. **15** (2019) e1008432. doi: 10.1371/journal.pgen.1008432.
20. Duforet-Frebourg N, Slatkin M. Isolation-by-distance-and-time in a stepping-stone model. *Theor.*

Popul. Biol. **108** (2016) 24–35. doi: 10.1016/j.tpb.2015.11.003.
21. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations, *Mol. Biol. Evol*. **28** (2011) 2239–2252. doi: 10.1093/molbev/msr048.
22. Estavoyer M, François O. Theoretical analysis of principal components in an umbrella model of intraspecific evolution. *Theor. Popul. Biol*. **148** (2022) 11–21. doi: 10.1016/j.tpb.2022.08.002.
23. Fernandes DM, Mittnik A, Olalde I, Lazaridis I, Cheronet O, Rohland N, Mallick S, Bernardos R, Broomandkhoshbacht N, Carlsson J, Culleton BJ, Ferry M, Gamarra B, Lari M, Mah M, Michel M, Modi A, Novak M, Oppenheimer J, Sirak KA, Stewardson K, Mandl K, Schattke C, Özdoğan KT, Lucci M, Gasperetti G, Candilio F, Salis G, Vai S, Camarós E, Calò C, Catalano G, Cueto M, Forgia V, Lozano M, Marini E, Micheletti M, Miccichè RM, Palombo MR, Ramis D, Schimmenti V, Sureda P, Teira L, Teschler-Nicola M, Kennett DJ, Lalueza-Fox C, Patterson N, Sineo L, Coppa A, Caramelli D, Pinhasi R, Reich D. The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean, *Nat. Ecol. Evol*. **4** (2020) 334–345. doi: 10.1038/s41559-020-1102-0.
24. Fernandes DM, Sirak KA, Ringbauer H, Sedig J, Rohland N, Cheronet O, Mah M, Mallick S, Olalde I, Culleton BJ, Adamski N, Bernardos R, Bravo G, Broomandkhoshbacht N, Callan K, Candilio F, Demetz L, Carlson KSD, Eccles L, Freilich S, George RJ, Lawson AM, Mandl K, Marzaioli F, McCool WC, Oppenheimer J, Özdogan KT, Schattke C, Schmidt R, Stewardson K, Terrasi F, Zalzala F, Antúnez CA, Canosa EV, Colten R, Cucina A, Genchi F, Kraan C, La Pastina F, Lucci M, Maggiolo MV, Marcheco-Teruel B, Maria CT, Martínez C, París I, Pateman M, Simms TM, Sivoli CG, Vilar M, Kennett DJ, Keegan WF, Coppa A, Lipson M, Pinhasi R, Reich D. A genetic history of the pre-contact Caribbean, *Nature*. **590** (2021) 103–110. doi: 10.1038/s41586-020-03053-2.
25. Flegontov P, Işıldak U, Maier R, Yüncü E, Changmai P, Reich D. Modeling of African population history using *f*-statistics is biased when applying all previously proposed SNP ascertainment schemes. *PLoS Genet*. **19** (2023) e1010931. doi: 10.1371/journal.pgen.1010931.
26. Frichot E, Schoville S, Bouchard G, François O. Correcting principal component maps for effects of spatial autocorrelation in population genetic data, *Front. Genet*. **3** (2012) 254. doi: 10.3389/fgene.2012.00254.
27. Geleta M, Mas Montserrat D, Giro-i-Nieto X, Ioannidis AG. Deep variational autoencoders for population genetics. *bioRxiv*. (2023). doi: 10.1101/2023.09.27.558320.
28. Gill H, Lee J, Jeong C. Reconstructing the genetic relationship between ancient and present-day Siberian populations. *Genome Biol Evol*. **16** (2024) evae063. doi: 10.1093/gbe/evae063.
29. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Ž, Gušic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S. A draft sequence of the Neandertal genome, *Science*. **328** (2010) 710–722. doi: 10.1126/science.1188021.
30. Gnecchi-Ruscone GA, Szécsényi-Nagy A, Koncz I, Csiky G, Rácz Z, Rohrlach AB, Brandt G, Rohland N, Csáky V, Cheronet O, Szeifert B, Rácz TÁ, Benedek A, Bernert Z, Berta N, Czifra S, Dani J, Farkas Z, Hága T, Hajdu T, Jászberényi M, Kisjuhász V, Kolozsi B, Major P, Marcsik A, Kovacsóczy BN, Balogh C, Lezsák GM, Ódor JG, Szelekovszky M, Szeniczey T, Tárnoki J, Tóth Z, Tutkovics EK, Mende BG, Geary P, Pohl W, Vida T, Pinhasi R, Reich D, Hofmanová Z, Jeong C, Krause J. Ancient genomes reveal origin and rapid trans-Eurasian migration of 7th century Avar elites, *Cell*. **185** (2022) 1402–1413.e21. doi: 10.1016/j.cell.2022.03.007.
31. Grundler MC, Terhorst J, Bradburd GS. A geographic history of human genetic ancestry. *bioRxiv*. (2024). doi: 10.1101/2024.03.27.586858.
32. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu Q, Mittnik A, Bánffy E, Economou C, Francken M, Friederich S, Pena RG, Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O, Moiseyev

V, Nicklisch N, Pichler SL, Risch R, Rojo Guerra MA, Roth C, Szécsényi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D. Massive migration from the steppe was a source for Indo-European languages in Europe, *Nature*. **522** (2015) 207–211. doi: 10.1038/nature14317.

