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, Chittussiho 1077/10, 710 00, Ostrava, Czechia*

[2] *Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branišovská 1160/31, 370 05, České Budějovice, Czechia*

[3] *Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Beutenbergstraße 11, 07745, Jena, Germany*

[4] *Department of Biological Sciences, Middle East Technical University, Üniversiteler Mahallesi, Dumlupınar Bulvarı No: 1, 06800, Çankaya/Ankara, Türkiye*

[5] *Department of Biology, Eberly College of Science, Pennsylvania State University, 168 Curtin Road, University Park, PA 16802, USA*

[6] *Department of Human Evolutionary Biology, Harvard University, 11 Divinity Avenue, Cambridge, MA 02138, USA*

* The authors contributed equally

§ Correspondence to: Pavel Flegontov

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

Running head: Performance of *qpAdm* on complex genetic histories

Keywords: archaeogenetics, genetic admixture, *qpAdm*, simulation, admixture graphs, stepping-stone models

Abstract

qpAdm is a statistical tool that is often used for testing large sets of alternative admixture models for a target population. Despite its popularity, *qpAdm* remains untested on two-dimensional stepping-stone landscapes and in situations with low pre-study odds (low ratio of true to false models). We tested high-throughput *qpAdm* protocols with typical properties such as number of source combinations per target, model complexity, model feasibility criteria, etc. Those protocols were applied to admixture-graph-shaped and stepping-stone simulated histories sampled randomly or systematically. We demonstrate that false discovery rates of high-throughput *qpAdm* protocols exceed 50% for many parameter combinations since: 1) pre-study odds are low and fall rapidly with increasing model complexity; 2) complex migration networks violate the assumptions of the method, hence there is poor correlation between *qpAdm* *p*-values and model optimality, contributing to low but non-zero false positive rate and low power; 3) although admixture fraction estimates between 0 and 1 are largely restricted to symmetric configurations of sources around a target, a small fraction of asymmetric highly non-optimal models have estimates in the same interval, contributing to the false positive rate. We also re-interpret large sets of *qpAdm* models from two studies in terms of source-target distance and symmetry and suggest improvements to *qpAdm* protocols: 1) temporal stratification of targets and proxy sources in the case of admixture-graph-shaped histories; 2) focused exploration of few models for increasing pre-study odds; 3) dense landscape sampling for increasing power and stringent conditions on estimated admixture fractions for decreasing the false positive rate.

Article Summary

Proliferation in the archaeogenetic literature of protocols for detection of admixed groups based a so-called *qpAdm* algorithm became disconnected from performance testing: the only extensive study of *qpAdm* on simulated data showed that it performs well under an unrealistically simple demographic scenario. We found that false discoveries of gene flows by *qpAdm* on a collection of random admixture-graph-shaped histories and on complex stepping-stone landscapes are very common and provide guidelines for design of *qpAdm* protocols in archaeogenetic studies.

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, 2022; Soraggi and Wiuf 2019) 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 have been devoted to testing the performance of these diverse methods on simulated genetic data (Alexander *et al.* 2009; McVean 2009; Martin *et al.* 2015; Lazaridis *et al.* 2017; Moreno-Mayar *et al.* 2018; Soraggi and Wiuf 2019; Ning *et al.* 2020b; Harney *et al.* 2021; Speidel *et al.* 2025), 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; Novembre and Stephens 2008; Duforet-Frebourg and Slatkin 2016).

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 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). Formal tests for genetic admixture such as f_3-statistics and D/f_4-statistics are often used to test for the possibility that a certain cline spotted in PC space might be a result of migration and admixture of previously isolated ancestries and does not reflect isolation by distance or recurrent bottlenecks (Novembre and Stephens 2008; Frichot *et al.* 2012; Duforet-Frebourg and Slatkin 2016; Estavoyer and François 2022). *D*- and f_4-statistics, which are identical except for the denominator and are not affected by genetic drift, test if an

89 unrooted four-population tree fits the data (Reich *et al.* 2009; Green *et al.* 2010; Durand *et*
90 *al.* 2011; Patterson *et al.* 2012). A statistically significant deviation of the statistic from 0
91 (estimated using jackknife or bootstrap resampling) means that either the assumed tree
92 topology is wrong, or gene flow occurred between a pair of branches in the tree, assuming
93 that at most one mutation per locus occurred in the history of the four populations (Durand
94 *et al.* 2011; Patterson *et al.* 2012) and SNP ascertainment bias (Flegontov *et al.* 2023) is
95 absent. However, interpretation of these statistics is ambiguous since gene flow directionality
96 remains unknown, and two pairs of branches can be responsible for a deviation of the statistic
97 from 0 (Lipson 2020). Since gene flow may be mediated by ghost groups related only distantly
98 to the sampled groups at the branch tips (Tricou *et al.* 2022), excluding one pair of branches
99 due to geographical and temporal plausibility of gene flow is also difficult. And interpretating
100 deviations of *D*- and f_4-statistics from 0 (or lack thereof) becomes almost impossible if both
101 branch pairs are connected by detectable gene flows, a typical situation on isolation-by-
102 distance landscapes. In general, unambiguous interpretation of non-rejection of the null
103 hypothesis (no deviation from 0) should be avoided.

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

113 Considering these complications, more sophisticated tests for genetic admixture are needed.
114 The *qpAdm* method introduced by Haak *et al.* (2015) was developed with the admixture graph
115 paradigm in mind. It is based on matrices of f_4-statistics and does not require detailed
116 knowledge of population phylogeny beyond a few assumptions (Lazaridis *et al.* 2016; Harney
117 *et al.* 2021). Recently, *qpAdm* was updated to accommodate more accurate f_4-statistics
118 derived from genome-wide genealogical trees for haplotypes rather than from allele
119 frequencies (Speidel *et al.* 2025). The *qpAdm* method tests admixture models in the form of

120 combinations of proxy ancestral groups ("sources" or "references"; Lazaridis *et al.* 2016) for
121 a "target" (or "test") population, given a genetically diverse set of "outgroup" populations,
122 and infers ancestry proportions in the target group contributed by the lineages represented
123 by the proxy sources ("outgroups" are often termed "right" populations for convenience since
124 they are usually not outgroups in the phylogenetic sense, and they were termed "references"
125 by Harney *et al.* 2021). See **Box 1** for definitions of various *qpAdm*-related terms used in this
126 study.

127
128 **Box 1.** Terminology used in this study for describing simulations, admixture models, results of screens
129 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 or isolation-by-resistance 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 François 2022), or via a multifurcation, i.e. star radiation (in this study). 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. Thus, gene-flow intensities are non-uniform and our SSL approximate isolation-by-resistance landscapes (McRae 2006). In this study all SSL are based on finite hexagonal lattices where demes are located at vertices of triangles tiling the plane (not be confused with honeycomb tiling), and node degree varies from 3 to 6. We note that rapid long-distance migrations (gene-flow edges connecting non-neighboring demes directly) are absent in such simulations.
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 the 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 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 up to 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 criteria above.
predicted 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 predicted positive (feasible, fitting).
false discovery rate (FDR)	The fraction of feasible admixture models of a certain complexity (e.g., two-way) that are classified as false in the case of AGS and as non-optimal in the case of SSL simulations. For topological criteria used for classifying admixture models on AGS histories into true and false ones see Results and Methods. For approaches to calculating FDR in the SSL context, see the definition for "optimal and non-optimal models" in **Box 2**. A general formula for FDR is as follows: $FDR = 1 - PPV = \frac{FPR}{\left(\frac{true}{false} - FNR \times \frac{true}{false} + FPR\right)}$ (***equation 1***) • PPV is positive predictive value; • FPR is false positive rate, also known as type I error probability (α) = 1 – true negative rate (TNR); • FNR is false negative rate, also known as type II error probability (β) = 1 – true positive rate (TPR) = 1 – power; • $\frac{true}{false}$ is the ratio of the number of true or optimal models to false or non-optimal models in the set of tested models, also known as pre-study odds (Ioannidis 2005).
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.
high-throughput *qpAdm* protocol or admixture screen	A protocol which for a target group tests all possible combinations of *k* proxy sources from a set of *n* alternative sources. When proxy sources are sampled over a large genetic landscape (in an exploratory study) and not from a small list of probable ancestry sources for a target (in a focused study), such study designs lead to low pre-study odds, especially for large *k*, and hence to high FDR according to ***equation 1*** above.
"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. 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 (Harney *et al.* 2021).
"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.
"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 a 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 model).

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 a target's history and average number of admixture events in sources' histories (In the context of an AGS history).

Ten 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, 2022; 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, 2022b; Gnecchi-Ruscone *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.* 2025 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).

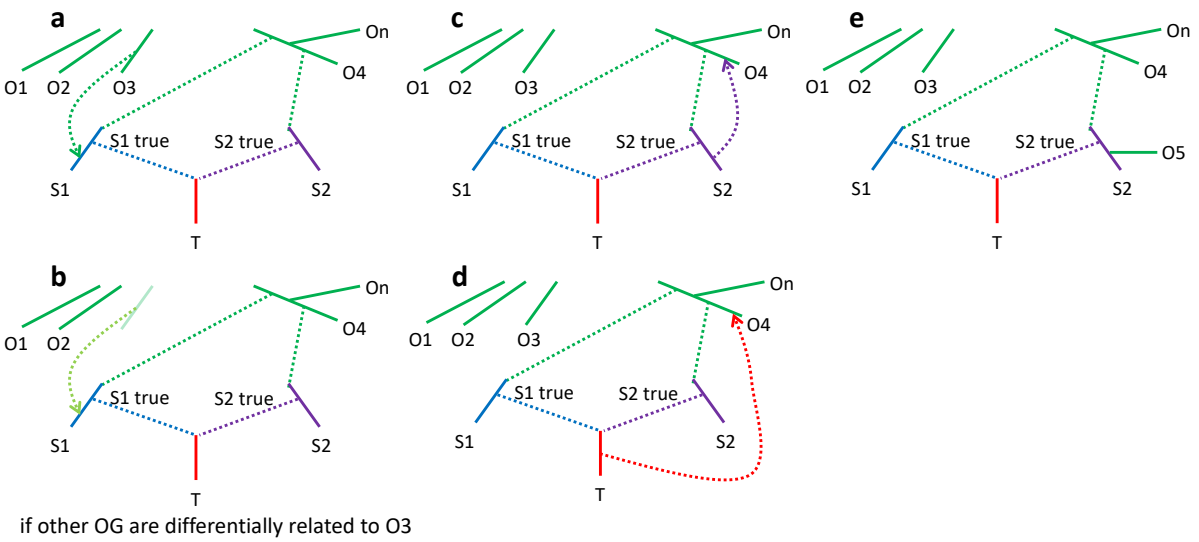


Figure 1. Admixture graphs showing an exhaustive list of assumption violations of the standard *qpAdm* protocol that may lead to rejection of the true simple model, and thus prompt the researcher to test overly complex models. (**a**) A gene flow from an outgroup O* to a proxy source after the divergence

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. Abbreviations: OG, outgroups.

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

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

181 Harney *et al.* (2021) demonstrated on AGS simulated histories that, if the *qpAdm* assumptions
182 are satisfied, it is highly favorable for statistical power of the method (for distinguishing
183 between alternative proxy sources that are unequally distant genetically from the true
184 ancestry source) to move at least some alternative proxy sources between the "left" and
185 "right" sets. In other words, having "on the right" populations that do not violate the
186 topological assumptions of *qpAdm*, but are closely related to proxy sources "on the left",
187 increases the statistical power greatly (see also Ning *et al.* 2020b and Williams *et al.* 2024 for
188 demonstrations of this on simple and complex AGS simulated histories, respectively, and
189 Speidel *et al.* 2025 for another exploration of a *qpAdm*-like algorithm on simple AGS
190 simulations).

191 This new type of *qpAdm* protocol, termed the "rotating" protocol, has been adopted in
192 archaeogenetics widely (see, e.g., Skoglund *et al.* 2017; Harney *et al.* 2019; Narasimhan *et al.*
193 2019; Olalde *et al.* 2019; Carlhoff *et al.* 2021; Fernandes *et al.* 2021; Librado *et al.* 2021;
194 Bergström *et al.* 2022; Lazaridis *et al.* 2022; Oliveira *et al.* 2022; Taylor *et al.* 2023; Zeng *et al.*
195 2023; Speidel *et al.* 2025). The most extreme version of the "rotating" protocol simply divides
196 a set of reference populations into all possible combinations of "right" and "proxy source"
197 subsets of certain sizes and rotates these combinations through successive *qpAdm* analyses.

198 Additional constraints can be placed on the rotating combinations such as restricting a set of
199 groups (usually highly divergent from the target) to remain in the “right” set in all models.
200 When evaluating the totality of multiple *qpAdm* tests, the simplest feasible models (e.g., one-
201 way, i.e., unadmixed) are usually favored, and increasingly complex models are explored
202 upon the rejection of simpler models. Model rejection is made according to a chosen *p*-value
203 threshold such that *qpAdm* models are considered feasible or "fitting" the data when the *p*-
204 value is above such a threshold (Skoglund *et al.* 2017; Harney *et al.* 2018; Narasimhan *et al.*
205 2019; Olalde *et al.* 2019; Yang *et al.* 2020; Carlhoff *et al.* 2021; Fernandes *et al.* 2021; Librado
206 *et al.* 2021; Zhang *et al.* 2021; Bergström *et al.* 2022; Lazaridis *et al.* 2022; Oliveira *et al.* 2022;
207 Nakatsuka *et al.* 2023; Taylor *et al.* 2023; Zeng *et al.* 2023; Kerdoncuff *et al.* 2024; Speidel *et*
208 *al.* 2025). As an additional criterion of a fitting model, all estimated admixture proportions
209 (see, e.g., Harney *et al.* 2018; Olalde *et al.* 2019; Yang *et al.* 2020; Zhang *et al.* 2021; Lazaridis
210 *et al.* 2022; Oliveira *et al.* 2022; Carlhoff *et al.* 2023; Zeng *et al.* 2023; Kerdoncuff *et al.* 2024;
211 Speidel *et al.* 2025), or proportions $\pm$ 2 standard errors (Narasimhan *et al.* 2019), may be

212 required to lie between 0 and 1. We note that estimated admixture fractions (abbreviated as
213 EAF) not restricted to [0, 1] are reported by *qpAdm* by default, and this setting is used in a
214 great majority of published studies (see Harney *et al.* 2021 for a justification of this unusual
215 property of the algorithm). It is important to remember that the statistical significance of the
216 *qpAdm/qpWave* test is, strictly speaking, a function of model rejection, and thus the failure
217 to reject a model may have underlying causes other than approximating the true situation
218 well enough (such as lack of statistical power or a lack of suitable "right" groups that capture
219 the divergent ancestry sources amongst the "left" set of populations).

220 A less exhaustive version of the rotating *qpAdm* protocol, termed "model competition" (e.g.,
221 Narasimhan *et al.* 2019; Fernandes *et al.* 2020; Carlhoff *et al.* 2021; Sirak *et al.* 2021; Zhang
222 *et al.* 2021; Maróti *et al.* 2022; Brielle *et al.* 2023; Lee *et al.* 2023), is used as widely as the
223 basic rotating protocol. It involves an initial (standard non-rotating) *qpAdm* analysis on a few
224 source populations (see **Box 1**). Upon identifying a list of plausible sources for a target, the
225 algorithm re-tests feasible models for this target rotating these plausible sources between
226 the "left" and "right" sets with the expectation of improving the power to reject models
227 including proxy sources that are genetically distant from the true sources.

228 The rotating *qpAdm* protocol and model competition are increasingly used as central
229 methods for testing admixture hypotheses proposed after inspecting distributions of
230 individuals in PC spaces, outcomes of *ADMIXTURE* analyses, and *f/D*-statistics indicative of a
231 gene flow rather than a simple tree relationship. Yet, the first study reporting detailed testing
232 of *qpAdm* on simulated data (Harney *et al.* 2021) was performed in extremely favorable
233 conditions: the simulated AGS history included just two non-nested admixture events; the
234 sources for the principal target group diverged about 1,200 generations ago (almost 35,000
235 years ago in the case of humans); the proxy sources were strictly cladal with the actual
236 ancestral groups for the target; several groups differentially related to these ancestry sources
237 were available; the simulated data were free of ascertainment bias since sites were sampled
238 in a random way; one million sites were used for most analyses; and only 50/50% simulated
239 admixture proportions were tested for some analyses. That study confirmed that the method
240 behaves as expected under these ideal conditions and offered some important guidelines on
241 the choice and number of "right" populations for optimal specificity of the method and on
242 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 for the jackknife algorithm, and to a particular type of SNP ascertainment bias: 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 (Harney *et al.* 2021) or nine (Speidel *et al.* 2025) groups was explored in the few studies aimed at testing *qpAdm* performance. Meanwhile, false discovery rate (FDR) of the method and violations of the topological assumptions of *qpAdm* (**Fig. 1**) remained virtually unexplored, and two-dimensional stepping-stone landscapes (SSL, see **Box 1**) and FDR in situations when the ratio of true to false models in the set of models tested (also termed “pre-study odds” by Ioannidis 2005, see **Box 1**) is low were completely untouched by these simulation studies (Harney *et al.* 2021; Speidel *et al.* 2025). 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 or two-dimensional SSL? This question motivated the current study exploring both simulated AGS and SSL histories and a companion paper focused on AGS histories (Williams *et al.* 2024).

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

Relying on general principles, we argue that any high-throughput *qpAdm* protocol (**Box 1**) on poorly understood complex demographic relationships approximated as admixture graphs and/or SSL is questionable as a formal test for admixture since a *p*-value threshold is used to *reject*, but not to *accept* models, and it is safer to interpret those models as a certain number of gene flows connecting "left" and "right" sets in any direction, not in the form of proxy sources and admixture proportions for a target. A model feasibility criterion including both

p-values and EAF is a complex construct relying on the topological assumptions outlined in **Fig. 1**. We expect that taking "left" and "right" sets that are not well-separated in time or contemporaneous (**Fig. S1**), and where relationships among groups are poorly understood (which is almost always true for exploratory studies), enriches the system for “left-to-right” gene flows, which in turn leads to frequent rejection of true simple admixture models. Since the behavior of *qpAdm* EAF under diverse demographic scenarios is poorly understood, it is possible that a large fraction of non-rejected complex models emerges as feasible, resulting in false signals of admixture. Another approach probably leading to the same outcome (enriching the system for “left-to-right” gene flows and other violations of the topological assumptions) is selection of pairs of “right” populations maximizing deviations of 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 or “left-to-right” gene flows are impossible to tell apart. This approach is sometimes used for 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 two-dimensional 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 grid of fixed size and test various types of high-throughput *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 re-analyzed from the SSL perspective large sets of published *qpAdm* models based on real data (Zeng *et al.* 2023; Speidel *et al.* 2024) and interpreted the resulting patterns using insights from the extensive exploration of simulated SSL data.