33. Harney É, May H, Shalem D, Rohland N, Mallick S, Lazaridis I, Sarig R, Stewardson K, Nordenfelt S, Patterson N, Hershkovitz I, Reich D. Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation, *Nat. Commun*. **9** (2018) 3336. doi: 10.1038/s41467-018-05649-9.
34. Harney É, Nayak A, Patterson N, Joglekar P, Mushrif-Tripathy V, Mallick S, Rohland N, Sedig J, Adamski N, Bernardos R, Broomandkhoshbacht N, Culleton BJ, Ferry M, Harper TK, Michel M, Oppenheimer J, Stewardson K, Zhang Z, Harashawaradhana, Bartwal MS, Kumar S, Diyundi SC, Roberts P, Boivin N, Kennett DJ, Thangaraj K, Reich D, Rai N. Ancient DNA from the skeletons of Roopkund Lake reveals Mediterranean migrants in India, *Nat. Commun*. **10** (2019) 3670. doi: 10.1038/s41467-019-11357-9.
35. Harney É, Patterson N, Reich D, Wakeley J. Assessing the performance of qpAdm: a statistical tool for studying population admixture, *Genetics*. **217** (2021) iyaa045. doi: 10.1093/GENETICS/IYAA045.
36. Kerdoncuff E, Skov L, Patterson N, Zhao W, Lueng YY, Schellenberg GD, Smith JA, Dey S, Ganna A, Dey AB, Kardia SLR, Lee J, Moorjani P. 50,000 years of evolutionary history of India: Insights from ~2,700 whole genome sequences. *bioRxiv*. (2024). doi: 10.1101/2024.02.15.580575.
37. Lawson DJ, van Dorp L, Falush D. A tutorial on how not to over-interpret *STRUCTURE* and *ADMIXTURE* bar plots, *Nat Commun*. **9** (2018) 3258. doi: 10.1038/s41467-018-05257-7.
38. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, Connell S, Stewardson K, Harney E, Fu Q, Gonzalez-Fortes G, Jones ER, Roodenberg SA, Lengyel G, Bocquentin F, Gasparian B, Monge JM, Gregg M, Eshed V, Mizrahi AS, Meiklejohn C, Gerritsen F, Bejenaru L, Blüher M, Campbell A, Cavalleri G, Comas D, Froguel P, Gilbert E, Kerr SM, Kovacs P, Krause J, McGettigan D, Merrigan M, Merriwether DA, O'Reilly S, Richards MB, Semino O, Shamoon-Pour M, Stefanescu G, Stumvoll M, Tönjes A, Torroni A, Wilson JF, Yengo L, Hovhannisyan NA, Patterson N, Pinhasi R, Reich D. Genomic insights into the origin of farming in the ancient Near East, *Nature*. **536** (2016) 419–424. doi: 10.1038/nature19310.
39. Lazaridis I, Mittnik A, Patterson N, Mallick S, Rohland N, Pfrengle S, Furtwängler A, Peltzer A, Posth C, Vasilakis A, McGeorge PJP, Konsolaki-Yannopoulou E, Korres G, Martlew H, Michalodimitrakis M, Özsait M, Özsait N, Papathanasiou A, Richards M, Roodenberg SA, Tzedakis Y, Arnott R, Fernandes DM, Hughey JR, Lotakis DM, Navas PA, Maniatis Y, Stamatoyannopoulos JA, Stewardson K, Stockhammer P, Pinhasi R, Reich D, Krause J, Stamatoyannopoulos G. Genetic origins of the Minoans and Mycenaeans, *Nature*. **548** (2017) 214–218. doi: 10.1038/nature23310.
40. Lazaridis I, Alpaslan-Roodenberg S, Acar A, Açıkkol A, Agelarakis A, Aghikyan L, Akyüz U, Andreeva D, Andrijašević G, Antonović D, Armit I, Atmaca A, Avetisyan P, Aytek Aİ, Bacvarov K, Badalyan R, Bakardzhiev S, Balen J, Bejko L, Bernardos R, Bertsatos A, Biber H, Bilir A, Bodružić M, Bonogofsky M, Bonsall C, Borić D, Borovinić N, Bravo Morante G, Buttinger K, Callan K, Candilio F, Carić M, Cheronet O, Chohadzhiev S, Chovalopoulou ME, Chryssoulaki S, Ciobanu I, Čondić N, Constantinescu M, Cristiani E, Culleton BJ, Curtis E, Davis J, Demcenco TI, Dergachev V, Derin Z, Deskaj S, Devejyan S, Djordjević V, Duffett Carlson KS, Eccles LR, Elenski N, Engin A, Erdoğan N, Erir-Pazarcı S, Fernandes DM, Ferry M, Freilich S, Frînculeasa A, Galaty ML, Gamarra B, Gasparyan B, Gaydarska B, Genç E, Gültekin T, Gündüz S, Hajdu T, Heyd V, Hobosyan S, Hovhannisyan N, Iliev I, Iliev L, Iliev S, İvgin İ, Janković I, Jovanova L, Karkanas P, Kavaz-Kındığılı B, Kaya EH, Keating D, Kennett DJ, Deniz Kesici S, Khudaverdyan A, Kiss K, Kılıç S, Klostermann P, Kostak Boca Negra Valdes S, Kovačević S, Krenz-Niedbała M, Krznarić Škrivanko M, Kurti R, Kuzman P, Lawson AM, Lazar C, Leshtakov K, Levy TE, Liritzis I, Lorentz KO, Łukasik S, Mah M,

Mallick S, Mandl K, Martirosyan-Olshansky K, Matthews R, Matthews W, McSweeney K, Melikyan V, Micco A, Michel M, Milašinović L, Mittnik A, Monge JM, Nekhrizov G, Nicholls R, Nikitin AG, Nikolov V, Novak M, Olalde I, Oppenheimer J, Osterholtz A, Özdemir C, Özdoğan KT, Öztürk N, Papadimitriou N, Papakonstantinou N, Papathanasiou A, Paraman L, Paskary EG, Patterson N, Petrakiev I, Petrosyan L, Petrova V, Philippa-Touchais A, Piliposyan A, Pocuca Kuzman N, Potrebica H, Preda-Bălănică B, Premužić Z, Price TD, Qiu L, Radović S, Raeuf Aziz K, Rajić Šikanjić P, Rasheed Raheem K, Razumov S, Richardson A, Roodenberg J, Ruka R, Russeva V, Şahin M, Şarbak A, Savaş E, Schattke C, Schepartz L, Selçuk T, Sevim-Erol A, Shamoon-Pour M, Shephard HM, Sideris A, Simalcsik A, Simonyan H, Sinika V, Sirak K, Sirbu G, Šlaus M, Soficaru A, Söğüt B, Sołtysiak A, Sönmez-Sözer Ç, Stathi M, Steskal M, Stewardson K, Stocker S, Suata-Alpaslan F, Suvorov A, Szécsényi-Nagy A, Szeniczey T, Telnov N, Temov S, Todorova N, Tota U, Touchais G, Triantaphyllou S, Türker A, Ugarković M, Valchev T, Veljanovska F, Videvski Z, Virag C, Wagner A, Walsh S, Włodarczak P, Workman JN, Yardumian A, Yarovoy E, Yavuz AY, Yılmaz H, Zalzala F, Zettl A, Zhang Z, Çavuşoğlu R, Rohland N, Pinhasi R, Reich D, Davtyan R. The genetic history of the Southern Arc: A bridge between West Asia and Europe, *Science*. **377** (2022) eabm4247. doi: 10.1126/science.abm4247.

41. Lee J, Miller BK, Bayarsaikhan J, Johannesson E, Ventresca Miller A, Warinner C, Jeong C. Genetic population structure of the Xiongnu Empire at imperial and local scales. *Sci. Adv*. **9** (2023) eadf3904. doi: 10.1126/sciadv.adf3904.

42. Librado P, Khan N, Fages A, Kusliy MA, Suchan T, Tonasso-Calvière L, Schiavinato S, Alioglu D, Fromentier A, Perdereau A, Aury JM, Gaunitz C, Chauvey L, Seguin-Orlando A, Der Sarkissian C, Southon J, Shapiro B, Tishkin AA, Kovalev AA, Alquraishi S, Alfarhan AH, Al-Rasheid KAS, Seregély T, Klassen L, Iversen R, Bignon-Lau O, Bodu P, Olive M, Castel JC, Boudadi-Maligne M, Alvarez N, Germonpré M, Moskal-Del Hoyo M, Wilczyński J, Pospuła S, Lasota-Kuś A, Tunia K, Nowak M, Rannamäe E, Saarma U, Boeskorov G, Lõugas L, Kyselý R, Peške L, Bălășescu A, Dumitrașcu V, Dobrescu R, Gerber D, Kiss V, Szécsényi-Nagy A, Mende BG, Gallina Z, Somogyi K, Kulcsár G, Gál E, Bendrey R, Allentoft ME, Sirbu G, Dergachev V, Shephard H, Tomadini N, Grouard S, Kasparov A, Basilyan AE, Anisimov MA, Nikolskiy PA, Pavlova EY, Pitulko V, Brem G, Wallner B, Schwall C, Keller M, Kitagawa K, Bessudnov AN, Bessudnov A, Taylor W, Magail J, Gantulga JO, Bayarsaikhan J, Erdenebaatar D, Tabaldiev K, Mijiddorj E, Boldgiv B, Tsagaan T, Pruvost M, Olsen S, Makarewicz CA, Valenzuela Lamas S, Albizuri Canadell S, Nieto Espinet A, Iborra MP, Lira Garrido J, Rodríguez González E, Celestino S, Olària C, Arsuaga JL, Kotova N, Pryor A, Crabtree P, Zhumatayev R, Toleubaev A, Morgunova NL, Kuznetsova T, Lordkipanize D, Marzullo M, Prato O, Bagnasco Gianni G, Tecchiati U, Clavel B, Lepetz S, Davoudi H, Mashkour M, Berezina NY, Stockhammer PW, Krause J, Haak W, Morales-Muñiz A, Benecke N, Hofreiter M, Ludwig A, Graphodatsky AS, Peters J, Kiryushin KY, Iderkhangai TO, Bokovenko NA, Vasiliev SK, Seregin NN, Chugunov KV, Plasteeva NA, Baryshnikov GF, Petrova E, Sablin M, Ananyevskaya E, Logvin A, Shevnina I, Logvin V, Kalieva S, Loman V, Kukushkin I, Merz I, Merz V, Sakenov S, Varfolomeyev V, Usmanova E, Zaibert V, Arbuckle B, Belinskiy AB, Kalmykov A, Reinhold S, Hansen S, Yudin AI, Vybornov AA, Epimakhov A, Berezina NS, Roslyakova N, Kosintsev PA, Kuznetsov PF, Anthony D, Kroonen GJ, Kristiansen K, Wincker P, Outram A, Orlando L. The origins and spread of domestic horses from the Western Eurasian steppes, *Nature*. **598** (2021) 634–640. doi: 10.1038/s41586-021-04018-9.