Results

Since the space of simulation parameters, *qpAdm* protocols, and interpretative approaches we explore is complex, we first provide a section-by-section guide to our results (see **Boxes 1** and **2** for a terminology glossary):

1. *AGS histories: approach to simulation and testing of qpAdm protocols* – in this section we introduce our AGS simulations and four high-throughput *qpAdm* protocols (rotating and non-rotating, distal and proximal).
2. *AGS histories: case studies illustrating false and true feasible qpAdm models* – we introduce our classification of fitting (feasible) models into true and false ones that is based on a set of topological criteria. We discuss selected false and true fitting *qpAdm* models along with PCA and *ADMIXTURE* results for the respective simulated demes.
3. *AGS histories: influence of temporal stratification and data amount/quality on the performance of qpAdm protocols* – we first make an overview of all *qpAdm* testing outcomes for two-way (two-source) admixture models by placing them in the coordinates of *qpAdm p*-values and a metric derived from EAF. Next, we apply a composite *qpAdm* model feasibility criterion borrowed from the literature and show that, given the pre-study odds in our experiments with 12 or 6 alternative proxy sources per target, FDR exceeds 50% when proxy sources have admixture history more complex than that of the target, and that situation is more common in proximal than in distal admixture models.
4. *Introducing SSL histories* – we describe a very different kind of simulations: two-dimensional stepping-stone landscapes based on a hexagonal lattice (formed by triangles tiling the plane), with non-uniform gene flow intensities between neighboring demes.
5. *SSL histories: randomized sampling of landscapes* – we explain why sets of targets, proxy sources and outgroups drawn randomly from a two-dimensional SSL make a good conservative baseline for a typical archaeogenetic sample set. We introduce *qpAdm* protocols on SSL that reproduce high-throughput *qpAdm* protocols popular in the literature (with a typical range of model complexities and typical numbers of source combinations per target) and result in low pre-study odds for complex models.
    a. *qpAdm performance at the level of models: model optimality metrics in spaces of estimated admixture fractions and p-values* – Since strict delineation of true and false models in the SSL context is impossible, we make a detour in our analyses and start

from a bird's-eye view on millions of two- to four-way randomized *qpAdm* models we tested on various types of SSL. For that, we place them in the space of *p*-values vs. EAF introduced in *section 3* and inspect how geometric properties of admixture models (also termed "model optimality metrics") map on this space. We show that a great majority of models with all EAF between 0 and 1 correspond to symmetric arrangements of sources around targets.

b. *qpAdm performance at the level of models: feasible models in spaces of model optimality metrics* – we apply various composite model feasibility criteria (based on both EAF and *p*-values, as is typically done in the *qpAdm* literature) and place feasible models in a space formed by key optimality metrics introduced above: maximal or average source-target (ST) distance and minimal source-target-source (STS) angle. We juxtapose in this space sets of models to be tested and sets of feasible models, define a threshold separating optimal from non-optimal models and proceed to calculate pre-study odds, false positive rate (FPR), false negative rate (FNR), and FDR. We show that it is usually impossible to extract a needle (models with symmetric sources close to the target) from a haystack of models with asymmetrically arranged distant sources, given that the variance of *p*-values is high (resulting in high FNR and low but non-zero FPR) and that EAF between 0 and 1 are reported for a small percentage of these abundant non-optimal models (also contributing to FPR).

c. *qpAdm performance at the level of experiments* – we go beyond individual admixture models and look at "experiments", an approach common in the *qpAdm* literature: all possible source combinations are tested for a target, moving from simple to more complex models, and an experiment is stopped as soon as at least one model satisfying a composite feasibility criterion is found. We demonstrate that on randomly sampled SSL results of experiments are often misleading: depending on the choice of feasibility criteria and on SSL properties, either accurate but simple (two-way) models are found, or inaccurate complex (three- or four-way) models are reported, or experiments progress even further. We discuss factors influencing complexity of models at which experiments end, including average "left-right" spatial distance, and an important difference between rotating and non-rotating protocols in this respect.

364 6. *SSL histories: systematic sampling of landscapes* – we move from randomized to
365 systematic construction of admixture models on SSL, with proxy sources and outgroups
366 drawn from circles of demes at various distances around the target.
367 a. *Systematic sampling: model optimality metrics in the spaces of estimated admixture*
368 *fractions and p-values* – in this subsection we apply the analytic approaches
369 introduced in *sections 5a* and *5b* to the systematic models and explore the influence
370 of average "left-right" distances on *qpAdm* results.
371 b. *Systematic sampling: exploring truly ideal six-way models* – we move to even more
372 complex, six-way, models and show that separating highly non-optimal and optimal
373 complex models with *p*-value thresholds is achievable only with the non-rotating
374 *qpAdm* protocol but not the rotating one.
375 c. *Systematic sampling: qpAdm performance at the level of experiments* – we show that
376 high-throughput *qpAdm* protocols applied to SSL can produce counter-intuitive
377 results: when sampling of the target's vicinity or the whole landscape is dense, simple
378 (two-way) models are most frequent outcomes of *qpAdm* experiments, but when
379 sampling of the target's vicinity is poor, experiments tend to stop at much more
380 complex, and inaccurate, models. We discuss how these patterns are explained by
381 average "left-right" spatial distance influencing model *p*-values.
382 7. *Geographical analysis of large sets of qpAdm models on real data* – in this section we
383 explore geometric properties of admixture models from two sets of high-throughput
384 *qpAdm* experiments on real human archaeogenetic data (Zeng *et al.* 2023; Speidel *et al.*
385 2024). Interestingly, distributions of fitting *qpAdm* models in the spaces of optimality
386 metrics are similar on simulated and real landscapes matched by sampling sparsity. This
387 illustrates low pre-study odds and high FDR in published *qpAdm* admixture screens and
388 suggests that our conclusions derived on simulated SSL are relevant for critical
389 reassessment of the existing archaeogenetic literature.

390

391 ***1. AGS histories: approach to simulation and testing of qpAdm protocols***

392 Below we explore, relying on simulated AGS histories, performance (mainly FDR) of *qpAdm*
393 protocols representing the spectrum of protocols used in the literature (for an overview of
394 our AGS workflow see **Fig. S2**). The most extreme example is a protocol where all groups are

395 rotated and all are treated alternatively as outgroups, targets, and proxy sources, i.e., there
396 is no temporal stratification between the latter categories. We term this protocol "proximal
397 rotating". Although such an extreme situation is, to our knowledge, rare among published
398 *qpAdm* protocols (see Carlhoff *et al.* 2021; Oliveira *et al.* 2022; Zeng *et al.* 2023), we use it to
399 illustrate the potential effects of poor temporal stratification of targets and proxy sources in
400 published rotating protocols (**Fig. S1**, **Suppl. text 1**). Models with targets pre-dating proxy
401 sources are encountered in high-throughput *qpAdm* screens, but do not constitute a majority
402 of models (Narasimhan *et al.* 2019; Librado *et al.* 2021; Bergström *et al.* 2022; Lazaridis *et al.*
403 2022; Taylor *et al.* 2023; Zeng *et al.* 2023; Speidel *et al.* 2025). We also explore FDR of the
404 proximal non-rotating (Harney *et al.* 2018; van de Loosdrecht *et al.* 2018; Narasimhan *et al.*
405 2019; Prendergast *et al.* 2019; Sikora *et al.* 2019; Wang *et al.* 2020; Carlhoff *et al.* 2021; Wang
406 *et al.* 2021; Zhang *et al.* 2021; Changmai *et al.* 2022a, 2022b; Maróti *et al.* 2022; Brielle *et al.*
407 2023; Lee *et al.* 2023; Nakatsuka *et al.* 2023), distal rotating (Narasimhan *et al.* 2019; Librado
408 *et al.* 2021; Bergström *et al.* 2022; Lazaridis *et al.* 2022; Taylor *et al.* 2023; Zeng *et al.* 2023;
409 Speidel *et al.* 2025), and distal non-rotating protocols (Haak *et al.* 2015; Mathieson *et al.* 2015,
410 2018; Lazaridis *et al.* 2016; Antonio *et al.* 2019; Sikora *et al.* 2019; Marcus *et al.* 2020; Yang *et*
411 *al.* 2020; Papac *et al.* 2021; Yaka *et al.* 2021; Patterson *et al.* 2022). In the distal protocols,
412 only *qpAdm* models where the target group's sampling date is strictly contemporaneous with,
413 or post-dates, sampling of both proxy sources were considered.

414 We tested performance of these *qpAdm* protocols on relatively complex AGS simulated
415 histories: 13 populations connected with each other by admixture graphs of random
416 topology, including 10 pulse-like admixture events. Further details on the simulated
417 population histories are presented in **Fig. S2**, in Methods and illustrated by examples in **Figs.**
418 **2** and **S3**. To explore the influence of data amount and population divergence levels on
419 *qpAdm* performance, we generated three independent sets of ten simulation replicates for
420 each graph topology (referred to as setups no. 1, 2, and 3 below) and single simulation
421 replicates for two further setups (no. 4 and 5; see **Fig. S2** for details and **Fig. S4** for the number
422 of SNP loci polymorphic in the simulated datasets). The simulations with the 3,000-generation
423 maximal history depth resulted in median F_{ST} at the inter-continental levels of human genetic
424 differentiation (~0.1), and those with the 800-generation depth matched intra-continental
425 levels (median F_{ST} from 0.022 to 0.051, see **Fig. S5**).

426 A typical archaeogenetic dataset is composed of pseudo-haploid data with high proportions
427 of missing sites and with widely different group sizes, including singleton groups. To generate
428 more realistic "noisy" data, we also performed randomized subsampling of SNP datasets for
429 simulation setups no. 1, 2, and 3 (300-, 1,000-, and 3,000-Mbp-sized genomes, respectively),
430 for one simulation replicate per each setup (**Fig. S2**, see Methods for details). The resulting
431 data were pseudo-haploid, had missing data rates ranging from 5% to 95% across individuals,
432 and had uneven group sizes ranging from 1 to 10 individuals. Ten independent subsampled
433 datasets were generated for each simulated topology (for counts of polymorphic loci see **Fig.**
434 **S4**).

435 The *qpAdm* protocols we applied to the simulated data were focused on the simplest
436 admixture models, one- and two-way. The median number of two-source admixture events
437 in a population from our AGS simulations is 1 (average = 1.17), suggesting that two-way
438 models are broadly adequate for this type of history, especially considering that by far not all
439 internal branches in AGS histories were sampled (**Fig. S3**). Histories that are more complex
440 than two-way mixtures were also not rare in the data, accounting for 38% of all models tested
441 (**Table S1a**). The model feasibility criteria followed Narasimhan *et al.* (2019) (**Box 1**). Thus, we
442 tested all possible two-way admixture models for 40 complex population histories (34,320
443 proximal and distal rotating models per simulation setup and replicate).

444 The non-rotating *qpAdm* approach was implemented as follows: for each simulated graph the
445 six most ancient groups were selected as a fixed "right" set (ties were resolved in alphabetical
446 order; these "right" sets remained unchanged for a given graph topology across independent
447 simulations) and for the remaining seven groups all possible one-way and two-way admixture
448 models were tested (4,200 proximal and distal models per simulation setup and replicate).
449 Thus, models tested with the non-rotating protocol represent a subset of models tested with
450 the rotating protocol.

451

452 ***2. AGS histories: case studies illustrating false and true feasible qpAdm models***

453 In the context of complex and random admixture graph topologies it is hard to draw a strict
454 boundary between true and false admixture models composed of a target and two proxy
455 sources. However, we classified the feasible *qpAdm* models into false and true ones relying

456 on a set of rules. By far the most common class of false feasible *qpAdm* models (referred to
457 as "false positive" or FP models), comprising 50.9% of all FP models generated by the
458 proximal rotating protocol across all the simulation and subsampling replicates (for setups
459 no. 1 and 2), occurs when the target group is falsely rejected as forming a clade with one or
460 both proxy sources whilst they are simulated as clades. In other words, these are false
461 rejections of single-source models.

462 An example of such a situation is shown in **Fig. 2**. The clade relationship between the target
463 (*J*) and source (*M*) is rejected by the rotating protocol due to "left-to-right" gene flows
464 violating the topological assumptions of *qpAdm*, and more complex models ("*J = A + M*" and
465 "*J = L + M*") are evaluated as true. When the true one-way model "*J = M*" is tested, *M*
466 represents a proxy for the only true source of ancestry in *J*, and outgroups *A*, *C*, *D*, and *L* are
467 partially derived from the target lineage, resulting in rejection of the one-way model with *p*-
468 values from ~10^{-32} to ~10^{-50} (in the case of 1000-Mbp-sized genomes and high-quality data).
469 Rejections of two-way models "*J = A + M*" and "*J = L + M*" according to *p*-values are rare (**Fig.**
470 **2c**), and nearly all EAF ± 2 SE are between 0 and 1 even on low-quality pseudo-haploid data
471 (**Fig. 2d**), that is why these models are classified as fitting in nearly all cases (in 34 and 36 of
472 40 simulation/subsampling replicates, respectively; setups no. 1 and 2 were considered).
473 Groups *A* and *L* are 270 generations younger than *J*, and >50% of their ancestry was simulated
474 as derived from the *J* lineage (**Fig. 2a**). These FP models illustrate not only cladality rejection,
475 but also incorrect gene flow direction suggested by the *qpAdm* results.

476 These results illustrate risks of applying rotating *qpAdm* protocols to sets including ancient
477 groups "on the left" and much younger groups "on the right" (Carlhoff *et al.* 2021; Oliveira *et*
478 *al.* 2022), since those may be descended partially from the "left" groups (**Fig. 1c,d,e**) and thus
479 lead to *qpAdm* assumption violations and false positive findings of gene flow. PCA on four
480 datasets (**Fig. 2b**) and unsupervised *ADMIXTURE* results at some *K* values (**Fig. 2f**) support
481 these FP models. Thus, a researcher co-analyzing ancient and modern groups due to lack of
482 ancient samples from certain regions may encounter this situation in practice and, given
483 support by all methods in the standard archaeogenetic toolkit (*qpAdm*, PCA, and
484 *ADMIXTURE*), is likely to conclude that FP models are true. We illustrate another similar
485 situation often leading to rejection of cladality and acceptance of two-way or even more

complex models (in 33-37 of 40 simulation/subsampling replicates, depending on model and *qpAdm* protocol; setups no. 1 and 2 were considered) in **Fig. S6a**.

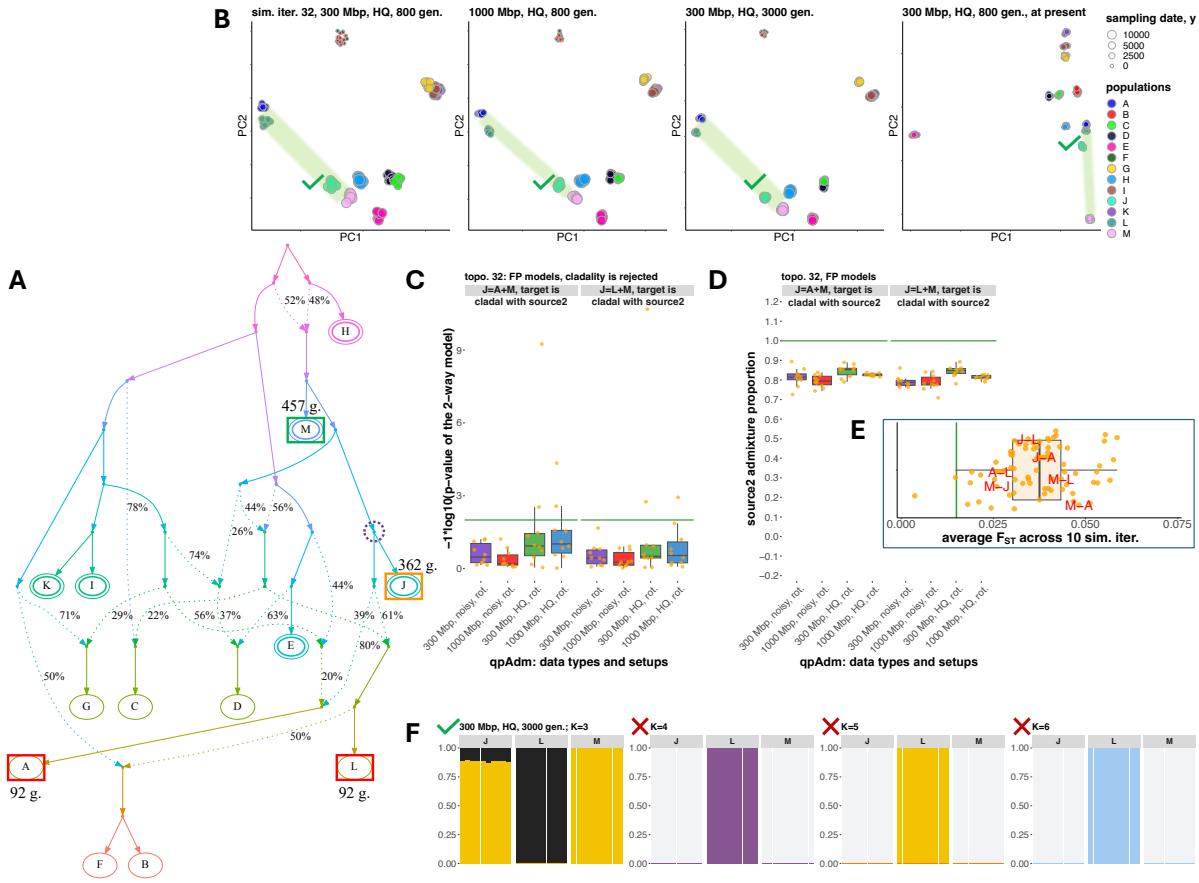


Figure 2. A case study illustrating the most common class of false positive *qpAdm* models supported by the proximal rotating protocol. Models of this type include at least one proxy ancestry source that is simulated as fully cladal with the target. The other proxy source may be simulated as a descendant of the target lineage (as shown here), may belong to an unrelated lineage (**Fig. S6a**), 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. S3**). 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 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 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,

510 grouped by simulation setups and *qpAdm* protocols as indicated on the x-axis. Green lines show the
511 *p*-value threshold of 0.01 used in this study for model rejection. The *qpAdm* models shown here were
512 not tested using the non-rotating protocol since the target falls into the set of six "right" populations
513 chosen for that protocol. **D.** Boxplots summarizing EAF across simulation or subsampling replicates,
514 grouped by simulation setups and *qpAdm* protocols. The admixture proportions are shown either for
515 the first or second proxy source, as indicated on the y-axis; green lines show the simulated admixture
516 proportion. **E.** All F_{ST} values for population pairs from the simulated history. Average F_{ST} values are
517 shown across 10 simulation replicates for 1000-Mbp-sized genomes and high-quality data. Population
518 pairs formed by components of the illustrated two-way *qpAdm* model(s) are labelled in red. The green
519 line shows median F_{ST} (0.015) for the Bronze Age West Asian populations analyzed by Lazaridis *et al.*
520 (2016) as an example of human genetic diversity at the sub-continental scale. **F.** Ancestry proportions
521 estimated with unsupervised *ADMIXTURE* for the groups constituting the *qpAdm* model(s). For
522 brevity, results are shown for one or two *qpAdm* models, for one simulated dataset indicated in the
523 plot title, and for four selected *K* values. If admixture model(s) are supported by this analysis at a given
524 *K* value, a green tick mark is placed beside the target group in the plot, and a red cross mark is placed
525 otherwise.

526

527 Other topological classes of FP models can be concisely described as follows (and are more
528 precisely defined in Methods): 1) gene flow goes from the target lineage (after the last
529 admixture event in its history) to a proxy source lineage (**Figs. 2** and **S6a,b**); 2) a proxy source
530 included in the model is symmetrically related to all real sources of ancestry in the target (**Fig.**
531 **S6c,d**), or both proxy sources represent the same true source and are symmetrically related
532 to all the other true sources (**Fig. S6e**); 3) both proxy sources represent distinct real sources,
533 however a proxy source is a heavily admixed group sharing less than 40% of ancestry with the
534 real source (**Fig. S6a,d**); 4) the target gets a gene flow from a deep-branching source not
535 represented by any sampled population, and an inappropriate proxy source is included in the
536 fitting model (**Fig. S6f,g**). Violations of the topological assumptions of *qpAdm* encountered in
537 these FP model classes and corresponding *qpAdm*, PCA, and *ADMIXTURE* results are
538 illustrated in **Suppl. text 2**. Definitions and examples of true positive (TP) models are also
539 presented in **Suppl. text 2**.

540

541 ***3. AGS histories: influence of temporal stratification and data amount/quality on the***
542 ***performance of qpAdm protocols***

543 To make a qualitative overview of all the *qpAdm* results on AGS histories and to gain intuition
544 about factors influencing frequency of false results, we placed models in the coordinates of
545 admixture fractions estimated by *qpAdm* (EAF) and *p*-values (**Fig. S7**). Both high- and low-

546 quality data generated by the most realistic simulation setup were considered (3,000-Mbp-
547 sized genomes, maximal population divergence depth at 3,000 generations). Since estimated
548 admixture fractions reported by *qpAdm* may vary from <<–1 to >>1 (see a justification for this
549 unusual property of the algorithm in Harney *et al.* 2021) and to make this visualization
550 approach scalable to models more complex than two-way, we use one number instead of
551 several EAF per an *n*-way model: maximal absolute difference between EAF in that model and
552 equal fractions $\frac{1}{n}$, abbreviated as max|*EAF–EF*| (**Box 1**). Each two-way model was tested many
553 times: on different simulation setups, data qualities, replicates and *qpAdm* protocols. We
554 explore the max|*EAF–EF*| vs. *p*-value spaces for three model classes separately: models that
555 were consistently non-fitting in all our analyses, models that were fitting at least once and
556 classified as false, and models that were fitting at least once and classified as true (**Fig. S7**).
557 For comparison, we also visualized models populating the space of EAF standard errors (SE)
558 reported by *qpAdm* and *p*-values (**Fig. S8**). Remarkably, the models consistently rejected (not
559 only on the simulations used for making **Fig. S7,** but throughout this study) demonstrated EAF
560 outside the (0, 1) interval in nearly all cases, especially on high-quality data, while models
561 that were fitting at least once and classified as either false or true are characterized by EAF
562 largely in the (0, 1) interval, but their *p*-values vary widely (**Fig. S7**). Of the consistently
563 rejected models, 40% are models whose target populations have 0 admixture events in their
564 history (**Table S1a**).

565 This qualitative overview suggests that EAF are much less variable than *p*-values, and we
566 quantify this using standard deviations (SD) of $\log_{10}$(*p*-value) or EAF across 10 (or 20, if a
567 model was evaluated under both rotating and non-rotating protocols)
568 simulation/subsampling replicates (**Fig. S9**). *P*-values of false, true, and consistently negative
569 models vary by many orders of magnitude between replicates, and SD grows with increasing
570 amount of data, up to ca. 30-60 orders of magnitude in the case of simulations that reproduce
571 typical archaeogenetic data most faithfully: noisy data derived from 3,000-Mbp-sized
572 genomes (**Fig. S9**). In contrast, median EAF SDs are consistently between 0.01 and 0.1 for
573 models that were fitting at least once in this study (with very small differences between false
574 and true models), and between 0.1 and 1 for consistently rejected models (**Fig. S9**). This
575 analysis confirms that there are two classes of two-way *qpAdm* models on AGS data that
576 encompass a great majority of all models: those with relatively stable EAF far from (0, 1) and

577 those with stable EAF within (0, 1). The latter models are universally fitting very rarely (**Fig.**
578 **S10**) due to the high variability of *qpAdm p*-values.

579 It is important to understand if some of these three model classes are enriched for certain
580 features of target's and proxy sources' admixture histories, and to understand how this is
581 shaped by the choice of *qpAdm* protocol and by temporal stratification of targets and proxy
582 sources. According to our definitions, proximal models, and especially proximal models
583 tested with the rotating protocol, are enriched for targets sampled deep in the past (see **Fig.**
584 **S7** for a mapping on the max|*EAF–EF*| vs. *p*-value space and **Table S1** for a quantitative
585 summary and statistical tests). As admixture events were simulated at random points in time
586 and placed randomly on the graphs, proximal models tested with the rotating protocol are
587 actually enriched for unadmixed targets (**Fig. S7**, **Table S1a**). In contrast, proximal non-
588 rotating and distal rotating/non-rotating models demonstrate unimodal, not monotonically
589 decaying, distributions of target history complexity (**Table S1a**).

590 For the same fundamental reasons, proximal models that were consistently non-fitting or
591 occasionally fitting and false are characterized by fewer admixture events in a target's history
592 as compared to average number of admixture events in sources' histories (median difference
593 is –1 in most cases), but this is not the case for proximal models that were occasionally fitting
594 and true (median difference is around 0, see **Table S1b**). This distinction between true
595 proximal models and the other two model classes is highly significant for both the rotating
596 and non-rotating protocols (**Table S1b**). Statistically significant trends in the same direction
597 are observed for distal models as well, but in this case median differences between number
598 of admixture events in a target's history and average number of admixture events in sources'
599 histories (abbreviated as AEH_T – *av.* AEH_S) are positive (**Table S1b**). The chance of
600 encountering violations of the *qpAdm* assumptions (**Fig. 1**) grows when admixture history of
601 proxy sources is more complex than that of a target group, explaining these results.
602 Considering all models on the 3,000-Mbp-sized simulations, AEH_T – *av.* AEH_S is just weakly
603 correlated to max|*EAF–EF*| or *p*-values: Spearman's correlation coefficients for various
604 *qpAdm* protocols and data qualities range from –0.49 to 0.3 in five practically relevant
605 sections of the max|*EAF–EF*| vs. *p*-value space (**Fig. S7**, **Table S1c**). Thus, although
606 false/consistently non-fitting and true models differ significantly according to the AEH_T – *av.*
607 AEH_S metric, using *p*-values or EAF for predicting which model is true and which is false is not

a fruitful approach, especially considering negative correlation between $AEH_T - av.\ AEH_S$ and *p*-values seen for some sections of this space and for some *qpAdm* protocols (**Table S1c**).

While FP *qpAdm* models are expected for complex genetic histories, the practical usage of high-throughput *qpAdm* protocols relies on an assumption that false positives are rare. As discussed above, due to a fundamental property of AGS histories where complexity (number of admixture events) grows with time, FDR of high-throughput *qpAdm* protocols is expected to depend on enrichment of the sets of tested models for those where proxy source history is more complex than that of the target. First, we calculated FDR on the complete set of simulated AGS histories, per each simulation/subsampling replicate, and stratified the results by amount and quality of data, by model types, and by *qpAdm* protocols (**Fig. 3**).

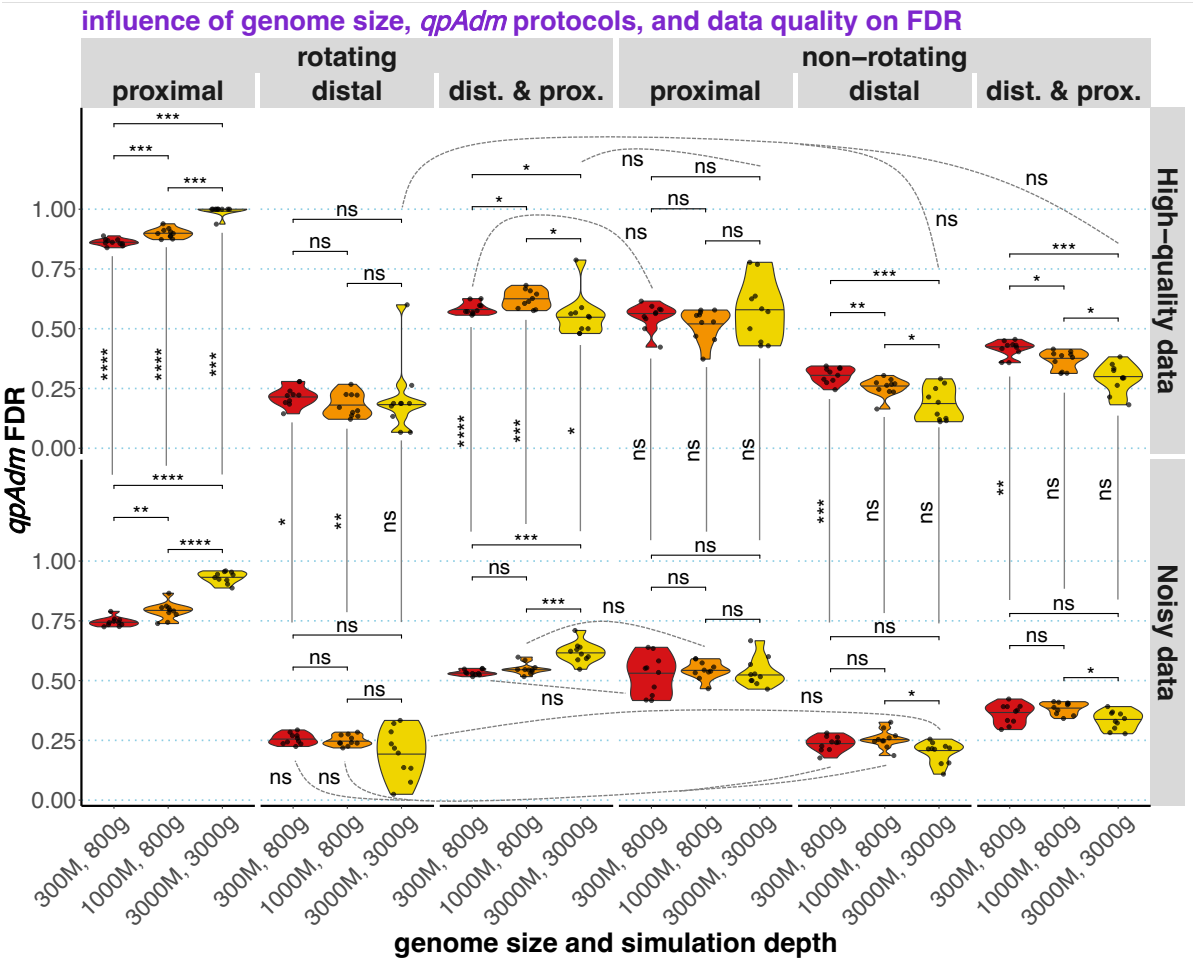


Figure 3. Distributions of FDR values over 10 simulation replicates (for high-quality data) or over 10 subsampling replicates derived from a single simulation replicate (for low-quality data). The distributions are summarized as violin plots with medians and stratified by *qpAdm* protocols (rotating and non-rotating), by model subsets (proximal, distal, or both model types included), by data quality, and by three combinations of simulated genome sizes (300, 1,000, or 3,000 Mbp) and maximal simulation depths (800 or 3,000 generations). The following statistical comparisons were performed

625 with the two-sided Wilcoxon test: across genome sizes, with protocol and data quality fixed (36 non-
626 paired tests); across data qualities, with protocol and genome size fixed (18 non-paired tests); and
627 across protocols, with genome size and data quality fixed (90 paired tests). Paired tests were used in
628 the latter case since the corresponding sets of models are not independent: non-rotating models form
629 a subset of rotating models, and distal and proximal models can be considered together. *P*-values
630 (adjusted for multiple testing using the Holm method) are coded in the following way: ns, >0.05; *,
631 ≤0.05; **, ≤0.01; ***, ≤0.001; ****, ≤0.0001. Most comparisons across protocols are significant at the
632 0.05 level (omitted for clarity), and only non-significant comparisons of this type are shown with
633 dashed lines.

634

635 Indeed, considering distal models only (tested with both rotating and non-rotating protocols)
636 results in the lowest median FDR across replicates, from 16.4% to 31.2% (**Fig. 3**). FDRs based
637 on proximal models tested with the rotating protocol varied from 72.5% to 100% and grew
638 significantly with increasing amount of data. A large fraction of false proximal models
639 emerges as fitting due to false rejections of one-way models because of violations of the
640 topological assumptions (**Fig. 1**), for instance, due to "left-to-right" gene flows, and these
641 prompt the investigator to test more complex two-way models, which are often fitting (**Suppl.**
642 **text 2**). And model rejection is more efficient when more data is available, explaining the
643 effect on FDR observed here.

644 FDRs based on proximal models tested with the non-rotating protocol were lower (medians
645 from 52.1% to 57.7%) and did not depend on the amount of data, suggesting that false model
646 rejections due to assumption violations play a less important role here, which is expected for
647 "right" and "left" population sets which are stratified chronologically. Similarly, the amount
648 of data had no statistically significant influence on FDR when distal models were considered,
649 except for the non-rotating protocol on high-quality data (increasing the amount of data
650 reduced FDR, **Fig. 3**). We also note that random subsampling of SNPs and individuals and a
651 less robust algorithm for calculating f_4-statistics on the low-quality data (with different
652 statistics calculated on different SNP sets) did not lead to dramatic increases/decreases in
653 FDR (**Fig. 3**). The observation that the decrease in FDR with increasing amount of data is small
654 and often non-significant is possibly explained in the following way. Positive models are
655 defined based on both *p*-values and EAF (**Box 1**): while EAF are relatively stable (**Fig. S9**), *p*-
656 values of individual models decrease exponentially with the amount of data (**Fig. S6**).
657 However, model rejection is not always "beneficial" for protocol's performance assessed via
658 FDR since both models with inappropriate proxy sources (false models) and true models with

659 violations of the topological assumptions (**Fig. 1**) are rejected more often when more data is
660 added.

661 In summary, even in optimal conditions (simulation depth of 87000 years on human
662 timescale, 3,000-Mbp-sized genomes, high-quality diploid data) we did not achieve negligible
663 FDR when distal models were tested on a collection of random AGS histories, and high data
664 amount/quality makes FDR based on proximal models grow. In nearly all conditions tested
665 there is very high variability in FDR across simulated AGS topologies, with many topologies
666 demonstrating FDR of either 0% or 100% (**Fig. S11**), and hence it is hard to predict if real
667 genetic histories are prone to generating FP signals of admixture when *qpAdm* protocols are
668 applied. For an illustration of three topologies most problematic for distal *qpAdm* protocols
669 see **Fig. S12**. These differences in *qpAdm* performance across topologies are probably driven
670 at least in part by the fact that some random AGS histories have much "denser" violations of
671 the *qpAdm* topological assumptions than others, but this frequency of assumption violations
672 was not quantified in our study since it requires another layer of manual topological analysis.

673 The fraction of two-way admixture models that are inappropriate according to our topological
674 criteria (the fraction of negatives) in a random sample of 400 proximal and distal rotating
675 models from all the simulated AGS histories is 82.3% (pre-study odds = 0.215), which allows
676 us to approximately estimate not only FDR (48%–78.7%), but the FPR of the proximal rotating
677 protocol = number of false positives / number of negatives per simulation replicate [858
678 models × 40 graphs × fraction of negatives] = 0.04%–2.6% across simulation parameter
679 combinations and replicates summarized in **Fig. 3**. The FNR of the proximal rotating protocol
680 = number of false negatives / number of positives per simulation replicate [858 models × 40
681 graphs × fraction of negatives] = 89.9% – 99.8%. Here are the same statistics for the distal
682 non-rotating protocol: the fraction of negatives in a random sample of 400 relevant models,
683 74.3% (pre-study odds = 0.346); total number of distal non-rotating two-way models across
684 all graph topologies, 1804; FDR across simulation parameter combinations and replicates
685 from **Fig. 3**, 10.9%–34.4%; FPR, 0.2%–3.2%; FNR, 75.6% – 96.5%. FDR depends on FPR and
686 FNR through ***equation 1*** which also includes pre-study odds (R). Given the relatively high R
687 for two-way models based on our AGS histories and low FPR, the principal contributor to the
688 high FDR values compromising historical interpretation of such screens for admixture is high
689 FNR via the term $R - FNR \times R$ in the denominator of ***equation 1***. And high FNR results from

690 false rejection of true two-way models due to violations of the topological assumptions (**Fig.**
691 **1**) and due to high variance in *p*-values across simulation/subsampling replicates (**Figs. S8,**
692 **S9**). Importantly, variance in *p*-values is especially high in the case of the most realistic
693 simulations: noisy data derived from 3,000-Mbp-sized genomes (**Fig. S9**).

694 As we show in **Suppl. text 3** (**Fig. S13, Tables S2** and **S3**), it is impossible to decrease FDR of
695 the proximal rotating protocol dramatically by combining it with other methods (PCA and
696 unsupervised *ADMIXTURE*) or by adjusting *p*-value thresholds. However, it is possible to
697 decrease FDR of distal *qpAdm* protocols to ca. 0% if positive *qpAdm* results are further
698 validated by PCA and the unsupervised *ADMIXTURE* (Alexander *et al.* 2009) methods (**Table**
699 **S4, Suppl. text 4**). In that supplemental section we also explore the performance of proximal
700 non-rotating *qpAdm* protocols combined with model competition (Narasimhan *et al.* 2019)
701 (**Table S5, Suppl. text 4**).