43. Librado P, Tressières G, Chauvey L, Fages A, Khan N, Schiavinato S, Calvière-Tonasso L, Kusliy MA, Gaunitz C, Liu X, Wagner S, Der Sarkissian C, Seguin-Orlando A, Perdereau A, Aury JM, Southon J, Shapiro B, Bouchez O, Donnadieu C, Collin YRH, Gregersen KM, Jessen MD, Christensen K, Claudi-Hansen L, Pruvost M, Pucher E, Vulic H, Novak M, Rimpf A, Turk P, Reiter S, Brem G, Schwall C, Barrey É, Robert C, Degueurce C, Horwitz LK, Klassen L, Rasmussen U, Kveiborg J, Johannsen NN, Makowiecki D, Makarowicz P, Szeliga M, Ilchyshyn V, Rud V, Romaniszyn J, Mullin VE, Verdugo M, Bradley DG, Cardoso JL, Valente MJ, Telles Antunes M, Ameen C, Thomas R, Ludwig A, Marzullo M, Prato O, Bagnasco Gianni G, Tecchiati U, Granado J,

Schlumbaum A, Deschler-Erb S, Mráz MS, Boulbes N, Gardeisen A, Mayer C, Döhle HJ, Vicze M, Kosintsev PA, Kyselý R, Peške L, O'Connor T, Ananyevskaya E, Shevnina I, Logvin A, Kovalev AA, Iderkhangai TO, Sablin MV, Dashkovskiy PK, Graphodatsky AS, Merts I, Merts V, Kasparov AK, Pitulko VV, Onar V, Öztan A, Arbuckle BS, McColl H, Renaud G, Khaskhanov R, Demidenko S, Kadieva A, Atabiev B, Sundqvist M, Lindgren G, López-Cachero FJ, Albizuri S, Trbojević Vukičević T, Rapan Papeša A, Burić M, Rajić Šikanjić P, Weinstock J, Asensio Vilaró D, Codina F, García Dalmau C, Morer de Llorens J, Pou J, de Prado G, Sanmartí J, Kallala N, Torres JR, Maraoui-Telmini B, Belarte Franco MC, Valenzuela-Lamas S, Zazzo A, Lepetz S, Duchesne S, Alexeev A, Bayarsaikhan J, Houle JL, Bayarkhuu N, Turbat T, Crubézy É, Shingiray I, Mashkour M, Berezina NY, Korobov DS, Belinskiy A, Kalmykov A, Demoule JP, Reinhold S, Hansen S, Wallner B, Roslyakova N, Kuznetsov PF, Tishkin AA, Wincker P, Kanne K, Outram A, Orlando L. Widespread horse-based mobility arose around 2200 BCE in Eurasia. *Nature*. **631** (2024) 819–825. doi: 10.1038/s41586-024-07597-5.

44. Lindig WH. Tree burial among the Seri Indians. *Ethnology*. **3** (1964) 284–286. doi: 10.2307/3772884.
45. Lipson M. Applying f_4-statistics and admixture graphs: Theory and examples, *Mol. Ecol. Resour*. **20** (2020) 1658–1667. doi: 10.1111/1755-0998.13230.
46. Maier R, Flegontov P, Flegontova O, Işıldak U, Changmai P, Reich D. On the limits of fitting complex models of population history to *f*-statistics. *eLife*. **12** (2023) e85492. doi: 10.7554/eLife.85492.
47. Marcus JH, Posth C, Ringbauer H, Lai L, Skeates R, Sidore C, Beckett J, Furtwängler A, Olivieri A, Chiang CWK, Al-Asadi H, Dey K, Joseph TA, Liu CC, Der Sarkissian C, Radzevičiūtė R, Michel M, Gradoli MG, Marongiu P, Rubino S, Mazzarello V, Rovina D, La Fragola A, Serra RM, Bandiera P, Bianucci R, Pompianu E, Murgia C, Guirguis M, Orquin RP, Tuross N, van Dommelen P, Haak W, Reich D, Schlessinger D, Cucca F, Krause J, Novembre J. Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia, *Nat. Commun*. **11** (2020) 939. doi: 10.1038/s41467-020-14523-6.
48. Maróti Z, Neparáczki E, Schütz O, Maár K, Varga GIB, Kovács B, Kalmár T, Nyerki E, Nagy I, Latinovics D, Tihanyi B, Marcsik A, Pálfi G, Bernert Z, Gallina Z, Horváth C, Varga S, Költő L, Raskó I, Nagy PL, Balogh C, Zink A, Maixner F, Götherström A, George R, Szalontai C, Szenthe G, Gáll E, Kiss AP, Gulyás B, Kovacsóczy BN, Gál SS, Tomka P, Török T. The genetic origin of Huns, Avars, and conquering Hungarians, *Curr. Biol*. **32** (2022) 2858–2870.e7. doi: 10.1016/j.cub.2022.04.093.
49. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA-BABA statistics to locate introgressed loci, *Mol. Biol. Evol*. **32** (2014) 244–257. doi: 10.1093/molbev/msu269.
50. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, Sirak K, Gamba C, Jones ER, Llamas B, Dryomov S, Pickrell J, Arsuaga JL, de Castro JM, Carbonell E, Gerritsen F, Khokhlov A, Kuznetsov P, Lozano M, Meller H, Mochalov O, Moiseyev V, Guerra MA, Roodenberg J, Vergès JM, Krause J, Cooper A, Alt KW, Brown D, Anthony D, Lalueza-Fox C, Haak W, Pinhasi R, Reich D. Genome-wide patterns of selection in 230 ancient Eurasians, *Nature*. **528** (2015) 499–503. doi: 10.1038/nature16152. Epub 2015 Nov 23. PMID: 26595274; PMCID: PMC4918750.
51. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, Olalde I, Broomandkhoshbacht N, Candilio F, Cheronet O, Fernandes D, Ferry M, Gamarra B, Fortes GG, Haak W, Harney E, Jones E, Keating D, Krause-Kyora B, Kucukkalipci I, Michel M, Mittnik A, Nägele K, Novak M, Oppenheimer J, Patterson N, Pfrengle S, Sirak K, Stewardson K, Vai S, Alexandrov S, Alt KW, Andreescu R, Antonović D, Ash A, Atanassova N, Bacvarov K, Gusztáv MB, Bocherens H, Bolus M, Boroneanţ A, Boyadzhiev Y, Budnik A, Burmaz J, Chohadzhiev S, Conard NJ, Cottiaux R, Čuka M, Cupillard C, Drucker DG, Elenski N, Francken M, Galabova B, Ganetsovski G, Gély B, Hajdu T, Handzhyiska V, Harvati K, Higham T, Iliev S, Janković I, Karavanić I, Kennett DJ, Komšo D, Kozak A, Labuda D, Lari M, Lazar C, Leppek M, Leshtakov K, Vetro DL, Los D, Lozanov I, Malina M, Martini F, McSweeney K, Meller H, Menđušić M, Mirea P, Moiseyev V,