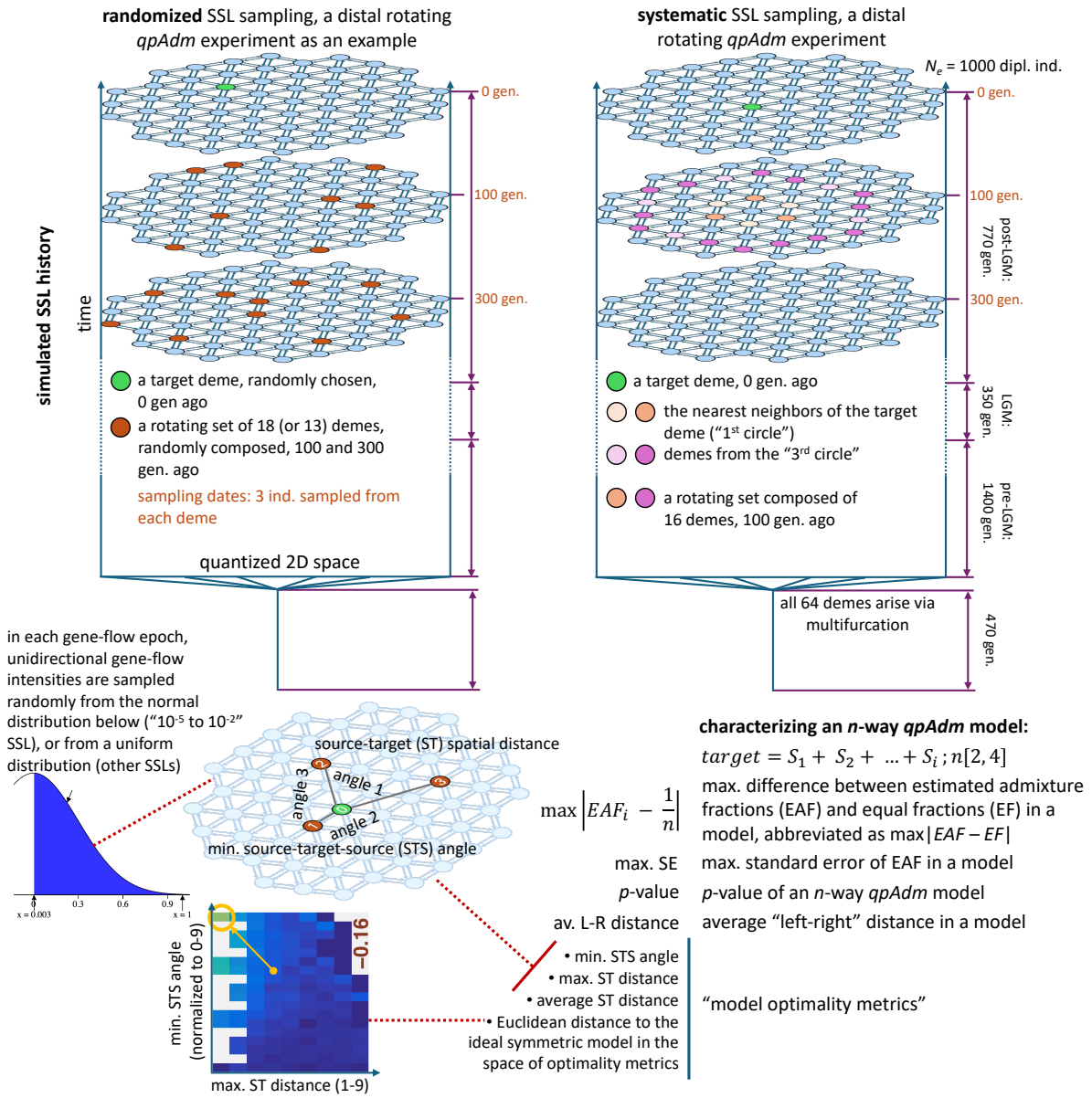
702

703 ***4. Introducing SSL histories***

704 Since effects of prolonged gene flows in the context of a simple AGS history were studied by
705 Harney *et al.* (2021), here we explore a more extreme case also investigated in that study
706 briefly, namely stepping-stone landscapes (SSL; **Box 1**). We simulated SSL that are more
707 realistic than uniform gene-flow scenarios and/or one-dimensional strings of demes
708 (Novembre and Stephens 2008; Duforet-Frebourg and Slatkin 2016; Harney *et al.* 2021;
709 Tournebize and Chikhi 2023; Speidel *et al.* 2025), namely two-dimensional landscapes with
710 gene-flow intensities that are non-uniform both in space and time (see similar stepping-stone
711 grids in, for example, Petkova *et al.* 2016 and Al-Asadi *et al.* 2019). Approximately circular
712 landscapes composed of 64 panmictic demes of constant effective size (1,000 diploid
713 individuals) were simulated as arising via multifurcation ("star radiation") and then evolving
714 for ca. 2,500 generations with non-uniform bidirectional gene-flows (**Fig. 4**).

715 Four types of SSL were simulated, characterized by varying ranges of gene-flow intensities,
716 with 10 simulation/gene-flow intensity replicates per range: (ca. 10^{-5} to 10^{-2}), (ca. 10^{-5} to 10^{-}
717 ^{4}), (ca. 10^{-4} to 10^{-3}), and (ca. 10^{-3} to 10^{-2}) (**Fig. S14b,c, Table S6**). Each intensity range
718 corresponds to gene-flow between two neighboring demes in one direction per generation.
719 An intensity value means that a given fraction of effective population size in a generation

720 comes from migrants from a specified neighboring deme. In all the cases, gene-flow
721 intensities were sampled at the start of each "gene-flow epoch" from uniform distributions,
722 except for the ca. 10^{-5} to 10^{-2} range: in that case intensities were sampled from a normal
723 distribution centered at 0 (**Table S6**).



724

725 **Figure 4.** Diagrams summarizing principles of SSL simulations, SSL-based *qpAdm* setups, and
726 approaches we used for interpreting *qpAdm* results in this context. Key features of the simulations
727 are as follows: 1) origin of all the demes via a multifurcation; 2) the 2D space is quantized into 64
728 demes of small and stable effective population size; 3) gene-flow intensities between neighboring
729 demes are generated by random sampling from normal or uniform distributions (in certain intervals),
730 at the beginning of "gene-flow epochs"; 4) at least three gene-flow epochs were simulated ("pre-
731 LGM", "LGM", and "post-LGM"); 5) all the demes are sampled at three time points ("slices" of the
732 landscape) in the second half of the "post-LGM" epoch. *qpAdm* experiments (rotating and non-
733 rotating, distal and proximal) were constructed by subsampling this initial set of deme samples
734 randomly or systematically, as illustrated by the examples on the left and the right, respectively. Each

735 *qpAdm* model (two- to six-way) was then characterized by 9 numbers: 3 of those were derived from
736 *qpAdm* outputs, and 6 were properties of the model in the context of the landscape. The former
737 metrics are as follows: max|*EAF*–*EF*|, max. SE, and *p*-value. Five of the latter numbers are termed
738 "model optimality metrics": min. STS angle, max. ST distance, average ST distance, Euclidean distance
739 to the ideal symmetric model (marked by the orange circle) in the space of model optimality metrics
740 (illustrated in the lower left corner of the figure) based on min. STS angle and max. or average ST
741 distance. Yet another metric is defined on the landscape but is not included in the optimality metrics,
742 that is average distance from demes in the "left" set to demes in the "right" set (average "left-right"
743 distance).

744

745 In the case of *f*-statistics, simulating $n \times m$ demes and sampling *m* of them evenly is equivalent
746 to simulating *m* demes, but with proportionally lower gene-flow intensities. Since the size of
747 the landscapes (64 demes) was the same across the gene-flow intensity brackets, these
748 landscape types, judging by the F_{ST} distributions (**Fig. S14a**), can be thought of as representing
749 regions of different size: from very sparse global sampling (10^{-5} to 10^{-4}, median F_{ST} = 0.41,
750 which is close to the highest F_{ST} between African and non-African humans) to dense sampling
751 in a sub-continental region (10^{-3} to 10^{-2}, median F_{ST} = 0.02, corresponding, e.g., to the Bronze
752 Age West Asian populations analyzed by Lazaridis *et al.* 2016). Thus, by varying gene-flow
753 intensity (**Fig. S14b,c**) we explored different levels of "'landscape sparsity" (**Box 2**): a
754 landscape property that, as we show below, is probably the single most important predictor
755 of *qpAdm* performance on SSL in the absence of rapid long-distance migrations (direct gene-
756 flows between non-neighboring demes). The "10^{-5} to 10^{-2}" type of landscapes, enriched for
757 low-intensity gene flows due to sampling from a normal distribution, but also including high-
758 intensity flows (**Fig. S14b,c**), is probably closer to real-world *qpAdm* protocols (**Suppl. text 1**)
759 where worldwide "right" groups are routinely combined with groups from a small region of
760 interest where intensive admixture is hypothesized. Median F_{ST} for this landscape type is 0.06
761 (**Fig. S14a**), which is typical for the commonly studied human groups from the Stone- and
762 Bronze Age Europe (F_{ST} between 0.05 and 0.09, Speidel *et al.* 2025).

763 To keep the simulations computationally tractable, the genome size was restricted to 500
764 Mbp with mutation and recombination rates that are realistic for humans (see Methods). See
765 **Fig. S14a** and **Fig. S15** for F_{ST} distributions and counts of polymorphic sites in these
766 simulations, respectively. In the following sections we explore effects on performance of
767 *qpAdm* protocols of the following variables: 1) landscape sparsity; 2) randomized vs.
768 systematic sampling of demes for *qpAdm* experiments (see examples in **Fig. 4**); 3) the four

769 basic types of high-throughput *qpAdm* admixture screens described above (distal and
770 proximal, rotating and non-rotating, summarized in **Fig. S16**); 4) and a wide range of
771 composite *qpAdm* model feasibility criteria (see Methods).

772

773 ***5. SSL histories: randomized sampling of landscapes***

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

786 As stated above, we explored four *qpAdm* protocols (**Fig. S16**). For example, in the case of
787 the distal rotating protocol, for a target deme randomly picked from the "present" sampling
788 point, we sampled 13 or 18 demes randomly from the two older sampling points combined
789 (100 and 300 generations before present), and they formed a rotating set of proxy sources
790 and "right" groups (**Fig. 4**). All possible one- to four-way *qpAdm* models on this subset of 13
791 or 18 demes were tested, and this collection of models for one target we termed "a *qpAdm*
792 experiment" (**Box 1**). The counts of alternative proxy sources per target (10 or 15 for the non-
793 rotating and 13 or 18 for rotating protocols) and the model complexities explored match
794 those in typical high-throughput *qpAdm* protocols from the literature (e.g., Narasimhan *et al.*
795 2019; Bergström *et al.* 2022; Lazaridis *et al.* 2022; Zeng *et al.* 2023; Speidel *et al.* 2025). In
796 general, the protocols in this study were designed to reproduce pre-study odds typical for
797 these published protocols, and we explore this issue more directly in *section 7* below. Below
798 we describe our analysis of *qpAdm* performance (based on 12,800 experiments and
799 27,511,200 individual *qpAdm* models) and propose guidelines for interpreting *qpAdm* results

800 on randomly sampled SSL, both at the level of models and experiments (the FDR results on
801 the AGS histories above are based on models, not experiments, but experiments were
802 considered in **Suppl. text 2**).

803 We characterized performance of high-throughput *qpAdm* protocols on SSL in two ways. A
804 threshold separating optimal and non-optimal models allows FDR and other terms from
805 ***equation 1*** to be calculated, but a drawback is that the estimates are threshold-dependent.
806 To avoid this artificial classification of models, we first used an indirect approach and ranked
807 models by various "optimality metrics" based on (maximal or average) spatial distances
808 between the target and sources (max. or average ST distance) and on angles formed by pairs
809 of sources and the target (minimal STS angle; **Fig. 4, Box 2**). Spatial distance is a noisy
810 approximation for genetic distance, but since 1) gene flows between non-neighboring demes
811 were not allowed, 2) gene-flow intensities on the landscapes (in the last epoch including our
812 sampling points) varied in relatively narrow ranges for all the SSL types except for "10^{-5} to 10^{-}
813 ^{2}" (**Fig. S14c, Table S6**), and 3) our main analyses were randomized and replicated (gene-flow
814 intensities were re-sampled randomly for each simulation replicate and sets of demes for
815 *qpAdm* experiments were drawn randomly), relationships between genetic (F_{ST}) and spatial
816 distances are strong, approximately linear (Pearson's *r* from 0.65 to 0.76), and remain
817 identical across *qpAdm* protocols on the same SSL type and very similar across the SSL types,
818 apart from "10^{-5} to 10^{-4}" (**Fig. S14d,e**). Thus, differences in *qpAdm* performance which we are
819 most interested in, namely those across *qpAdm* protocols, gene-flow intensity ranges, and
820 model complexities, depend on factors other than the shape of the genetic–spatial distance
821 relationship. Moreover, genetic distance measures such as F_{ST} and haplotype sharing
822 statistics on SSL depend on gene flows in both directions, but *qpAdm* results are traditionally
823 interpreted in terms of source-to-target gene flows only.

824

825 **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.
average “left-right” distance	Average spatial distance between demes in the “left” and “right” sets in a *qpAdm* model. Smaller distance increases the probability of encountering “proxy-to-right” and “right-to-proxy” gene flows bypassing the target and of “target-to-right” gene flows bypassing the proxy sources or, more formally, of observing genetic drift shared exclusively by the proxy sources and “right” populations (but not by the target) or shared exclusively by the target and “right” populations (but not by the proxy sources). These violations of the *qpAdm* assumptions (**Fig. 1**) make *p*-values smaller on average.
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 geometry of the underlying hexagonal 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 direct 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.
optimal and non-optimal models	For calculating pre-study odds, FPR, FNR, and FDR in the SSL context, we must set a threshold for classifying models into optimal and non-optimal ones. Such a threshold was based on distance to the ideal symmetric model in the space of max. ST distances and min. STS angles and was set at ≤ 5 (it equals the

	median value of this metric in the whole dataset divided by 2 and rounded). Obviously, other thresholds can be used for illustrating the same patterns in the data. Alternatively, we considered models including only the nearest neighbors and/or direct ancestors of a target as optimal and the other models as non-optimal.
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".

5a. qpAdm performance at the level of models: model optimality metrics in spaces of estimated admixture fractions and p-values

Our exploration of *qpAdm* FDR 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; Speidel *et al.* 2025). 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; Lazaridis *et al.* 2022; Brielle *et al.* 2023; Carlhoff *et al.* 2023; Speidel *et al.* 2025). Another part of the composite feasibility criteria, namely conditions on estimated admixture fractions (EAF), are also used very often (see Introduction for a brief overview of the literature). With these considerations in mind, we first inspect results (all *qpAdm* protocols pooled) for the four landscape types in the coordinates of max|*EAF*–*EF*| and *p*-values (**Figs. 5, S17**). On these two-dimensional spaces we mapped median values of the "optimality metrics" characterizing *qpAdm* models: maximal (**Fig. 5**) or average (**Fig. S17**) ST distance in a model and minimal STS angle (**Figs. 5, S17**).

It is immediately obvious that global genetic landscapes sampled very sparsely (approximated by SSL with gene-flow intensities per generation from ca. 10^{-5} to 10^{-4}) are unsuitable for *qpAdm* analyses since the bins with low *p*-values and EAF between 0 and 1 are not enriched

848 in optimal models (those with small max. ST distances and/or symmetrically arranged
849 sources) (**Fig. 5**) and absolute Spearman's rank correlation coefficient (ρ) for max|*EAF–EF*| or
850 *p*-value vs. any optimality metric does not exceed 0.1 (if max|*EAF–EF*| ≤0.5 and *p*-value ≥10^{-}
851 ^{5}; see **Table S7**). Although this high level of sparsity is probably rare in practice, it may
852 approach the current sampling situation for Paleolithic humans (Posth *et al.* 2023; Zeng *et al.*
853 2023; Allentoft *et al.* 2024) and for many species other than humans. In contrast, for the other
854 landscape sparsity levels we simulated there are strong correlations between max|*EAF–EF*|
855 and the model optimality metrics, for some model complexities at least (**Fig. 5, Table S7**; see
856 also correlation coefficients stratified by *qpAdm* protocols in **Table S8**). For instance, in the
857 case of two-way models there is a sharp boundary between models with EAF between 0 and
858 1 and those outside of this interval (the latter corresponding to max|*EAF–EF*| >0.5): the area
859 to the left of this boundary is enriched in symmetric models, while the area to the right of it
860 is enriched in asymmetric models, and this is true for all three denser SSLs (**Fig. 5**). This
861 corresponds to significant Spearman's ρ between –0.65 and –0.81 for max|*EAF–EF*| vs. min.
862 STS angle for two-way models on the "10^{-4} to 10^{-3}", "10^{-5} to 10^{-2}", and "10^{-3} to 10^{-2}" landscape
863 types (*p*-values ≥10^{-5} or ≥10^{-55}; **Table S7**).

864 This result shows that the method usually fails to report EAF in the interval (0, 1) for models
865 with asymmetrically arranged proxy sources. There is a similar, albeit more complex,
866 relationship of max|*EAF–EF*| with maximal (**Fig. 5**) or average (**Fig. S17**) ST distances. These
867 relationships of max|*EAF–EF*| with the angle- and distance-based optimality metrics
868 gradually disappear with growing model complexity and/or landscape sparsity (**Figs. 5, S17**)
869 as illustrated by Spearman's ρ (**Table S7**). Thus, a limit to model complexity that makes sense
870 to explore with high-throughput *qpAdm* on SSL is rather low in most of our simulations: just
871 two or three sources (see more on this topic below in the paragraphs focused on FDR).

872 In the dimension of *p*-values, there also exists a relationship with the model optimality
873 metrics, but it is much weaker (**Figs. 5, S17**). For example, in the practically relevant range of
874 *p*-values (10^{-5} to 1) and for max|*EAF–EF*| ≤0.5, *p*-value is not correlated with max. ST distance
875 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}"
876 landscapes (**Fig. 5**). There is weak correlation with average ST distance in the same part of the
877 space (significant Spearman's ρ up to –0.27, **Table S7**). This relationship also decays with
878 growing model complexity and landscape sparsity (**Table S7**).