Petrova V, Price TD, Simalcsik A, Sineo L, Šlaus M, Slavchev V, Stanev P, Starović A, Szeniczey T, Talamo S, Teschler-Nicola M, Thevenet C, Valchev I, Valentin F, Vasilyev S, Veljanovska F, Venelinova S, Veselovskaya E, Viola B, Virag C, Zaninović J, Zäuner S, Stockhammer PW, Catalano G, Krauß R, Caramelli D, Zariņa G, Gaydarska B, Lillie M, Nikitin AG, Potekhina I, Papathanasiou A, Borić D, Bonsall C, Krause J, Pinhasi R, Reich D. The genomic history of southeastern Europe, *Nature*. **555** (2018) 197–203. doi: 10.1038/nature25778.

52. McInnes L, Healy J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*. (2018). doi: 10.48550/arXiv.1802.03426.
53. McVean G. A genealogical interpretation of principal components analysis, *PLoS Genet*. **5** (2009) e1000686. doi: 10.1371/journal.pgen.1000686.
54. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T, Rasmussen S, Margaryan A, Iraeta Orbegozo M, Mylopotamitaki D, Wooller M, Bataille C, Becerra-Valdivia L, Chivall D, Comeskey D, Devièse T, Grayson DK, George L, Harry H, Alexandersen V, Primeau C, Erlandson J, Rodrigues-Carvalho C, Reis S, Bastos MQR, Cybulski J, Vullo C, Morello F, Vilar M, Wells S, Gregersen K, Hansen KL, Lynnerup N, Mirazón Lahr M, Kjær K, Strauss A, Alfonso-Durruty M, Salas A, Schroeder H, Higham T, Malhi RS, Rasic JT, Souza L, Santos FR, Malaspinas AS, Sikora M, Nielsen R, Song YS, Meltzer DJ, Willerslev E. Early human dispersals within the Americas, *Science*. **362** (2018) eaav2621. doi: 10.1126/science.aav2621.
55. Nakatsuka N, Holguin B, Sedig J, Langenwalter PE 2nd, Carpenter J, Culleton BJ, García-Moreno C, Harper TK, Martin D, Martínez-Ramírez J, Porcayo-Michelini A, Tiesler V, Villapando-Canchola ME, Valdes Herrera A, Callan K, Curtis E, Kearns A, Iliev L, Lawson AM, Mah M, Mallick S, Micco A, Michel M, Workman JN, Oppenheimer J, Qiu L, Zalzala F, Rohland N, Punzo Diaz JL, Johnson JR, Reich D. Genetic continuity and change among the Indigenous peoples of California. *Nature*. **624** (2023) 122–129. doi: 10.1038/s41586-023-06771-5.
56. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M, Kim AM, Olivieri LM, Coppa A, Vidale M, Mallory J, Moiseyev V, Kitov E, Monge J, Adamski N, Alex N, Broomandkhoshbacht N, Candilio F, Callan K, Cheronet O, Culleton BJ, Ferry M, Fernandes D, Freilich S, Gamarra B, Gaudio D, Hajdinjak M, Harney É, Harper TK, Keating D, Lawson AM, Mah M, Mandl K, Michel M, Novak M, Oppenheimer J, Rai N, Sirak K, Slon V, Stewardson K, Zalzala F, Zhang Z, Akhatov G, Bagashev AN, Bagnera A, Baitanayev B, Bendezu-Sarmiento J, Bissembaev AA, Bonora GL, Chargynov TT, Chikisheva T, Dashkovskiy PK, Derevianko A, Dobeš M, Douka K, Dubova N, Duisengali MN, Enshin D, Epimakhov A, Fribus AV, Fuller D, Goryachev A, Gromov A, Grushin SP, Hanks B, Judd M, Kazizov E, Khokhlov A, Krygin AP, Kupriyanova E, Kuznetsov P, Luiselli D, Maksudov F, Mamedov AM, Mamirov TB, Meiklejohn C, Merrett DC, Micheli R, Mochalov O, Mustafokulov S, Nayak A, Pettener D, Potts R, Razhev D, Rykun M, Sarno S, Savenkova TM, Sikhymbaeva K, Slepchenko SM, Soltobaev OA, Stepanova N, Svyatko S, Tabaldiev K, Teschler-Nicola M, Tishkin AA, Tkachev VV, Vasilyev S, Velemínský P, Voyakin D, Yermolayeva A, Zahir M, Zubkov VS, Zubova A, Shinde VS, Lalueza-Fox C, Meyer M, Anthony D, Boivin N, Thangaraj K, Kennett DJ, Frachetti M, Pinhasi R, Reich D. The formation of human populations in South and Central Asia, *Science*. **365** (2019) eaat7487. doi: 10.1126/science.aat7487.
57. Nelson D, Kelleher J, Ragsdale AP, Moreau C, McVean G, Gravel S. Accounting for long-range correlations in genome-wide simulations of large cohorts, *PLoS Genet*. **16** (2020) e1008619. doi: 10.1371/journal.pgen.1008619.
58. Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. *PLoS Comput Biol*. **15** (2019) e1006907. doi: 10.1371/journal.pcbi.1006907.
59. Ning C, Li T, Wang K, Zhang F, Li T, Wu X, Gao S, Zhang Q, Zhang H, Hudson MJ, Dong G, Wu S, Fang Y, Liu C, Feng C, Li W, Han T, Li R, Wei J, Zhu Y, Zhou Y, Wang CC, Fan S, Xiong Z, Sun Z, Ye M, Sun L, Wu X, Liang F, Cao Y, Wei X, Zhu H, Zhou H, Krause J, Robbeets M, Jeong C, Cui Y. Ancient genomes from northern China suggest links between subsistence changes and human

migration, *Nat. Commun*. **11** (2020a) 2700. doi: 10.1038/s41467-020-16557-2.
60. Ning C, Fernandes D, Changmai P, Flegontova O, Yüncü E, Maier R, Altınışık NE, Kassian AS, Krause J., Lalueza-Fox C, Manica A, Potter BA, Robbeets M, Sirak K, Siska V, Vajda EJ, Vyazov LA, Wang K, Wang L, Wu X, Xiao X, Zhang F, Reich D, Schiffels S, Pinhasi R, Cui Y, Flegontov P. The genomic formation of First American ancestors in East and Northeast Asia, *bioRxiv*. (2020b). doi: 10.1101/2020.10.12.336628.
61. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation, *Nat. Genet*. **40** (2008) 646–649. doi: 10.1038/ng.139.
62. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, Dulias K, Edwards CJ, Gandini F, Pala M, Soares P, Ferrando-Bernal M, Adamski N, Broomandkhoshbacht N, Cheronet O, Culleton BJ, Fernandes D, Lawson AM, Mah M, Oppenheimer J, Stewardson K, Zhang Z, Jiménez Arenas JM, Toro Moyano IJ, Salazar-García DC, Castanyer P, Santos M, Tremoleda J, Lozano M, García Borja P, Fernández-Eraso J, Mujika-Alustiza JA, Barroso C, Bermúdez FJ, Viguera Mínguez E, Burch J, Coromina N, Vivó D, Cebrià A, Fullola JM, García-Puchol O, Morales JI, Oms FX, Majó T, Vergès JM, Díaz-Carvajal A, Ollich-Castanyer I, López-Cachero FJ, Silva AM, Alonso-Fernández C, Delibes de Castro G, Jiménez Echevarría J, Moreno-Márquez A, Pascual Berlanga G, Ramos-García P, Ramos-Muñoz J, Vijande Vila E, Aguilella Arzo G, Esparza Arroyo Á, Lillios KT, Mack J, Velasco-Vázquez J, Waterman A, Benítez de Lugo Enrich L, Benito Sánchez M, Agustí B, Codina F, de Prado G, Estalrrich A, Fernández Flores Á, Finlayson C, Finlayson G, Finlayson S, Giles-Guzmán F, Rosas A, Barciela González V, García Atiénzar G, Hernández Pérez MS, Llanos A, Carrión Marco Y, Collado Beneyto I, López-Serrano D, Sanz Tormo M, Valera AC, Blasco C, Liesau C, Ríos P, Daura J, de Pedro Michó MJ, Diez-Castillo AA, Flores Fernández R, Francès Farré J, Garrido-Pena R, Gonçalves VS, Guerra-Doce E, Herrero-Corral AM, Juan-Cabanilles J, López-Reyes D, McClure SB, Merino Pérez M, Oliver Foix A, Sanz Borràs M, Sousa AC, Vidal Encinas JM, Kennett DJ, Richards MB, Werner Alt K, Haak W, Pinhasi R, Lalueza-Fox C, Reich D. The genomic history of the Iberian Peninsula over the past 8000 years, *Science*. **363** (2019) 1230–1234. doi: 10.1126/science.aav4040.
63. Oliveira S, Nägele K, Carlhoff S, Pugach I, Koesbardiati T, Hübner A, Meyer M, Oktaviana AA, Takenaka M, Katagiri C, Murti DB, Putri RS, Mahirta, Petchey F, Higham T, Higham CFW, O'Connor S, Hawkins S, Kinaston R, Bellwood P, Ono R, Powell A, Krause J, Posth C, Stoneking M. Ancient genomes from the last three millennia support multiple human dispersals into Wallacea. *Nat. Ecol. Evol*. **6** (2022) 1024–1034. doi: 10.1038/s41559-022-01775-2.
64. Papac L, Ernée M, Dobeš M, Langová M, Rohrlach AB, Aron F, Neumann GU, Spyrou MA, Rohland N, Velemínský P, Kuna M, Brzobohatá H, Culleton B, Daněček D, Danielisová A, Dobisíková M, Hložek J, Kennett DJ, Klementová J, Kostka M, Krištuf P, Kuchařík M, Hlavová JK, Limburský P, Malyková D, Mattiello L, Pecinovská M, Petriščáková K, Průchová E, Stránská P, Smejtek L, Špaček J, Šumberová R, Švejcar O, Trefný M, Vávra M, Kolář J, Heyd V, Krause J, Pinhasi R, Reich D, Schiffels S, Haak W. Dynamic changes in genomic and social structures in third millennium BCE central Europe, *Sci. Adv*. **7** (2021) eabi6941. doi: 10.1126/sciadv.abi6941.
65. Patterson N, Price AL, Reich D. Population structure and eigenanalysis, *PLoS Genet*. **2** (2006) e190. doi: 10.1371/journal.pgen.0020190.
66. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. Ancient admixture in human history, Genetics. **192** (2012) 1065–1093. doi: 10.1534/genetics.112.145037.
67. Patterson N, Isakov M, Booth T, Büster L, Fischer CE, Olalde I, Ringbauer H, Akbari A, Cheronet O, Bleasdale M, Adamski N, Altena E, Bernardos R, Brace S, Broomandkhoshbacht N, Callan K, Candilio F, Culleton B, Curtis E, Demetz L, Carlson KSD, Edwards CJ, Fernandes DM, Foody MGB, Freilich S, Goodchild H, Kearns A, Lawson AM, Lazaridis I, Mah M, Mallick S, Mandl K, Micco A, Michel M, Morante GB, Oppenheimer J, Özdoğan KT, Qiu L, Schattke C, Stewardson K, Workman JN, Zalzala F, Zhang Z, Agustí B, Allen T, Almássy K, Amkreutz L, Ash A, Baillif-Ducros C, Barclay A, Bartosiewicz L, Baxter K, Bernert Z, Blažek J, Bodružić M, Boissinot P, Bonsall C, Bradley P,

Brittain M, Brookes A, Brown F, Brown L, Brunning R, Budd C, Burmaz J, Canet S, Carnicero-Cáceres S, Čaušević-Bully M, Chamberlain A, Chauvin S, Clough S, Čondić N, Coppa A, Craig O, Črešnar M, Cummings V, Czifra S, Danielisová A, Daniels R, Davies A, de Jersey P, Deacon J, Deminger C, Ditchfield PW, Dizdar M, Dobeš M, Dobisíková M, Domboróczki L, Drinkall G, Đukić A, Ernée M, Evans C, Evans J, Fernández-Götz M, Filipović S, Fitzpatrick A, Fokkens H, Fowler C, Fox A, Gallina Z, Gamble M, González Morales MR, González-Rabanal B, Green A, Gyenesei K, Habermehl D, Hajdu T, Hamilton D, Harris J, Hayden C, Hendriks J, Hernu B, Hey G, Horňák M, Ilon G, Istvánovits E, Jones AM, Kavur MB, Kazek K, Kenyon RA, Khreisheh A, Kiss V, Kleijne J, Knight M, Kootker LM, Kovács PF, Kozubová A, Kulcsár G, Kulcsár V, Le Pennec C, Legge M, Leivers M, Loe L, López-Costas O, Lord T, Los D, Lyall J, Marín-Arroyo AB, Mason P, Matošević D, Maxted A, McIntyre L, McKinley J, McSweeney K, Meijlink B, Mende BG, Menđušić M, Metlička M, Meyer S, Mihovilić K, Milasinovic L, Minnitt S, Moore J, Morley G, Mullan G, Musilová M, Neil B, Nicholls R, Novak M, Pala M, Papworth M, Paresys C, Patten R, Perkić D, Pesti K, Petit A, Petriščáková K, Pichon C, Pickard C, Pilling Z, Price TD, Radović S, Redfern R, Resutík B, Rhodes DT, Richards MB, Roberts A, Roefstra J, Sankot P, Šefčáková A, Sheridan A, Skae S, Šmolíková M, Somogyi K, Somogyvári Á, Stephens M, Szabó G, Szécsényi-Nagy A, Szeniczey T, Tabor J, Tankó K, Maria CT, Terry R, Teržan B, Teschler-Nicola M, Torres-Martínez JF, Trapp J, Turle R, Ujvári F, van der Heiden M, Veleminsky P, Veselka B, Vytlačil Z, Waddington C, Ware P, Wilkinson P, Wilson L, Wiseman R, Young E, Zaninović J, Žitňan A, Lalueza-Fox C, de Knijff P, Barnes I, Halkon P, Thomas MG, Kennett DJ, Cunliffe B, Lillie M, Rohland N, Pinhasi R, Armit I, Reich D. Large-scale migration into Britain during the Middle to Late Bronze Age, *Nature*. **601** (2022) 588–594. doi: 10.1038/s41586-021-04287-4.