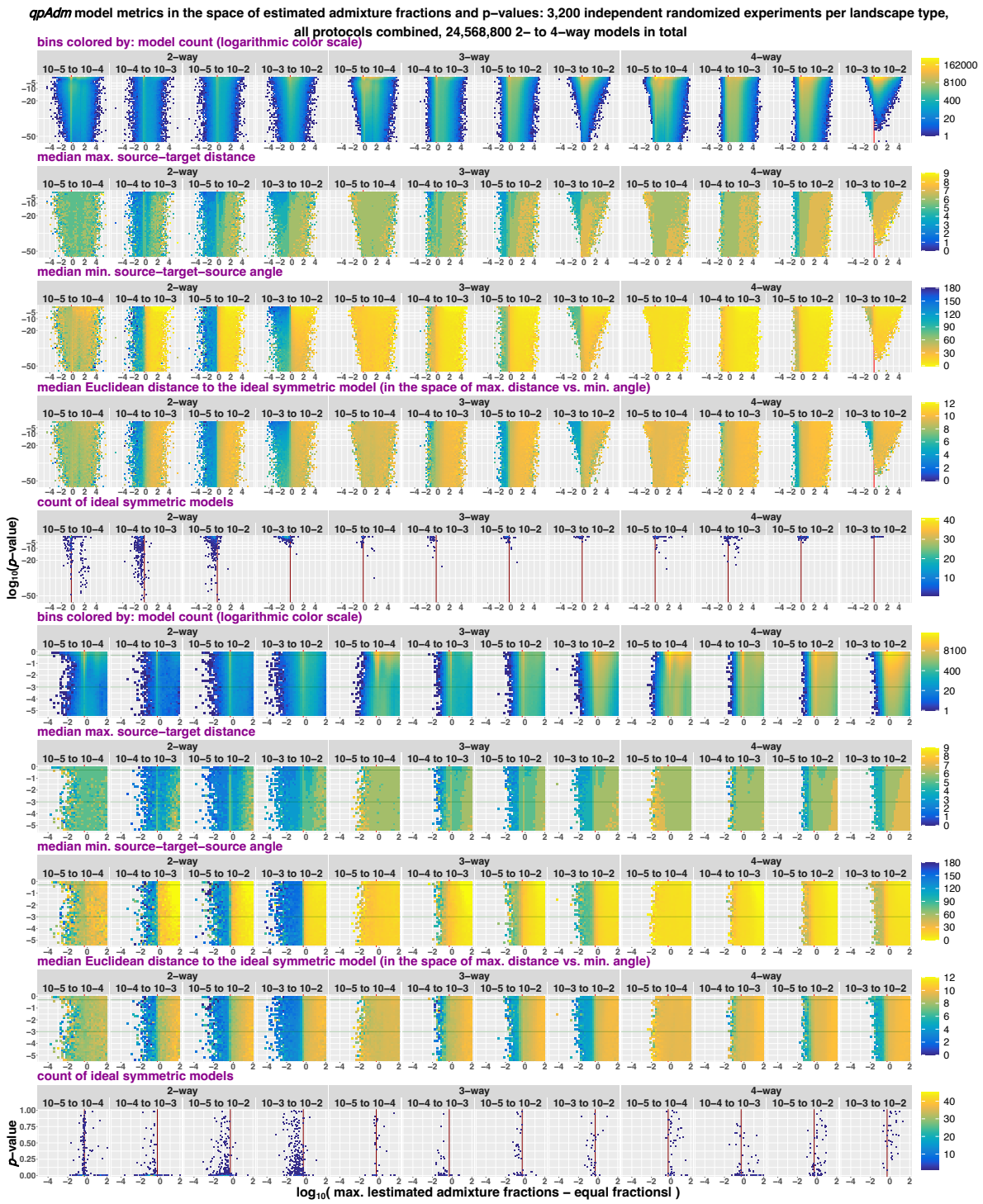


Figure 5. Randomized two to four-way *qpAdm* models in the space of max|*EAF*–*EF*| and *p*-values (both axes are logarithmic). Results for all *qpAdm* protocols are combined but stratified by model complexity and landscape type. Two sections of the space are shown: *p*-values from 10^{-55} to 1 (the 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 a 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).

We observe that our randomized *qpAdm* protocols favor 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. S18a**). 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. 4, 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; **Figs. 5, S17, Table S7**). While conditions on EAF 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 approach uniformity is the "10^{-3} to 10^{-2}" landscapes and three- and four-way models (**Figs. 5, S17, Table S9**).

Thus, the distributions of ideal randomized models on the *p*-value axis become uniform with decreasing landscape sparsity and increasing model complexity, which is an expected behavior since none of the randomized models we tested are truly ideal considering the topology of the landscape with an immediate ancestor and up to six neighboring demes for each target deme (**Fig. 4**, see however an exploration of six-way models in the systematic SSL sampling context below). Another reason for the non-uniformity of the *p*-value distributions are complex networks of gene flows that permeate the landscape and lead to low *p*-values of *qpAdm* models due to violations of the assumptions on the "left" – "right" gene flows (**Fig. 1**) that were derived originally on AGS histories (Haak *et al.* 2015; Harney *et al.* 2021), but not on SSL. In summary, in any *qpAdm* screening relying on stringent *p*-value thresholds such as

921 0.01 there probably exists a large class of two- to four-way models that have non-negative
922 admixture proportion estimates and are relatively close to optimal ones (**Fig. 5**) but are
923 rejected according to *p*-values. We explore performance of *qpAdm* protocols based on
924 composite feasibility criteria in the following sections.

925 We also analyze spaces of EAF standard errors reported by *qpAdm* vs. *p*-values (**Figs. S19,**
926 **S20, Tables S7, S8**) and, relying on similar approaches, explore randomized one-way models
927 (**Fig. S21**, **Table S10**) in **Suppl. text 5**.

928

929 *5b. qpAdm performance at the level of models: feasible models in spaces of model optimality*
930 *metrics*

931 In the previous section we undertook a very general overview of *qpAdm* results on SSL and
932 reached some preliminary conclusions on performance of high-throughput *qpAdm* protocols
933 used in practice for finding optimal models for target groups or individuals. Here we apply
934 the thresholds on *p*-values and EAF explicitly and explore a wide range of composite feasibility
935 criteria, looking at distributions of models passing these criteria in two-dimensional spaces
936 formed by the optimality metrics (**Fig. 4**). As explained in **Box 1**, composite feasibility criteria
937 may include not only conditions on EAF and *p*-value thresholds but also a condition that all
938 "trailing" simpler models are rejected according to *p*-values ("trailing" models should not be
939 confused with all simpler models tested for a given target in an experiment, see **Box 1**). As
940 another condition (rarely used in practice, to our knowledge), we tested a *p*-value threshold
941 for rejection of "trailing" models that differs from the "main" *p*-value threshold. In total, 36
942 composite feasibility criteria were tested, with *p*-value thresholds ranging from 0.001 to 0.5,
943 which covers the whole range of thresholds used in the literature explicitly or implicitly, and
944 with three types of conditions on EAF (see **Table S11** and Methods for a complete list). Results
945 for seven selected criteria are presented in the figures cited below. Since this analysis was
946 performed on combinations of source demes drawn randomly from SSL, the spaces of
947 optimality metrics are highly unevenly populated by models (**Fig. 6**; see the spaces based on
948 average ST distance in **Fig. S22**).

949 First, we explore the positive rate of *qpAdm* (**Box 1**), dividing the spaces of optimality metrics
950 into two-dimensional bins: the positive rate is simply a fraction of all models tested (within a

951 bin) passing a feasibility criterion (**Fig. S18a,b,d, S22a,c,e**). Under stringent criteria (such as
952 EAF $\pm$ 2 SE in the interval (0, 1), *p*-value $\geq$0.05, "trailing" models rejected), positive rates
953 approach 0 in most parts of the optimality metric space except for its upper left corner where
954 ideal symmetric models lie (**Fig. 4**). This is true for most levels of landscape sparsity and for
955 most model complexities we tested (**Figs. S18a, S22a**). These results are based on all the
956 *qpAdm* protocols combined, and we show results for rotating and non-rotating protocols
957 separately in **Figs. S18b,d** and **S22c,e**. For assessing practical significance of *qpAdm*'s
958 "preference" for nearly ideal symmetric models it is important to look at the system not only
959 through the lens of positive rates (**Fig. S18**), but also at density of fitting models in the space
960 of optimality metrics (**Figs. 6, S18c,e, S22b,d,f**). We use Spearman's ρ based on model counts
961 in bins to compare two-dimensional distributions (in the optimality metric space) of fitting
962 models and of all models prior to *qpAdm* testing (**Table S11**). Those highly non-uniform
963 distributions of fitting models (**Fig. 6**) are "co-created" by outcomes of *qpAdm* protocols (their
964 FPR and FNR) and by randomized construction of proxy source combinations on a two-
965 dimensional grid, and we consider this approach as a conservative approximation for high-
966 throughput *qpAdm* protocols on real-world archaeogenetic sampling (see a demonstration of
967 this in *section 7*). In other words, random landscape sampling creates "sets of all possible
968 admixture models for a target" where relatively symmetric models composed of target's
969 neighbors constitute a small minority, and where highly asymmetric models with distant
970 sources are by far the most abundant (**Fig. 6**). See also alternative visualizations (**Fig. S23**)
971 combining the approaches from **Figs. 5** and **6** and illustrating the same points.

972 In the cases of four- and three-way models, only very specific landscape properties and
973 feasibility criteria led to depletion of the sets of fitting models for non-optimal models (those
974 with small min. STS angles and at least one distant source), which is evidenced by Spearman's
975 ρ (for total model counts vs. fitting model counts in two-dimensional bins) not significantly
976 different from 0 under these conditions, but high and significant under nearly all the other
977 conditions tested (**Table S11**). These conditions are as follows: the densest landscape tested
978 ("10^{-3} to 10^{-2}") and EAF $\pm$ 2 SE $\in$ (0, 1), with no significant influence of the *p*-value threshold
979 on Spearman's ρ. These results are in line with our observations in the preceding section (**Fig.
980 5**), namely that EAF (and their SE correlated with them) and landscape sparsity, in contrast to
981 model *p*-values (which are highly variable due to gene flows in all directions on the landscapes

and due to average distances between “left” “right” deme sets fluctuating across models, as discussed in *section 5c*), 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 ∈ (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.

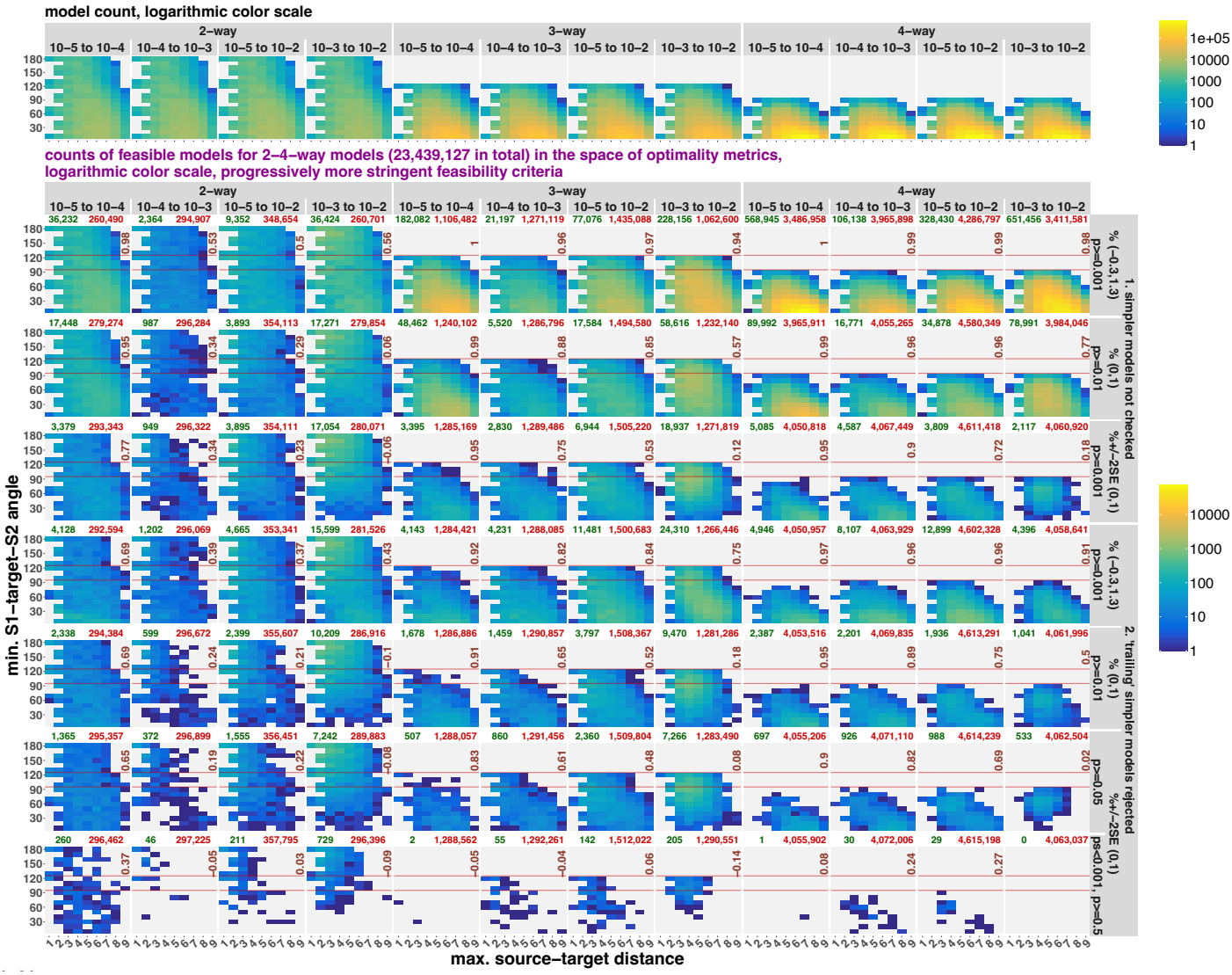


Figure 6. Randomized two to four-way *qpAdm* models in the space of model optimality metrics: max. ST distance and min. STS angle (models with the latter metric undefined were excluded from this analysis). Results for all *qpAdm* protocols are combined but stratified by model complexity and landscape 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.

Introducing very stringent feasibility criteria (such as *p*-value ≥0.5, "trailing" models rejected with p <0.001, EAF ± 2 SE ∈ (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. 6, Table S11**). As we discuss below in *section 6b*, despite *p*-values growing with increasing model complexity, fraction of models with EAF ∈ (0, 1) falls sharply 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 Eurasia in the case of humans (which corresponds approximately to the "10^{-3} to 10^{-2}" landscapes); stringent conditions on EAF such as EAF ± 2 SE ∈ (0, 1); a low *p*-value threshold for "fitting" models (such as 0.001) and no *p*-value thresholds for rejecting simpler "trailing" models. We note that, following this approach, it is almost impossible to find a single best model complexity level for a target and a single best model at that complexity level, but that is difficult to achieve in principle on SSL as we discuss in the next section.

Inspecting the spaces formed by average ST distance and min. STS angle, we observe the same trends (**Fig. S22b,d,f**). Considering the radius and shape of the SSL studied here (**Fig. 4**), the triangular region in the lower right corner of the optimality metric space (**Fig. S22b,d,f**) is populated by models where a target is located close to (or on) one edge of the landscape and at least some proxy sources are located close to the opposite edge of the landscape (average ST distance >4.5, which is the radius of the landscape). Complete elimination of those highly asymmetric models from sets of fitting models is more achievable in the case of two-way models, but is achievable for three- and especially four-way models only on the densest landscape "10^{-3} to 10^{-2}" and/or when the most stringent conditions on EAF are applied (**Fig. S22b, Table S11**).

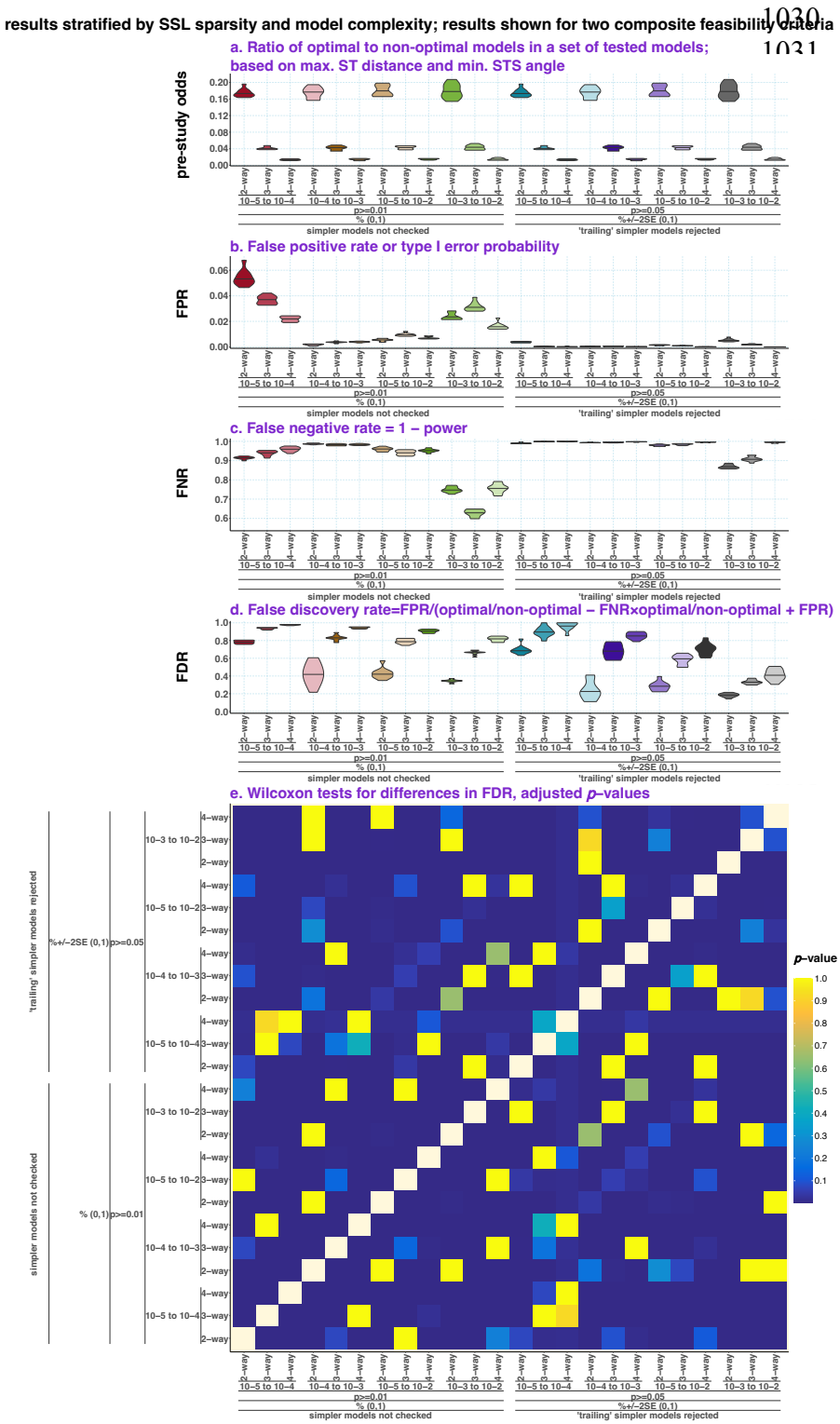


Figure 7. Pre-study odds (**a**), FPR (**b**), FNR (**c**), and FDR (**d**) are shown for two composite model feasibility criteria. The violin plots with medians visualize distributions of these metrics across simulation replicates. The results are grouped by SSL sparsity and model complexity (all *qpAdm* protocols are considered together). Results of pairwise comparisons of the FDR distributions are shown in the matrix form in panel **e** (*p*-values of non-paired Wilcoxon tests were adjusted for multiple testing using the Holm method).