68. Pavlinskaya LR (Ed.), 2009 *Sibirskij sbornik-1: Pogrebalnyj obryad narodov Sibiri i sopredelnych territorij (book 2)*. Museum of Antrhopology and Ethnography of the Russian Academy of Sciences, St. Petersburg.
69. Peter BM. Admixture, Population Structure, and *F*-Statistics, *Genetics*. **202** (2016) 1485–1501. doi: 10.1534/genetics.115.183913.
70. Peter BM. A geometric relationship of *F2*, *F3* and *F4*-statistics with principal component analysis, *Philos Trans R Soc Lond B Biol Sci*. **377** (2022) 20200413. doi: 10.1098/rstb.2020.0413.
71. Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet*. **48** (2016) 94–100. doi: 10.1038/ng.3464.
72. Podani J, Miklós I. Resemblance coefficients and the horseshoe effect in Principal Coordinates Analysis, *Ecology*. **83** (2002) 3331–3343. doi: 10.1890/0012-9658(2002)083[3331:RCATHE]2.0.CO;2.
73. Posth C, Yu H, Ghalichi A, Rougier H, Crevecoeur I, Huang Y, Ringbauer H, Rohrlach AB, Nägele K, Villalba-Mouco V, Radzeviciute R, Ferraz T, Stoessel A, Tukhbatova R, Drucker DG, Lari M, Modi A, Vai S, Saupe T, Scheib CL, Catalano G, Pagani L, Talamo S, Fewlass H, Klaric L, Morala A, Rué M, Madelaine S, Crépin L, Caverne JB, Bocaege E, Ricci S, Boschin F, Bayle P, Maureille B, Le Brun-Ricalens F, Bordes JG, Oxilia G, Bortolini E, Bignon-Lau O, Debout G, Orliac M, Zazzo A, Sparacello V, Starnini E, Sineo L, van der Plicht J, Pecqueur L, Merceron G, Garcia G, Leuvrey JM, Garcia CB, Gómez-Olivencia A, Połtowicz-Bobak M, Bobak D, Le Luyer M, Storm P, Hoffmann C, Kabaciński J, Filimonova T, Shnaider S, Berezina N, González-Rabanal B, González Morales MR, Marín-Arroyo AB, López B, Alonso-Llamazares C, Ronchitelli A, Polet C, Jadin I, Cauwe N, Soler J, Coromina N, Rufí I, Cottiaux R, Clark G, Straus LG, Julien MA, Renhart S, Talaa D, Benazzi S, Romandini M, Amkreutz L, Bocherens H, Wißing C, Villotte S, de Pablo JF, Gómez-Puche M, Esquembre-Bebia MA, Bodu P, Smits L, Souffi B, Jankauskas R, Kozakaitė J, Cupillard C, Benthien H, Wehrberger K, Schmitz RW, Feine SC, Schüler T, Thevenet C, Grigorescu D, Lüth F, Kotula A, Piezonka H, Schopper F, Svoboda J, Sázelová S, Chizhevsky A, Khokhlov A, Conard NJ, Valentin F, Harvati K, Semal P, Jungklaus B, Suvorov A, Schulting R, Moiseyev V, Mannermaa K, Buzhilova A, Terberger T, Caramelli D, Altena E, Haak W, Krause J. Palaeogenomics of Upper Palaeolithic to Neolithic European hunter-gatherers. *Nature*. 615 (2023) 117–126. doi: 10.1038/s41586-023-

05726-0.
74. Prendergast ME, Lipson M, Sawchuk EA, Olalde I, Ogola CA, Rohland N, Sirak KA, Adamski N, Bernardos R, Broomandkhoshbacht N, Callan K, Culleton BJ, Eccles L, Harper TK, Lawson AM, Mah M, Oppenheimer J, Stewardson K, Zalzala F, Ambrose SH, Ayodo G, Gates HL Jr, Gidna AO, Katongo M, Kwekason A, Mabulla AZP, Mudenda GS, Ndiema EK, Nelson C, Robertshaw P, Kennett DJ, Manthi FK, Reich D. Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science*. **365** (2019) eaaw6275. doi: 10.1126/science.aaw6275.
75. Reich D, Price AL, Patterson N. Principal component analysis of genetic data, *Nat. Genet*. **40** (2008) 491–492.
76. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history, *Nature*. **461** (2009) 489–494. doi: 10.1038/nature08365.
77. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, García LF, Triana O, Blair S, Maestre A, Dib JC, Bravi CM, Bailliet G, Corach D, Hünemeier T, Bortolini MC, Salzano FM, Petzl-Erler ML, Acuña-Alonzo V, Aguilar-Salinas C, Canizales-Quinteros S, Tusié-Luna T, Riba L, Rodríguez-Cruz M, Lopez-Alarcón M, Coral-Vazquez R, Canto-Cetina T, Silva-Zolezzi I, Fernandez-Lopez JC, Contreras AV, Jimenez-Sanchez G, Gómez-Vázquez MJ, Molina J, Carracedo A, Salas A, Gallo C, Poletti G, Witonsky DB, Alkorta-Aranburu G, Sukernik RI, Osipova L, Fedorova SA, Vasquez R, Villena M, Moreau C, Barrantes R, Pauls D, Excoffier L, Bedoya G, Rothhammer F, Dugoujon JM, Larrouy G, Klitz W, Labuda D, Kidd J, Kidd K, Di Rienzo A, Freimer NB, Price AL, Ruiz-Linares A. Reconstructing Native American population history, *Nature*. **488** (2012) 370–374. doi: 10.1038/nature11258.
78. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution, *Nat. Rev. Genet*. **13** (2012) 745–753. doi: 10.1038/nrg3295.
79. Sikora M, Pitulko VV, Sousa VC, Allentoft ME, Vinner L, Rasmussen S, Margaryan A, de Barros Damgaard P, de la Fuente C, Renaud G, Yang MA, Fu Q, Dupanloup I, Giampoudakis K, Nogués-Bravo D, Rahbek C, Kroonen G, Peyrot M, McColl H, Vasilyev SV, Veselovskaya E, Gerasimova M, Pavlova EY, Chasnyk VG, Nikolskiy PA, Gromov AV, Khartanovich VI, Moiseyev V, Grebenyuk PS, Fedorchenko AY, Lebedintsev AI, Slobodin SB, Malyarchuk BA, Martiniano R, Meldgaard M, Arppe L, Palo JU, Sundell T, Mannermaa K, Putkonen M, Alexandersen V, Primeau C, Baimukhanov N, Malhi RS, Sjögren KG, Kristiansen K, Wessman A, Sajantila A, Lahr MM, Durbin R, Nielsen R, Meltzer DJ, Excoffier L, Willerslev E. The population history of northeastern Siberia since the Pleistocene. *Nature*. **570** (2019) 182–188. doi: 10.1038/s41586-019-1279-z.
80. Sirak KA, Fernandes DM, Lipson M, Mallick S, Mah M, Olalde I, Ringbauer H, Rohland N, Hadden CS, Harney É, Adamski N, Bernardos R, Broomandkhoshbacht N, Callan K, Ferry M, Lawson AM, Michel M, Oppenheimer J, Stewardson K, Zalzala F, Patterson N, Pinhasi R, Thompson JC, Van Gerven D, Reich D. Social stratification without genetic differentiation at the site of Kulubnarti in Christian Period Nubia, *Nat. Commun*. **12** (2021) 7283. doi: 10.1038/s41467-021-27356-8.
81. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, Salie T, Rohland N, Mallick S, Peltzer A, Heinze A, Olalde I, Ferry M, Harney E, Michel M, Stewardson K, Cerezo-Román JI, Chiumia C, Crowther A, Gomani-Chindebvu E, Gidna AO, Grillo KM, Helenius IT, Hellenthal G, Helm R, Horton M, López S, Mabulla AZP, Parkington J, Shipton C, Thomas MG, Tibesasa R, Welling M, Hayes VM, Kennett DJ, Ramesar R, Meyer M, Pääbo S, Patterson N, Morris AG, Boivin N, Pinhasi R, Krause J, Reich D. Reconstructing Prehistoric African Population Structure, *Cell*. **171** (2017) 59–71.e21. doi: 10.1016/j.cell.2017.08.049.
82. Skourtanioti E, Ringbauer H, Gnecchi Ruscone GA, Bianco RA, Burri M, Freund C, Furtwängler A, Gomes Martins NF, Knolle F, Neumann GU, Tiliakou A, Agelarakis A, Andreadaki-Vlazaki M, Betancourt P, Hallager BP, Jones OA, Kakavogianni O, Kanta A, Karkanas P, Kataki E, Kissas K, Koehl R, Kvapil L, Maran J, McGeorge PJP, Papadimitriou A, Papathanasiou A, Papazoglou-Manioudaki L, Paschalidis K, Polychronakou-Sgouritsa N, Preve S, Prevedorou EA, Price G, Protopapadaki E, Schmidt-Schultz T, Schultz M, Shelton K, Wiener MH, Krause J, Jeong C, Stockhammer PW. Ancient DNA reveals admixture history and endogamy in the prehistoric