Admixture models on SSL lie on an "optimality continuum", but for illustrating our conclusions above in a more quantitative way it makes sense to introduce a threshold in the space of optimality metrics (such as that in **Fig. 6**) for classifying models into optimal and non-optimal ones, which enables us to compute pre-study odds, FPR, FNR, and FDR. The distance threshold we chose (**Box 2**) corresponds to a difference in max. ST distance (vs. the ideal model) up to 5 and/or to differences in min. STS angle up to 100°, 66.7°, and 50° (for two-, three-, and four-way models, respectively). Even considering this relaxed threshold, the SSL-based simulations designed to resemble typical high-

1072 throughput *qpAdm* protocols from the literature result in low pre-study odds: ca. 0.18, 0.04,
1073 and 0.015 for two-, three-, and four-way models, respectively (**Fig. 7**). In this figure we also
1074 show FPR, FNR, and FDR distributions across simulation replicates grouped by SSL sparsity
1075 and model complexity. See more fine-grained results stratified additionally by the number of
1076 alternative sources for a target and by *qpAdm* protocols in **Figs. S24-S26** (there results are
1077 shown for all the composite feasibility criteria from **Fig. 6** except for the most stringent one).

1078 In **Fig. 7** we compare results for two composite feasibility criteria of different stringency: EAF
1079 ∈ (0, 1), *p*-value ≥0.01, simpler models not checked; and EAF ± 2 SE ∈ (0, 1), *p*-value ≥0.05,
1080 "trailing" simpler models rejected. As expected, increasing the stringency changes FPR from
1081 0.2%–5.4% to 0.01%–0.5% (these are FPR values from **Fig. 7**, averaged across simulation
1082 replicates), and these differences are statistically significant. There is the opposite trend in
1083 FNR: it grows from 62.9%–98.7% to 86.9%–99.9% with increasing feasibility criteria
1084 stringency (**Fig. 7**). Due to low pre-study odds and very high FNR, the improvement in FDR
1085 with increasing feasibility criteria stringency is much less striking than the improvement in
1086 FPR, although most differences are statistically significant (**Fig. 7**): for example, for models of
1087 varying complexity on the "10^{-3} to 10^{-2}" landscapes FDR falls from 31.5%–81.8% to 18.4%–
1088 40.9%. These results also illustrate in another way the trends discussed in this section above,
1089 namely the significant differences in *qpAdm* performance between model complexities and
1090 some SSL sparsity levels (**Fig. 7**). Since pre-study odds fall rapidly with increasing model
1091 complexity, to keep FDR for three-way and more complex models below 50%, both FPR and
1092 FNR must remain relatively low for these model classes (***equation 1***), but this is achievable
1093 only in the case of densely sampled landscapes and stringent conditions on EAF (EAF ± 2 SE
1094 ∈ (0, 1)). The power of *qpAdm* model screens (1 – FNR) grows with SSL density, especially
1095 between the "10^{-5} to 10^{-2}" and "10^{-3} to 10^{-2}" landscape types (**Fig. 7**), and FPR is improved by
1096 orders of magnitude with the introduction of more stringent conditions on EAF. Compare, for
1097 example, the following criteria: EAF ∈ (-0.3, 1.3), *p*-value ≥0.001, simpler models not checked
1098 and EAF ± 2 SE ∈ (0, 1), *p*-value ≥0.001, simpler models not checked (**Fig. S24**). Although
1099 different *qpAdm* protocols yield significantly different FPR and FNR in some cases (for some
1100 SSL types and model complexities), changes in these variables mostly cancel out, and there
1101 are no statistically significant differences in FDR between *qpAdm* protocols and different sizes
1102 of proxy source or rotating sets (10 or 15, 13 or 18 demes) (**Fig. S24a-c**).

1103 Importantly, high *p*-value thresholds often used in the literature explicitly or implicitly (≥0.05)
1104 are counter-productive in the SSL context since they do not improve FDR significantly (**Fig.**
1105 **S24**) but result in very low overall positivity rates (**Table S11**). We note that due to the
1106 widespread usage of composite model feasibility criteria in the literature and in this study
1107 and since ideal models on SSL may be much more complex than those tested in practice (see
1108 *section 6b*), *p*-value thresholds are not equivalent to FPR and, moreover, predicted positive
1109 findings interpreted in typical *qpAdm* analyses are not rejections of the null hypothesis, but
1110 non-rejections (see Introduction).

1111

1112 *5c. qpAdm performance at the level of experiments*

1113 In this section *qpAdm* results are interpreted from the perspective of experiments which
1114 progress from simple to complex models (from one-way to two-, three-, and four-way models
1115 in our study) and are stopped (produce a positive result) as soon as at least one fitting model
1116 is found for a chosen target. Experiments are defined not only by a target, but also by a set
1117 of "right" demes or a rotating set. Approaches of this sort are often used in high-throughput
1118 screens for admixture that rely on *qpAdm* (see, e.g., Narasimhan *et al.* 2019; Bergström *et al.*
1119 2022; Lazaridis *et al.* 2022; Zeng *et al.* 2023; Speidel *et al.* 2025). When interpreting *qpAdm*
1120 results at the level of experiments, we paid attention to fractions of experiments producing
1121 at least one fitting model at various model complexity levels and to fractions of those
1122 "positive" outcomes that would be misleading in practice. In short, we define such misleading
1123 outcomes in the following way (see Methods for a precise definition): models rejected in
1124 those experiments are much closer to the ideal symmetric models, as compared to models
1125 fitting in those experiments. For declaring a positive result, it was required that all simpler
1126 models for the target (all possible source combinations) are rejected according to a chosen
1127 composite feasibility criterion.

1128 For a complete overview of our results at the level of experiments see **Figs. S27** (relying on
1129 max. ST distance) and **S28** (relying on average ST distance). A key observation is that the most
1130 optimal *qpAdm* setup (the densest, "10^{-3} to 10^{-2}", landscape, EAF ± 2 SE ∈ (0, 1), *p*-value
1131 ≥0.001, "trailing" models not checked) found in the previous section indeed results in the
1132 lowest fractions of positive experiments with "bad" outcomes, as compared to other

1133 conditions tested (compare **Figs. 8** and **S27**). Here we focus on the results for all the *qpAdm*
1134 protocols under these optimal conditions, and they share a common pattern: fitting models
1135 were found for 92% of targets; 55% of experiments ended at one-way models, and 36.5% at
1136 two-way models (on average across all simulation replicates and protocols; **Fig. 8**). Nearly all
1137 experiments of both types were not classified as producing misleading results (**Fig. 8**).
1138 Nevertheless, the outcome can hardly be viewed as satisfactory since one-way and two-way
1139 models are too simple for these landscapes with high-intensity gene flows.

1140 In general, increasing SSL density from "10^{-4} – 10^{-3}" to "10^{-5} – 10^{-2}" and to "10^{-3} – 10^{-2}" leads
1141 to complex models emerging less often as outcomes of experiments (**Figs. S27, S28**). For
1142 example, experiments on the "10^{-5} to 10^{-2}" landscapes yielded three- and four-way models
1143 significantly more often than experiments on the "10^{-3} to 10^{-2}" landscapes discussed above.
1144 The same effect has increased stringency of conditions on EAF. Raising the *p*-value threshold,
1145 on the contrary, increases the fraction of complex models as experiment outcomes, but at
1146 the cost of significantly higher rates of positive experiments with potentially misleading
1147 outcomes (**Figs. 8, S27, S28**). Introducing the stringent condition on EAF (EAF $\pm$ 2 SE $\in$ (0, 1))
1148 allowed low rates of "bad" positive outcomes (up to 10-20%) to be achieved even for high *p*-
1149 value thresholds (**Fig. S27c**).

1150 Another major trend is that a substantial fraction of non-rotating experiments did not end
1151 even at the stage of four-way models (from ca. 90% to 30%, depending mostly on the *p*-value
1152 threshold), in contrast to rotating experiments, and the distributions of experiments by
1153 model complexity at which they end are even more skewed towards simpler models in this
1154 case (this is effect is especially pronounced in the case of the 10^{-5} to 10^{-2} and sparser
1155 landscapes, see **Figs. S27, S28**). A fundamental difference between these two protocol types
1156 underlies this observation: a key point is that small average distance between "left" and
1157 "right" deme sets increases the probability of encountering violations of the *qpAdm*
1158 assumptions (e.g., "left-to-right" gene flows) that lower *p*-values and thus increase model
1159 rejection frequency (**Fig. 1** and **Box 2**). We explored this issue directly (**Fig. S29**) in the
1160 systematic SSL sampling context (see *section 6a*): for two to ten target demes in the center
1161 of the "10^{-5} to 10^{-2}" landscapes, we generated distal non-rotating and rotating experiments
1162 characterized by varying average "left-right" distances (proxy, rotating, and "right" demes
1163 were sampled from circles at varying distance around the targets as shown in **Fig. S16**).

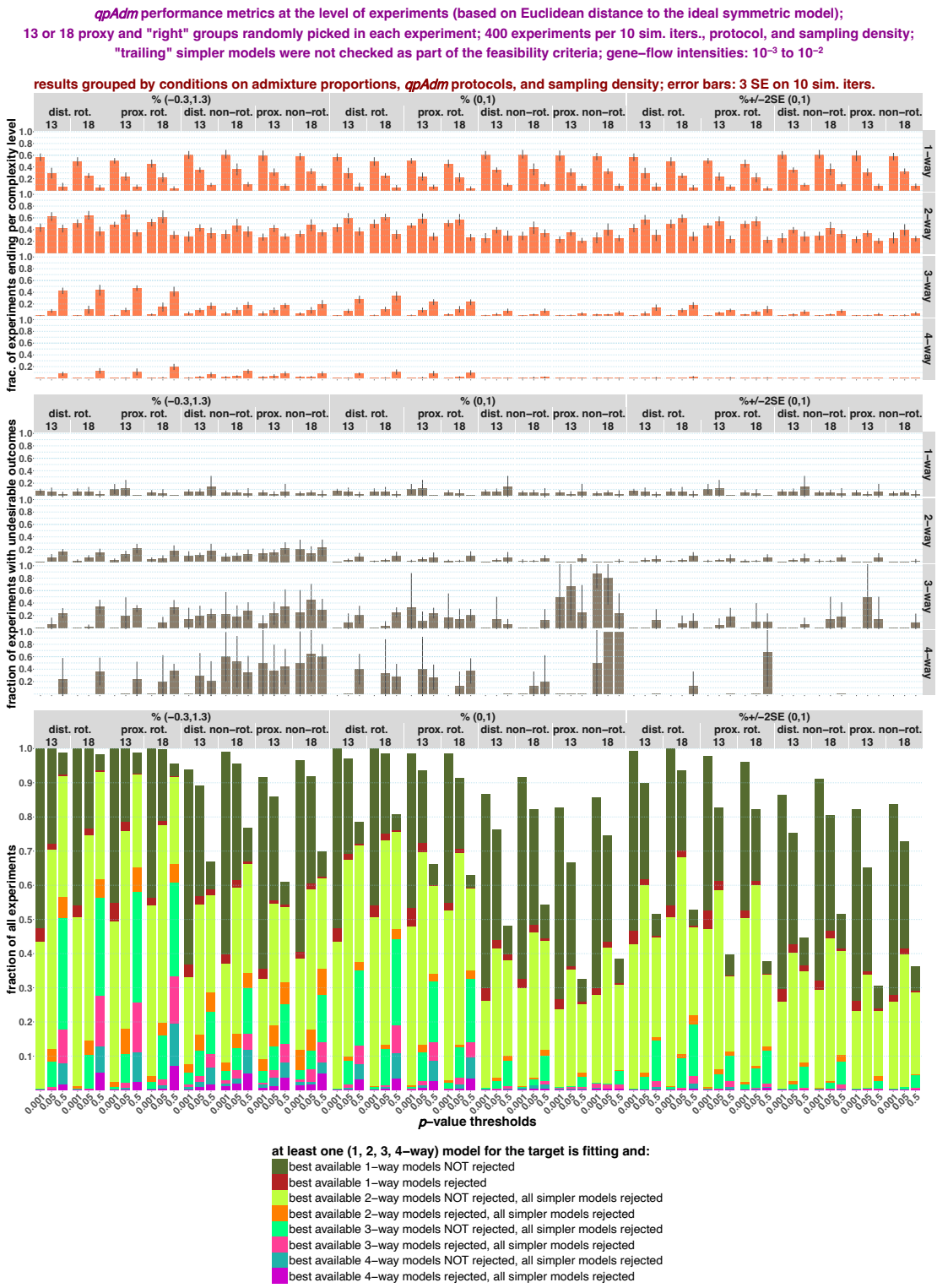


Figure 8. Randomized *qpAdm* results at the level of experiments visualized for "dense" landscapes (10^{-3} to 10^{-2}). The results are stratified by *qpAdm* protocols and landscape sampling density (13 or 18 demes in rotating sets and 10 or 15 demes in non-rotating "right" and proxy source sets, respectively), 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 (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 STS 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) are visualized by the stacked bar plots. Error bars are not shown for visual clarity in the latter plot.

In a non-rotating experiment, distances from the "right" demes to the target and to each alternative source remain fixed, but in a rotating experiment many bisections of the rotating set are tested; consequently, a wider range of average "left-right" distances is sampled (**Fig. S29d**), and such an experiment may end at a model with a large average "left-right" distance since it has a relatively low rejection probability. As we consider experiment outcomes on many randomly generated "left" and "right" sets, some non-rotating experiments end quickly because of large average "left-right" distance in most optimal source configurations available in those experiments, others end at much more complex models because of small average "left-right" distance (in most optimal models), and rotating experiments end somewhere in between since "left-right" distance is optimized within each experiment. For an illustration of this argument see contrasting *p*-value distributions for non-rotating and rotating systematic two- and three-way models matched by the number of target's nearest neighbors included (**Fig. S29d**). At the level of systematic experiments, we observe that there are "left-right" distances which cause nearly all experiments (of 20 performed in each case) to progress beyond four-way models, and the experiment ending rate grows with average "left-right" distance (**Fig. S29a**). For example, even some "inverted" proxy and "right" deme arrangements (the latter closer to the target than the former) demonstrate relatively high experiment ending rates, such as all proxies at a distance of 4 and "right" demes at distances of 1 and 2 (**Fig. S29a**). Indeed, *p*-value distributions are skewed towards values <<0.01 for small average "left-right" distances and become more uniform as "left-right" distance grows (**Fig. S29c,d**).

A general conclusion from this section is that it is very difficult to find setups for randomized *qpAdm* experiments in the SSL context 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

1206 outcomes of two experiments (**Fig. S30**) relying on one of the most optimal composite
1207 feasibility criteria (EAF $\pm$ 2 SE $\in$ (0, 1), *p*-value $\geq$0.001, "trailing" models not checked), where
1208 only one fitting two-way model was found in each (the "10^{-5} to 10^{-2}" landscapes, distal
1209 rotating protocol, rotating sets composed of 18 demes sampled 100 and/or 300 generations
1210 ago). Attempting to get more complex models (three-way) resulted in non-optimal models
1211 being accepted and much more optimal models being rejected (**Fig. S30**). Considering patchy
1212 sampling simulated here (**Fig. S30**), this situation may be wrongly interpreted as a signal of a
1213 long-distance migration, especially in a high-throughput setting when each experiment is not
1214 carefully assessed.

1215

1216 ***6. SSL histories: systematic sampling of landscapes***

1217 We aimed to explore how the availability of close ancestry sources (the nearest neighbors of
1218 a target deme, also termed "1st circle") influences outcomes of *qpAdm* experiments on the
1219 most realistic simulated landscapes ("10^{-5} to 10^{-2}", see F_{ST} distributions in **Fig. S14a**). In
1220 contrast to the randomized deme sets tested above, we used "right", proxy, and rotating sets
1221 composed of demes equidistant from the target and distributed uniformly on the landscape,
1222 forming circles around the target (**Figs. 4, S16**). Two kinds of *qpAdm* protocols were tested:
1223 distal rotating and distal non-rotating (**Fig. S16**).

1224

1225 *6a. Systematic sampling: model optimality metrics in the spaces of estimated admixture*
1226 *fractions and p-values*

1227 We first explored rotating sets composed of two deme circles around the target: at distances
1228 of 3 and 1 (for a description of the corresponding non-rotating setup see **Fig. S16**). We again
1229 inspect median values of the model optimality metrics in the spaces of max|*EAF–EF*| and *p*-
1230 values (**Figs. S31, S32**) and show the corresponding Spearman's ρ for the optimality metrics
1231 vs. max|*EAF–EF*| or *p*-values in **Table S7**, and correlation coefficients stratified by *qpAdm*
1232 protocols in **Table S8**. In these tables, the results relying on randomized and systematic deme
1233 sampling are juxtaposed. In **Figs. S31** and **S32** (based on max. and average ST distance,
1234 respectively), we also juxtapose results relying on randomized and systematic sampling of the
1235 "10^{-5} to 10^{-2}" landscapes. These two sampling approaches demonstrate similar results (**Fig.**

1236 **S32**), however there are generally stronger correlations between max|*EAF–EF*| or *p*-values
1237 and the optimality metrics in the case of systematic sampling (**Table S7**). These results suggest
1238 that a particular arrangement of "right" groups (evenly distributed on a circle around the
1239 target) improves performance of the method (that is, correlation between *p*-values and
1240 model optimality metrics), however the improvement is not striking. This arrangement of
1241 "right" demes also leads to rejection of a larger fraction of ideal two- or three-way models as
1242 compared to that under randomized deme sampling (**Table S9**). We note that for nearly all
1243 ideal two- and three-way models under systematic sampling, EAF ∈ (0, 1) (**Table S9**).

1244 Next, we broadened the sets of proxy sources tested to encompass all possible one- to four-
1245 deme sets at a particular time (this is feasible in the non-rotating context only) and inspected
1246 the influence of target-"right" distance on *qpAdm* performance: minimal target-"right"
1247 distance varied from 1 to 4 for two suitable target demes at the center of the landscapes,
1248 while keeping the size of the "right" set constant at 16 demes. We present the results in the
1249 usual way: in the spaces of max|*EAF–EF*| and *p*-values (**Figs. S33, S34**) and as Spearman's
1250 correlation coefficients between the model optimality metrics and max|*EAF–EF*| or *p*-values
1251 in various, more or less practically relevant, parts of these spaces (**Table S12**). These spaces
1252 are in general similar to those based on randomized *qpAdm* tests explored above (**Figs. 5** and
1253 **S17**), with simpler models (two-way) demonstrating better performance than three- or four-
1254 way models. The same difference in *qpAdm* performance is seen when models composed
1255 exclusively of target's nearest neighbors are considered true, and all the other models are
1256 considered false: pre-study odds fall with increasing model complexity, and FDR grows mainly
1257 for this reason (**Table S13**). The best performance of the method (strongest correlation
1258 between the model optimality metrics and max|*EAF–EF*| or *p*-values) is observed when
1259 "right" demes are in the 3rd or 4th circles around the target (see, for example, the correlation
1260 between *p*-values and average ST distances in **Table S12**). The distributions of ideal symmetric
1261 two- to four-way models also become less uniform on the *p*-value axis with decreasing target-
1262 "right" distance (**Figs. S33, S34**) since "right" demes move closer to "left" demes adjacent to
1263 the target, which increases the probability of "left-to-right" gene flows bypassing the target
1264 and causes *p*-values to drop on average (see the discussion in the preceding *section 5c*).