Aegean. *Nat Ecol Evol*. **7** (2023) 290–303. doi: 10.1038/s41559-022-01952-3.
83. Soraggi S, Wiuf C. General theory for stochastic admixture graphs and *F*-statistics, *Theor. Popul. Biol*. **125** (2019) 56–66. doi: 10.1016/j.tpb.2018.12.002.
84. Speidel L, Silva M, Booth T, Raffield B, Anastasiadou K, Barrington C, Götherström A, Heather P, Skoglund P. High-resolution genomic ancestry reveals mobility in early medieval Europe. *bioRxiv*. (2024). doi: https://doi.org/10.1101/2024.03.15.585102.
85. Taylor WTT, Librado P, American Horse CJ, Shield Chief Gover C, Arterberry J, Afraid of Bear-Cook AL, Left Heron H, Yellow Hair RM, Gonzalez M, Means B, High Crane S, Yellow Bull WW, Dull Knife B, Afraid of Bear A, Tecumseh Collin C, Ward C, Pasqual TA, Chauvey L, Tonasso-Calviere L, Schiavinato S, Seguin-Orlando A, Fages A, Khan N, Der Sarkissian C, Liu X, Wagner S, Leonard BG, Manzano BL, O'Malley N, Leonard JA, Bernáldez-Sánchez E, Barrey E, Charliquart L, Robbe E, Denoblet T, Gregersen K, Vershinina AO, Weinstock J, Rajić Šikanjić P, Mashkour M, Shingiray I, Aury JM, Perdereau A, Alquraishi S, Alfarhan AH, Al-Rasheid KAS, Trbojević Vukičević T, Buric M, Sauer E, Lucas M, Brenner-Coltrain J, Bozell JR, Thornhill CA, Monagle V, Perri A, Newton C, Hall WE, Conver JL, Le Roux P, Buckser SG, Gabe C, Belardi JB, Barrón-Ortiz CI, Hart IA, Ryder C, Sponheimer M, Shapiro B, Southon J, Hibbs J, Faulkner C, Outram A, Patterson Rosa L, Palermo K, Solé M, William A, McCrory W, Lindgren G, Brooks S, Eché C, Donnadieu C, Bouchez O, Wincker P, Hodgins G, Trabert S, Bethke B, Roberts P, Jones EL, Running Horse Collin Y, Orlando L. Early dispersal of domestic horses into the Great Plains and northern Rockies. *Science*. **379** (2023) 1316–1323. doi: 10.1126/science.adc9691.
86. Tricou T, Tannier E, de Vienne DM. Ghost Lineages Highly Influence the Interpretation of Introgression Tests, *Syst. Biol*. **71** (2022) 1147–1158. doi: 10.1093/sysbio/syac011.
87. van de Loosdrecht M, Bouzouggar A, Humphrey L, Posth C, Barton N, Aximu-Petri A, Nickel B, Nagel S, Talbi EH, El Hajraoui MA, Amzazi S, Hublin JJ, Pääbo S, Schiffels S, Meyer M, Haak W, Jeong C, Krause J. Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. *Science*. **360** (2018) 548–552. doi: 10.1126/science.aar8380.
88. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J. Machine Learning Res*. **9** (2008) 2579–2605.
89. Villalba-Mouco V, van de Loosdrecht MS, Rohrlach AB, Fewlass H, Talamo S, Yu H, Aron F, Lalueza-Fox C, Cabello L, Cantalejo Duarte P, Ramos-Muñoz J, Posth C, Krause J, Weniger GC, Haak W. A 23,000-year-old southern Iberian individual links human groups that lived in Western Europe before and after the Last Glacial Maximum. *Nat Ecol Evol*. **7** (2023) 597–609. doi: 10.1038/s41559-023-01987-0.
90. Wang K, Goldstein S, Bleasdale M, Clist B, Bostoen K, Bakwa-Lufu P, Buck LT, Crowther A, Dème A, McIntosh RJ, Mercader J, Ogola C, Power RC, Sawchuk E, Robertshaw P, Wilmsen EN, Petraglia M, Ndiema E, Manthi FK, Krause J, Roberts P, Boivin N, Schiffels S. Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa, *Sci. Adv*. **6** (2020) eaaz0183. doi: 10.1126/sciadv.aaz0183.
91. Wang CC, Yeh HY, Popov AN, Zhang HQ, Matsumura H, Sirak K, Cheronet O, Kovalev A, Rohland N, Kim AM, Mallick S, Bernardos R, Tumen D, Zhao J, Liu YC, Liu JY, Mah M, Wang K, Zhang Z, Adamski N, Broomandkhoshbacht N, Callan K, Candilio F, Carlson KSD, Culleton BJ, Eccles L, Freilich S, Keating D, Lawson AM, Mandl K, Michel M, Oppenheimer J, Özdoğan KT, Stewardson K, Wen S, Yan S, Zalzala F, Chuang R, Huang CJ, Looh H, Shiung CC, Nikitin YG, Tabarev AV, Tishkin AA, Lin S, Sun ZY, Wu XM, Yang TL, Hu X, Chen L, Du H, Bayarsaikhan J, Mijiddorj E, Erdenebaatar D, Iderkhangai TO, Myagmar E, Kanzawa-Kiriyama H, Nishino M, Shinoda KI, Shubina OA, Guo J, Cai W, Deng Q, Kang L, Li D, Li D, Lin R, Nini, Shrestha R, Wang LX, Wei L, Xie G, Yao H, Zhang M, He G, Yang X, Hu R, Robbeets M, Schiffels S, Kennett DJ, Jin L, Li H, Krause J, Pinhasi R, Reich D. Genomic insights into the formation of human populations in East Asia, *Nature*. **591** (2021) 413–419. doi: 10.1038/s41586-021-03336-2.
92. Williams MP, Flegontov P, Maier R, Huber CD. Testing times: Disentangling admixture histories in recent and complex demographies using ancient DNA. *Genetics*. (2024) iyae110. doi:

10.1093/genetics/iyae110.
93. Yaka R, Mapelli I, Kaptan D, Doğu A, Chyleński M, Erdal ÖD, Koptekin D, Vural KB, Bayliss A, Mazzucato C, Fer E, Çokoğlu SS, Lagerholm VK, Krzewińska M, Karamurat C, Gemici HC, Sevkar A, Dağtaş ND, Kılınç GM, Adams D, Munters AR, Sağlıcan E, Milella M, Schotsmans EMJ, Yurtman E, Çetin M, Yorulmaz S, Altınışık NE, Ghalichi A, Juras A, Bilgin CC, Günther T, Storå J, Jakobsson M, de Kleijn M, Mustafaoğlu G, Fairbairn A, Pearson J, Togan İ, Kayacan N, Marciniak A, Larsen CS, Hodder I, Atakuman Ç, Pilloud M, Sürer E, Gerritsen F, Özbal R, Baird D, Erdal YS, Duru G, Özbaşaran M, Haddow SD, Knüsel CJ, Götherström A, Özer F, Somel M. Variable kinship patterns in Neolithic Anatolia revealed by ancient genomes. *Curr. Biol*. **31** (2021) 2455–2468.e18. doi: 10.1016/j.cub.2021.03.050.
94. Yang MA, Fan X, Sun B, Chen C, Lang J, Ko YC, Tsang CH, Chiu H, Wang T, Bao Q, Wu X, Hajdinjak M, Ko AM, Ding M, Cao P, Yang R, Liu F, Nickel B, Dai Q, Feng X, Zhang L, Sun C, Ning C, Zeng W, Zhao Y, Zhang M, Gao X, Cui Y, Reich D, Stoneking M, Fu Q. Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science*. **369** (2020) 282–288. doi: 10.1126/science.aba0909.
95. Zeng TC, Vyazov LA, Kim A*, Flegontov P, Sirak K, Maier R, Lazaridis I, Akbari A, Frachetti M, Tishkin AA, Ryabogina NE, Agapov SA, Agapov DS, Alekseev AN, Boeskorov GG, Derevianko AP, Dyakonov VM, Enshin DN, Fribus AV, Frolov YV, Grushin SP, Khokhlov AA, Kiryushin KY, Kiryushin YF, Kitov EP, Kosintsev P, Kovtun IV, Makarov NP, Morozov VV, Nikolaev EN, Rykun MP, Savenkova TM, Shchelchkova MV, Shirokov V, Skochina SN, Sherstobitova OS, Slepchenko SM, Solodovnikov KN, Solovyova EN, Stepanov AD, Timoshchenko AA, Vdovin AS, Vybornov AV, Balanovska EV, Dryomov S, Hellenthal G, Kidd K, Krause J, Starikovskaya E, Sukenik R, Tatarinova T, Thomas MG, Zhabagin M, Callan K, Cheronet O, Fernandes D, Keating D, Francesca C, Iliev L, Kearns A, Özdoğan KT, Mah M, Micco A, Michel M, Olalde I, Zalzala F, Mallick S, Rohland N, Pinhasi R, Narasimhan V, Reich D. Postglacial genomes from foragers across Northern Eurasia reveal prehistoric mobility associated with the spread of the Uralic and Yeniseian languages. *bioRxiv*. (2023). doi: 10.1101/2023.10.01.560332.
96. Zhang F, Ning C, Scott A, Fu Q, Bjørn R, Li W, Wei D, Wang W, Fan L, Abuduresule I, Hu X, Ruan Q, Niyazi A, Dong G, Cao P, Liu F, Dai Q, Feng X, Yang R, Tang Z, Ma P, Li C, Gao S, Xu Y, Wu S, Wen S, Zhu H, Zhou H, Robbeets M, Kumar V, Krause J, Warinner C, Jeong C, Cui Y. The genomic origins of the Bronze Age Tarim Basin mummies. *Nature*. **599** (2021) 256–261. doi: 10.1038/s41586-021-04052-7.

References (for suppl. text)

97. de Barros Damgaard P, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, Peyrot M, Barjamovic G, Rasmussen S, Zacho C, Baimukhanov N, Zaibert V, Merz V, Biddanda A, Merz I, Loman V, Evdokimov V, Usmanova E, Hemphill B, Seguin-Orlando A, Yediay FE, Ullah I, Sjögren KG, Iversen KH, Choin J, de la Fuente C, Ilardo M, Schroeder H, Moiseyev V, Gromov A, Polyakov A, Omura S, Senyurt SY, Ahmad H, McKenzie C, Margaryan A, Hameed A, Samad A, Gul N, Khokhar MH, Goriunova OI, Bazaliiskii VI, Novembre J, Weber AW, Orlando L, Allentoft ME, Nielsen R, Kristiansen K, Sikora M, Outram AK, Durbin R, Willerslev E. The first horse herders and the impact of early Bronze Age steppe expansions into Asia, *Science*. **360** (2018) eaar7711. doi: 10.1126/science.aar7711.
98. Furtwängler A, Rohrlach AB, Lamnidis TC, Papac L, Neumann GU, Siebke I, Reiter E, Steuri N, Hald J, Denaire A, Schnitzler B, Wahl J, Ramstein M, Schuenemann VJ, Stockhammer PW, Hafner A, Lösch S, Haak W, Schiffels S, Krause J. Ancient genomes reveal social and genetic structure of Late Neolithic Switzerland. *Nat. Commun*. **11** (2020) 1915. doi: 10.1038/s41467-020-15560-x.
99. Jeong C, Balanovsky O, Lukianova E, Kahbatkyzy N, Flegontov P, Zaporozhchenko V, Immel A, Wang CC, Ixan O, Khussainova E, Bekmanov B, Zaibert V, Lavryashina M, Pocheshkhova E,

Yusupov Y, Agdzhoyan A, Koshel S, Bukin A, Nymadawa P, Turdikulova S, Dalimova D, Churnosov M, Skhalyakho R, Daragan D, Bogunov Y, Bogunova A, Shtrunov A, Dubova N, Zhabagin M, Yepiskoposyan L, Churakov V, Pislegin N, Damba L, Saroyants L, Dibirova K, Atramentova L, Utevska O, Idrisov E, Kamenshchikova E, Evseeva I, Metspalu M, Outram AK, Robbeets M, Djansugurova L, Balanovska E, Schiffels S, Haak W, Reich D, Krause J. The genetic history of admixture across inner Eurasia, *Nat. Ecol. Evol*. **3** (2019) 966–976. doi: 10.1038/s41559-019-0878-2.

100. Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas AS, Sikora M, Reuther JD, Irish JD, Malhi RS, Orlando L, Song YS, Nielsen R, Meltzer DJ, Willerslev E. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*. **553** (2018) 203–207. doi: 10.1038/nature25173.

101. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW Jr, Orlando L, Metspalu E, Karmin M, Tambets K, Rootsi S, Mägi R, Campos PF, Balanovska E, Balanovsky O, Khusnutdinova E, Litvinov S, Osipova LP, Fedorova SA, Voevoda MI, DeGiorgio M, Sicheritz-Ponten T, Brunak S, Demeshchenko S, Kivisild T, Villems R, Nielsen R, Jakobsson M, Willerslev E. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*. **505** (2014) 87–91. doi: 10.1038/nature12736.

102. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MT, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen C, Pierre TL, Grønnow B, Meldgaard M, Andreasen C, Fedorova SA, Osipova LP, Higham TF, Ramsey CB, Hansen TV, Nielsen FC, Crawford MH, Brunak S, Sicheritz-Pontén T, Villems R, Nielsen R, Krogh A, Wang J, Willerslev E. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*. **463** (2010) 757–762. doi: 10.1038/nature08835.