1265

1266

1267 *6b. Systematic sampling: exploring truly ideal six-way models*

1268 Due to computational limitations, in all our randomized and systematic *qpAdm* experiments
1269 described above the most complex models tested were four-way. However, on the SSL
1270 simulated in this study (**Fig. 4**), truly ideal models for most demes are six-way. Here we do
1271 not consider cases when a direct ancestor is available since this makes modelling population
1272 history trivial in our systematic setups based on the "10^{-5} to 10^{-2}" landscapes: instead of
1273 seven-way models, one-way models fit the data in nearly all cases. Here we explore all
1274 possible one- to four-way and six-way models for our standard systematic setup: distal
1275 rotating and distal non-rotating protocols based on demes from the 1st and 3rd circles around
1276 the target (all demes with complete 3rd circles were included as targets). More precisely,
1277 demes from the 1st and 3rd circles comprised 16-deme rotating or proxy sets (**Fig. S16**). Results
1278 based on this setup are also presented in **Figs. S31** and **S32** and **Tables S7-S9**, but here we
1279 show results in ideal conditions only, where all six demes from the 1st circle were available in
1280 *qpAdm* experiments (not to be confused with individual models). We first look at the results
1281 in the max|*EAF–EF*| vs. *p*-value spaces (**Figs. 9a** and **S35a,b**).

1282 We noted above that *p*-value is not an optimal criterion for ranking randomized or systematic
1283 *qpAdm* models on the "10^{-5} to 10^{-2}" landscapes (and most other landscape types tested) and
1284 that max|*EAF–EF*| correlates better with the model optimality metrics (**Fig. 5, Tables S7, S9**).
1285 For example, large fractions of "ideal" symmetric three-way models are rejected according
1286 to *p*-value thresholds such as 0.01 or 0.05 (**Table S9**), however this happens precisely because
1287 any three-way model is not truly ideal on our hexagonal grid and therefore *p*-value
1288 distributions are not uniform for this model complexity (**Figs. S33, S34**). In contrast, six-way
1289 models composed of target's nearest neighbors demonstrate uniform distributions of *p*-
1290 values (**Fig. 9a**), as expected for ideal models (Harney *et al.* 2021). Non-optimal sources in this
1291 setup come from the 3rd circle, the same circle all the "right" groups (200 generations deeper
1292 in time) in the non-rotating setup and most of the rotating set come from (**Fig. S16**). Thus,
1293 removing all demes from the 3rd circle from a *qpAdm* model eliminates detectable "left-to -
1294 right" gene flows and results in a high *p*-value.

1295

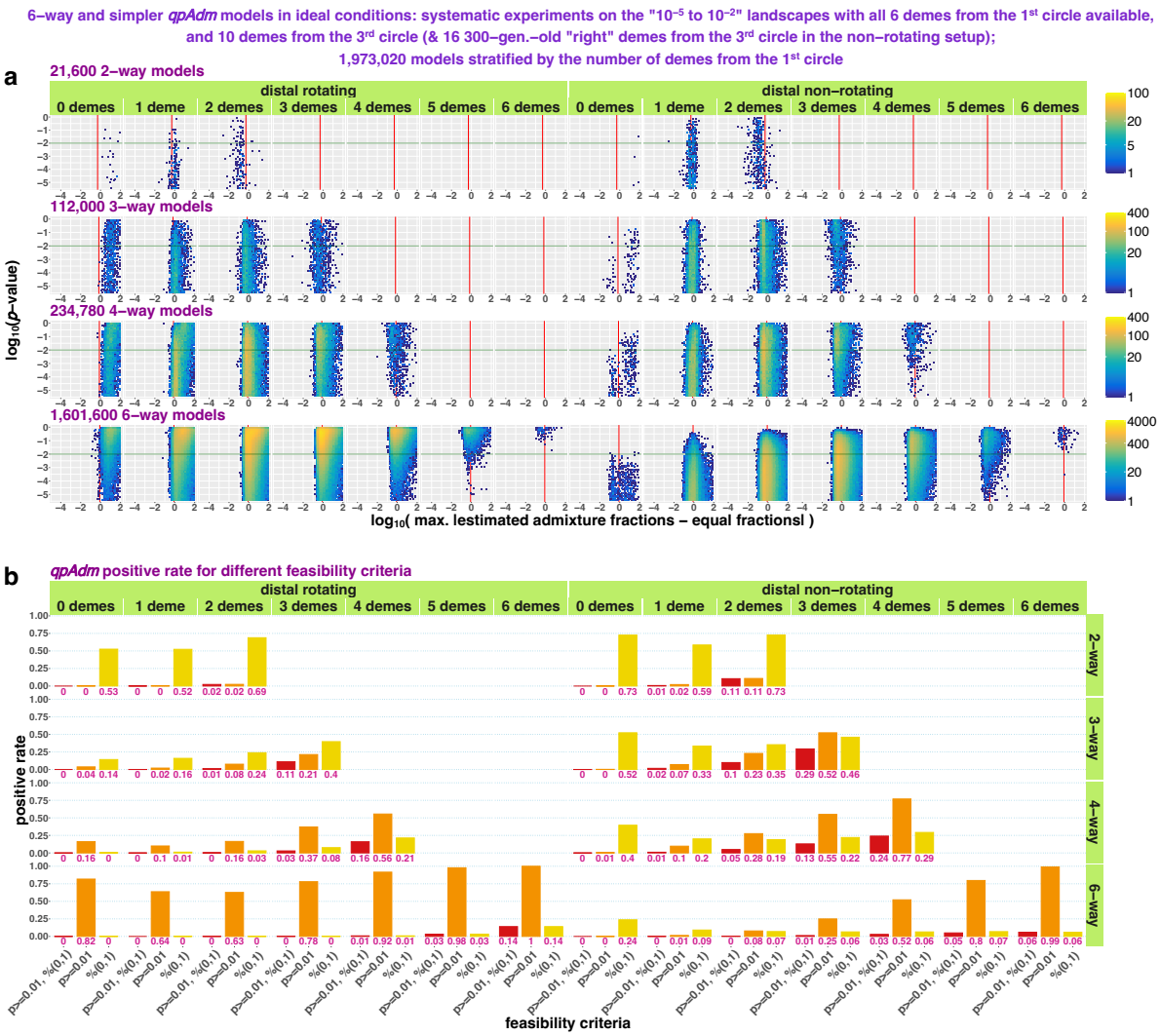


Figure 9. Systematic two to four- and six-way *qpAdm* models on the "10^{-5} to 10^{-2}" landscapes in the space of max|*EAF*–*EF*| and *p*-values (**a**). Results are stratified by model complexity, *qpAdm* protocol, and number of target's nearest neighbors in a model, which is the only optimality metric 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. S35a**. 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 fractions of tested models that satisfy three kinds of feasibility criteria marked on the x-axis. The fractions are also indicated below the bars.

To look at feasibility criteria systematically, we show *qpAdm* positive rates for models stratified by complexity and optimality, which is defined here as the number of demes from the 1st circle included in a model (**Figs. 9b, S35c**). 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. The systematic distal

1314 non-rotating protocol with the second feasibility criterion (the *p*-value threshold)
1315 demonstrates attractive properties in the case of three-, four-, and six-way models: there is
1316 strong correlation between the fraction of models passing the *p*-value threshold and model
1317 optimality, and the percentage of truly ideal models rejected equals the *p*-value threshold, as
1318 expected (**Fig. 9b**). In contrast, such a "clean" result was not achieved with the distal rotating
1319 protocol: the fraction of models passing the *p*-value threshold grows with increasing model
1320 complexity, but at the same time its correlation with model optimality disappears (**Fig. 9b**).
1321 The fraction of models satisfying the EAF condition demonstrates the opposite pattern: it is
1322 correlated with model optimality when the rotating protocol is used and uncorrelated if the
1323 non-rotating protocol is used (**Fig. 9b**). But 86% or 94% of ideal six-way models were rejected
1324 according to the condition EAF ∈ (0, 1) in the rotating and non-rotating setups, respectively
1325 (**Fig. 9b**), demonstrating that AF estimates become highly uncertain for complex models.
1326 Thus, in this particular systematic setup (**Fig. S16**), the non-rotating protocol with model
1327 feasibility criteria based solely on *p*-values is performing substantially better than the rotating
1328 one with model feasibility criteria based on EAF and/or *p*-values, but it is hard to say if this
1329 observation can be generalized to other regions of the huge parameter space we try to
1330 explore.

1331 It is possible look at the same system through the lens of FDR. For simplicity, admixture
1332 models composed exclusively of target's nearest neighbors were considered true and all the
1333 other models were considered false. Although (if feasibility conditions on both EAF and *p*-
1334 values are used) FPR is low (0% to 3.7%) and independent of model complexity (**Table S13**),
1335 FDR for two- and three-way are drastically different: 0% and 40%, and FDR reaches 98.4% in
1336 the case of six-way models (the distal rotating protocol; model feasibility criteria: *p*-value ≥
1337 0.1 and EAF ∈ (0, 1); see **Table S13**). These results are due to pre-study odds falling rapidly
1338 with increasing model complexity: from 0.143 for two-way, 0.037 for three-way, and $1.2\times10^{-}$
1339 ^{4} for six-way models in this setup (**Table S13**). FDR is very strongly (negatively) correlated with
1340 pre-study odds: Pearson's *r* ranges from –0.86 to –0.99 across various model feasibility
1341 criteria we explored (**Table S13**). More relaxed definitions of true models can be used, but
1342 the problem remains the same in principle. We also note that, when viewed from the FDR
1343 perspective, the composite feasibility criteria are superior to *p*-value thresholds and the distal
1344 rotating protocol is superior to the non-rotating one (**Table S13**). These conclusions differ

1345 from those based on **Fig. 9** since relative sizes of the model optimality classes were not
1346 considered in the latter interpretation, but they do influence FDR of high-throughput *qpAdm*
1347 protocols where all possible combinations of chosen sources are tested.

1348

1349 *6c. Systematic sampling: qpAdm performance at the level of experiments*

1350 In this section we discuss problems with interpretation of systematic *qpAdm* experiments'
1351 outcomes. For both the rotating and non-rotating protocols based on demes from the 1st and
1352 3rd circles around the target (**Fig. S16**), we observe the following trend: model complexity at
1353 which an experiment ends (the first feasible model is found) decreases with the number of
1354 close ancestry sources (demes from the 1st circle) available in that experiment (**Fig. 10a,b**).
1355 This trend is especially striking in the case of the distal non-rotating protocol: when 0 close
1356 sources were available, although we relied on a relatively relaxed composite feasibility
1357 criterion (*p*-value threshold at 0.01, EAF between 0 and 1), nearly all experiments did not end
1358 at the stage of four-way models, suggesting that five-way or even more complex models are
1359 needed (**Fig. 10b**). If one to four close sources were available, experiments most often ended
1360 at three-way models, and if all six nearest neighbors of a target were available, a great
1361 majority of experiments ended at two-way models.

1362 We also expanded the sets of available proxy sources in the distal non-rotating protocol to all
1363 possible combinations of demes at 100 generations before "present" and varied target-
1364 "right" distance (see *section 6a*). In line with the results above, most of those experiments
1365 with the widest selection of sources available ended at two-way models, and a majority of
1366 those fitting (EAF $\in$ (0, 1); *p*-value $\geq$0.01) two-way models included just one nearest neighbor
1367 of a target (**Fig. S36**). The fraction of experiments ending at three-way models grew with
1368 decreasing target-"right" distance, and no experiment ended at four-way models (**Fig. S36**).

1369 Naively, we would expect to see an opposite trend: *qpAdm* stops at more complex models
1370 when more true ancestry sources are available in a panel of populations tested. But in our
1371 experiments *qpAdm* results can be considered misleading in all cases. When no or few close
1372 ancestry sources are available, the method tends to pick distant sources in the context of
1373 complex models. When (nearly) all closest ancestry sources are available, the method
1374 oversimplifies the history of the target by picking just two or three sources.

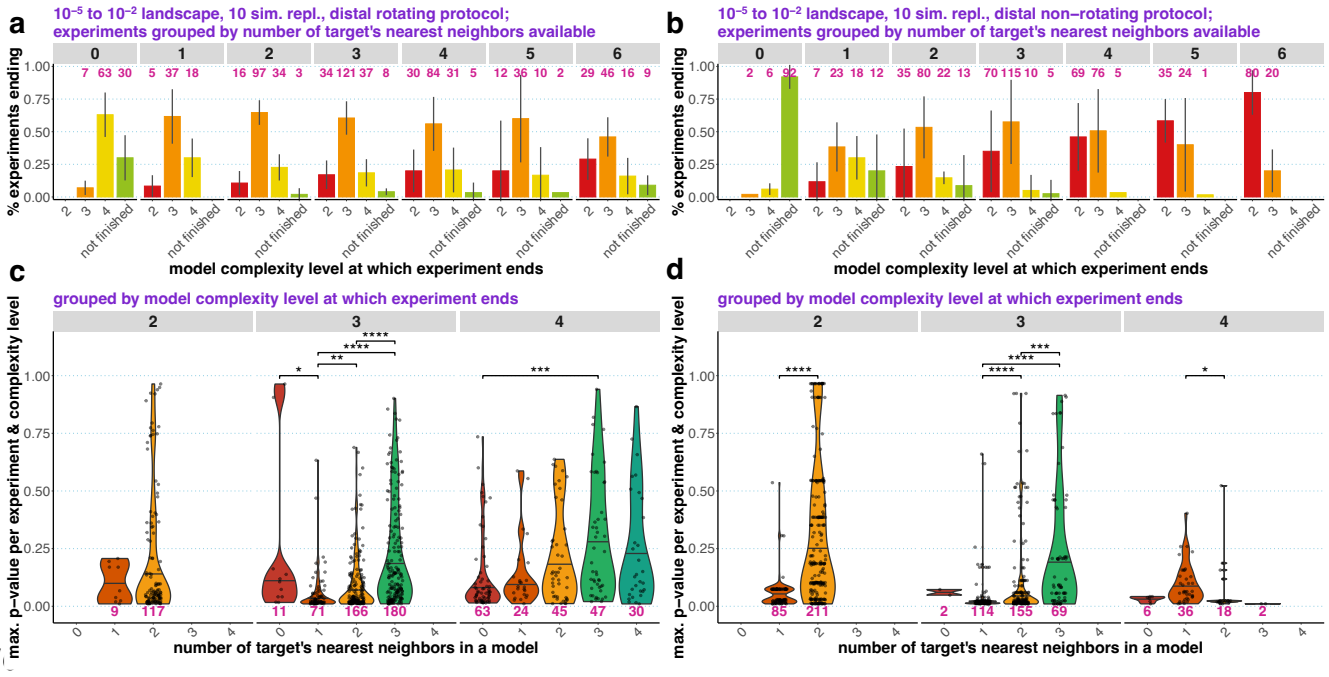


Figure 10. *qpAdm* performance in the case of systematic and symmetric landscape sampling, interpreted at the level of experiments. Results for the distal rotating (**a, c**) and distal non-rotating (**b, d**) protocols are shown. The protocols relied on the following composite feasibility criterion: *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 pairs of the distributions were compared 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.

This seemingly puzzling pattern (**Figs. 10, S36, S37**) can be explained by the relationship between average "left-right" spatial distance and *qpAdm p*-values (**Fig. S29**; see also *section 5c*). The rotating sets in the experiments discussed here were split in different ratios between the 1st and 3rd circles (**Fig. S16**). Consider, for instance, three- and four-way rotating models including no proxy sources from the 1st circle, that is, with all proxy sources coming from the 3rd circle (**Fig. S29d**). Even though both source and "right" demes are sampled from the same

1401 circle in this case, there exists a fraction of these rotating set bisections with average "left-
1402 right" distances >4 and *p*-values >0.01 (**Fig. S29d**), which corresponds to ca. 75% of
1403 experiments ending at three- and four-way models in the leftmost panel of **Fig. 10a**. In
1404 contrast, non-rotating three- and four-way models of the same type (with both proxy and
1405 "right" demes coming from the 3rd circle) demonstrate average "left-right" distances in a
1406 narrow range (3.5 to 3.75) and *p*-values mostly <<0.01 (**Fig. S29d**), and hence ca. 90% of
1407 experiments progress beyond four-way models (see the leftmost panel in **Fig. 10b**). And the
1408 opposite situation is observed at the other end of the spectrum for models composed of
1409 target's nearest neighbors only: three- and four-way non-rotating models demonstrate larger
1410 "left-right" distances on average and higher *p*-values as compared to rotating models (**Fig.**
1411 **S29d**), corresponding to the significantly different experiment ending rates in the rightmost
1412 panels in **Fig. 10a** and **10b**. When the whole landscape is sampled in an experiment (**Fig. S36**),
1413 average "left-right" distance varies widely in models from such an experiment, but only "left-
1414 right" distances in most optimal models composed of sources close to the target determine
1415 how fast experiments end. Thus, decreasing target-"right" distance also decreases "left-right"
1416 distance for optimal models, drives their *p*-values down (**Fig. 29c**), and causes experiments to
1417 end at more complex models. The patterns highlighted here are peculiar to our "circular"
1418 rotating and non-rotating setups, but they are instructive and help to understand the
1419 patterns observed on random landscape samples (see *section 5c*).

1420

1421 ***7. Geographical analysis of large sets of qpAdm models on real data***

1422 In this section we use the intuition and approaches from our analyses on SSL to interpret
1423 results of high-throughput *qpAdm* protocols applied to human archaeogenetic data from two
1424 preprints: a study of the Paleolithic to Bronze Age Siberia (Zeng *et al.* 2023) and a study of
1425 Medieval Europe relying on mostly Iron Age and Roman-period source groups (Speidel *et al.*
1426 2024). Considering that a much more ancient period was studied in the former paper, and
1427 that the archaeogenetic sampling in Siberia is much poorer as compared to Europe, the
1428 studies are markedly different in the sparsity of their geographic sampling (all the proxy
1429 sources/outgroups and "fixed outgroups" used by Speidel *et al.* come from Europe and
1430 Anatolia, except for "Ekven IA" from Chukotka; these are small regions as compared to North,
1431 Central, and East Eurasia in Zeng *et al.*).

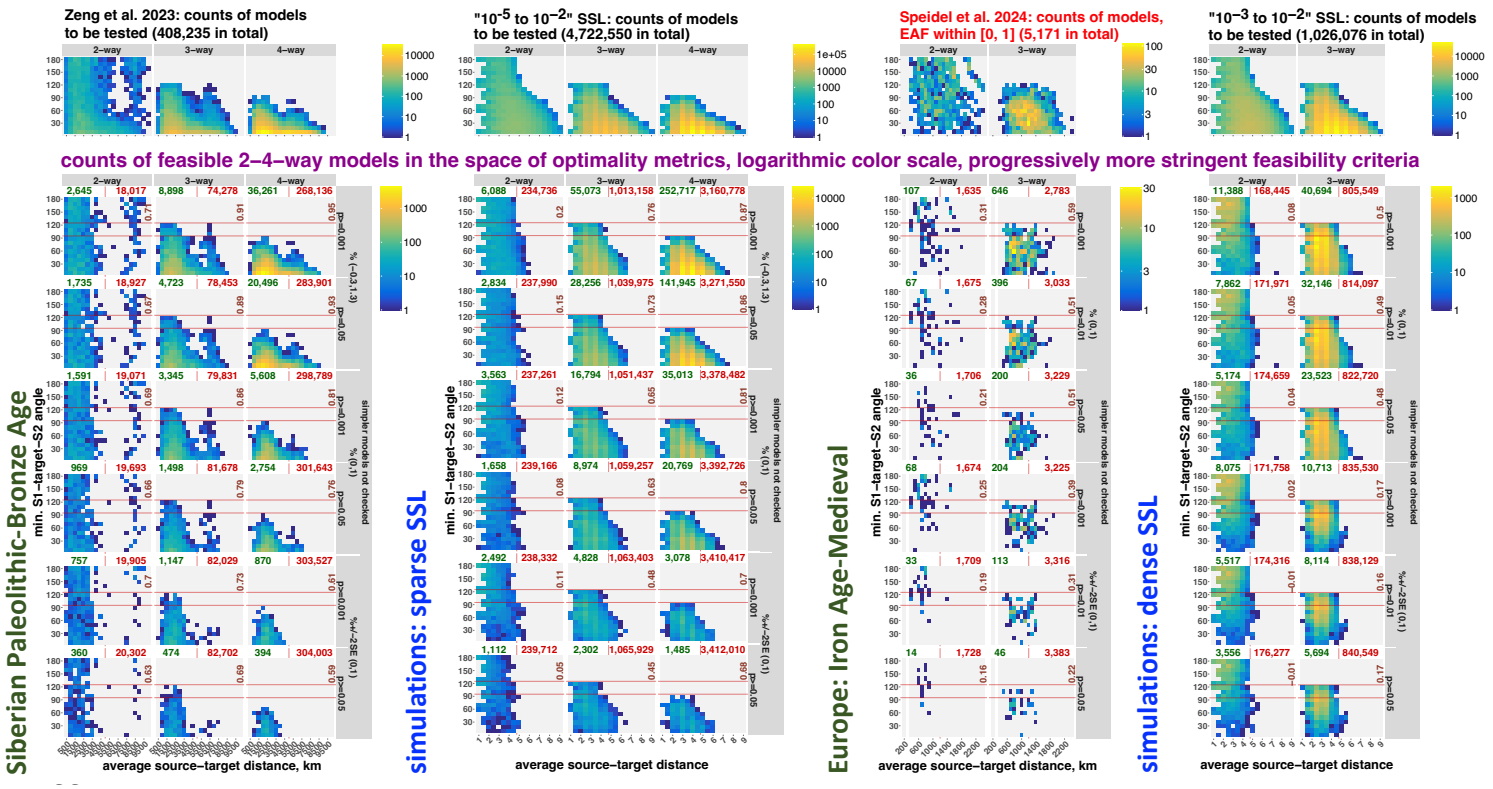


Figure 11. Comparing *qpAdm* models analyzed by the high-throughput protocols from Zeng *et al.* (2023) and Speidel *et al.* (2024) and results of the randomized *qpAdm* protocol on the "10^{-5} to 10^{-2}" and "10^{-3} to 10^{-2}" landscapes, respectively, by placing them in the spaces of optimality metrics (average ST distance vs. min. STS angle). Results are stratified by model complexity. The upper row of plots shows density of all published models in this space (all models tested in the case of Zeng *et al.* and models with EAF ∈ [0, 1] in the case of Speidel *et al.*), 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 x-axis is divided into 20 or 22 bins in the case of the real data and 18 bins in the case of the simulated data; the y-axis is divided into 18 bins (intervals [0°, 10°], (10°, 20°], and so forth). Spearman's correlation coefficients for counts of all published 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 midpoints (5,000 or 1,200 km) between minimal and maximal average ST distances on the real landscapes. 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

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

1459 metrics defined in the SSL context (**Box 2**). The analysis is based on ca. 422,000 two- to four-
1460 way models collected across 725 distal and proximal rotating experiments from Zeng *et al.*
1461 (some experiments in the cited study progressed far beyond four-way models, but to match
1462 our randomized protocol on SSL, these complexity classes were not considered) and on ca.
1463 5,200 two- to three-way models collected across 28 distal and proximal rotating experiments
1464 from Speidel *et al.* (see Suppl. table 3 in that study). All the model metrics (EAF, *p*-values, and
1465 geographical coordinates of individuals) were taken from the respective studies. We note
1466 that neither in the preprint (Speidel *et al.* 2024), nor in the final study version (Speidel *et al.*
1467 2025) unfiltered *qpAdm* results are reported since models with EAF ∉ [0, 1] were omitted
1468 ("we remove models with infeasible admixture proportions", to quote the study).

1469 Placing the real *qpAdm* models from Zeng *et al.* in the spaces of max|*EAF*–*EF*| and *p*-values
1470 and juxtaposing them with the most realistic sparse SSL simulated in this study ("10^{-5} to 10^{-}
1471 ^{2}"; **Fig. S14**), we observe that distributions of models in these spaces are strikingly similar (**Fig.**
1472 **S38**; we cannot use the results from Speidel *et al.* for this analysis since models with EAF ∉
1473 [0, 1] were not presented). Model optimality metrics such as min. STS angles and
1474 maximal/average ST distances are correlated with max|*EAF*–*EF*| and (weakly) with *p*-values
1475 on both the real and simulated data (**Fig. S38, Table S14**). It is not unexpected that on the
1476 simulated SSL data the correlations are stronger (**Table S14**) since, as mentioned above, on
1477 the real data we used just great circle distances, and not realistic travel distances on the
1478 landscape of Eurasia. Moreover, some groupings on the real data cover large areas of
1479 irregular shape (Zeng *et al.* 2023), adding to uncertainties about their centroids.

1480 At the next step, we explored how both real admixture models from Zeng *et al.* and simulated
1481 models populate the spaces of optimality metrics (average ST distances vs. min. STS angles,
1482 **Fig. 11**). Since in both cases all possible models constructed from dozens of alternative
1483 sources were tested for each target, the distributions of models tested have the hallmarks of
1484 random samples: both are skewed towards highly asymmetric models with distant sources
1485 (those with min. STS angles close to 0 and average ST distances approaching the landscape
1486 radius). And on both the real and simulated data, *qpAdm*, equipped with stringent composite
1487 feasibility criteria, does reject efficiently such models with average ST distances above a
1488 certain threshold (3,500 km for the real landscape from Zeng *et al.*). However, in the
1489 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. 11**). 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. 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 just for the simplest, two-way, models (**Figs. 6, 11, Table S11**), but in the case of the real data these distributions are correlated even for two-way models (**Fig. 11**).

Although we do not have access to the starting model distributions from Speidel *et al.*, we can juxtapose the distributions of fitting models in the optimality metric spaces to those on our simulated data. We do observe enrichment of the set of fitting three-way models for relatively symmetric source configurations and relatively small ST distances, which matches our results on the densest simulated landscapes, "10^{-3} to 10^{-2}" (**Fig. 11**). The *Twigstats* algorithm by Speidel *et al.* (2024, 2025) is different from the original *qpAdm* in one way only: it is based on more accurate (Speidel *et al.* 2025) f_2-statistics calculated from genome-wide genealogical trees for haplotypes rather than from allele frequencies, and these f_2-statistics are then passed to the *ADMIXTOOLS 2* package (Maier *et al.* 2023) for calculation of f_4-statistics and *qpAdm* (the same package was used by Zeng *et al.* and in our study). Our juxtaposition with the simulated results on SSL of varying sparsity (**Fig. 11**) suggests that denser landscape sampling is probably enough to explain the superior performance of *Twigstats* as compared to that of the classic *qpAdm* from Zeng *et al.* Moreover, it is unclear if the interpretations of fitting *qpAdm* models in terms of historical human migrations (Speidel *et al.* 2025) are justified: while some asymmetric source configurations and some distant proxy sources are retained by the composite feasibility criterion used in the original study (p-value >0.01, EAF ∈ [0, 1]), with increasing criteria stringency distant sources and asymmetric source configurations (min. STS angle <60°) are progressively pruned (**Fig. 11**). In other words, *qpAdm* models are retained based on their geometric properties, but historically-attested long-distance migrations are not expected to be aligned with them.

Discussion

1520 ***AGS histories: general conclusions***

1521 Temporal stratification tested in this study and practiced in the literature is of two sorts: 1)
1522 most or all populations in the "right" set are sampled deeper in the past than those in the
1523 "left" set (non-rotating protocols); 2) a target group post-dates (or is as old as) all its proxy
1524 sources (distal protocols). Since complexity of AGS histories (number of admixture events)
1525 grows over time, temporal stratification of targets and proxy sources is a very efficient way
1526 of reducing FDR of *qpAdm* protocols (**Fig. 3, Table S1b**) since it avoids situations where
1527 admixture history of proxy sources is more complex than that of the target, which increases
1528 the probability of encountering violations of the topological assumptions of *qpAdm* (**Fig. 1**),
1529 and temporal stratification is indeed common in the *qpAdm* literature. The distal rotating and
1530 non-rotating protocols invariably demonstrated FDR significantly lower than those of the
1531 rotating and non-rotating protocols without temporal stratification (**Fig. 3**). Both distal
1532 protocols demonstrated FDRs around 20% and not different significantly from each other, at
1533 least in the case of 3,000-Mbp-sized genomes (**Fig. 3**). In contrast, application of the proximal
1534 rotating protocol (**Box 1**) that can be summarized as "whatever is not on the right is on the
1535 left" without any attempt at temporal stratification of the "right" and "left" sets and of proxy
1536 sources and targets carries a risk of an FDR above 50%. Adding further levels of support
1537 (considering only models supported by PCA and/or an *ADMIXTURE* analysis) does not help to
1538 decrease FDR drastically in this case (**Table S2, Suppl. text 3**). Although the proximal model
1539 competition protocol (Narasimhan *et al.* 2019; Carlhoff *et al.* 2021; Zhang *et al.* 2021; Maróti
1540 *et al.* 2022; Brielle *et al.* 2023; Lee *et al.* 2023) was not tested on multiple simulation or
1541 subsampling replicates (**Table S5**), we note that it demonstrated FDR values higher than those
1542 of the distal non-rotating protocol (**Table S4**). It is hard to extrapolate this aspect of our
1543 results to real analyses and predict FDR on real data since they depend on pre-study odds,
1544 and they in turn depend on complexity and other topological properties of simulated
1545 histories, but only one graph complexity level was tested, that is 13 groups and 10 admixture
1546 events.

1547 A potential way of improving FDR of high-throughput *qpAdm* protocols applied to AGS
1548 histories is combining *qpAdm* with an unsupervised *ADMIXTURE* analysis (**Suppl. text 4**),
1549 however this method has known problems even on high-quality data and AGS histories
1550 (Lawson *et al.* 2018), and our recommendation is tentative. f_3-statistic is a simple method for

1551 proving that a population is admixed, and it demonstrated FDR values much lower (6%, see
1552 **Suppl. text 3**) than those of standalone *qpAdm* protocols, but f_3-statistics are applicable only
1553 to recent admixture events and/or populations of large effective size since post-admixture
1554 drift on the target lineage obscures the signal (Patterson *et al.* 2012; Peter 2022). Moreover,
1555 calculating f_3-statistics for a target composed of a single pseudo-haploid individual is
1556 impossible since a heterozygosity estimate is required (Maier *et al.* 2023), and such singleton
1557 groups are common in archaeogenetic studies. Researchers should also be aware that f_3-
1558 statistics are defined on unrooted trees, and that may lead to rare but strong false signals of
1559 admixture (**Fig. S6f**).

1560

1561 ***If population history conforms to SSL, qpAdm results are probably misleading in practice***

1562 Although it is very often assumed that the genetic history of humans and many other species
1563 is approximated by complex admixture graphs (see a short review in Maier *et al.* 2023), this
1564 assumption remains unproven. Strictly speaking, it remains unknown if admixture history of
1565 humans, wolves (Bergström *et al.* 2022), and horses (Librado *et al.* 2021, 2024; Taylor *et al.*
1566 2023), to cite some recent examples from the archaeogenetic literature, is approximated
1567 reasonably well by admixture graphs composed of isolated populations and episodes of
1568 admixture, or rather by two-dimensional stepping-stone models, or a combination thereof.

1569 We discuss here several aspects of *qpAdm* protocols on SSL demographies that, in our view,
1570 lie at the core of the "high FDR problem" affecting typical high-throughput *qpAdm* protocols:
1571 1) the ratio of true or optimal to false or non-optimal models (mostly highly asymmetric
1572 models with at least one distant proxy source) is low both for relatively small sets of 10 to 18
1573 alternative sources we constructed per simulated target and for the *qpAdm* screens on real
1574 data we re-visited and falls rapidly with increasing model complexity (**Figs. 6, 7, 11, Table**
1575 **S13**); 2) a small fraction of suboptimal *qpAdm* models demonstrates EAF within (0, 1) (**Figs.**
1576 **S39, S40, Suppl. text 6**) and/or *p*-values above the typical thresholds between 0.001 and 0.1,
1577 contributing to FPR <1% (**Figs. 6, 7, 9, 11**); 3) conversely, a large fraction of optimal *qpAdm*
1578 models demonstrates EAF outside (0, 1) (this fraction grows with increasing model complexity
1579 and reflects growing uncertainties in EAF estimates) and *p*-values below the typical
1580 thresholds (this fraction falls with increasing model complexity), contributing to FNR between
1581 50% and 100% (**Figs. 7, 9**). Given the low pre-study odds common in exploratory studies

1582 (Ioannidis 2005), even very small FPR values are enough to inflate FDR to ca. 50%–100%, with
1583 high FNR playing a role too (**Figs. 7, S24, Table S13, *equation 1***). Since many equally valid
1584 thresholds for classifying admixture models into optimal and non-optimal ones can be set in
1585 the SSL context, we also visualize and compare geometric properties of models in pre-study
1586 sets and in sets of fitting models (**Figs. 6, 11, S23**). We note that these three reasons behind
1587 high FDR (low pre-study odds, very high FNR, and non-zero FPR) are shared by all the high-
1588 throughput *qpAdm* protocols we tested: on AGS, on randomly and systematically sampled
1589 SSL, despite different simulation approaches and definitions of optimal and non-optimal
1590 models. An interplay between pre-study odds and the properties of the simulated landscapes
1591 (mainly sampling sparsity) and composite feasibility criteria affecting FPR and FNR (1 – power)
1592 creates the patterns that we explored on simulated and real data with an emphasis on
1593 practical applications, both at the level of individual *qpAdm* models (**Figs. 5, 7, 9**) and at the
1594 level of "experiments" (**Figs. 8, 10**). Below we discuss some of the most important patterns
1595 observed in this complex system.

1596 First, due to gene flows in all directions violating the topological assumptions of *qpAdm*
1597 designed in the AGS framework (**Fig. 1**), correlation between *p*-values and distance- and
1598 angle-based model optimality metrics on SSL is very poor, but EAF correlate with both types
1599 of optimality metrics better (**Fig. 5, Table S8**). In addition, our analysis on AGS histories
1600 showed that EAF are stable across simulation/subsampling replicates, while *qpAdm p*-values
1601 vary by many orders of magnitude (**Fig. S9**). Thus, considering both AGS- and SSL-based
1602 results we conclude that *p*-value-based model sorting (Narasimhan *et al.* 2019; Fernandes *et*
1603 *al.* 2020; Bergström *et al.* 2022; Lazaridis *et al.* 2022; Brielle *et al.* 2023; Carlhoff *et al.* 2023;
1604 Speidel *et al.* 2025), where slight *p*-value differences may affect if a model receives
1605 interpretation or not, is problematic.

1606 Another tricky aspect of the composite model feasibility criteria common in the literature is
1607 the fact that EAF lie largely within (0, 1) and correlate with simulated gene-flow intensities
1608 only for sources located symmetrically around the target, but for asymmetric sources they
1609 vary widely both outside and, complicating the matters even further, inside (0, 1) (**Figs. S39,**
1610 **S40, Suppl. text 6**). This property of admixture fraction estimation is not recognized in
1611 practice and thus may often confound interpretation of *qpAdm* results on SSL (if the
1612 interpretation of rejected and fitting source configurations is not explicitly spatial). For

1613 instance, a very distant proxy source may be chosen instead of a much closer one (**Fig. S30**)
1614 because the latter configuration of sources happens to be slightly less symmetric and yields
1615 a several orders of magnitude lower *p*-value due to complex patterns of "left-right" gene
1616 flows on SSL. And highlighting a distant source and rejecting a close source may lead to false
1617 findings of long-distance migrations.

1618 For many species, geographical regions and historical periods, archaeogenetic sampling is
1619 very sparse and may be random with respect to population density of ancient individuals in
1620 those regions/periods. Moreover, a general quantitative algorithm for selection of optimal
1621 "right" group sets for *qpAdm* analyses has never been published, and "right" groups are
1622 usually chosen based on qualitative criteria (Harney *et al.* 2021) or borrowed from study to
1623 study. Motivated by these considerations and by the occurrence in the literature of *qpAdm*
1624 protocols exploring thousands of alternative models for dozens of target groups (e.g.,
1625 Narasimhan *et al.* 2019; Lazaridis *et al.* 2022; Zeng *et al.* 2023; Speidel *et al.* 2025; the latter
1626 two studies were re-visited here), we aimed to find expectations for *qpAdm* performance
1627 under random sampling of "right", target, and source demes from SSL, and our central
1628 analysis relies on thousands of such random sets of "right" demes or random rotating sets of
1629 two sizes (**Figs. 5, 6, 8**). Highly asymmetric models with distant sources are by far the most
1630 common in random sets of demes drawn from simulated landscapes (**Figs. 6, S23**) and real
1631 archaeogenetic data (**Fig. 11**): that is, pre-study odds are low (**Fig. 7**). Efficient rejection of
1632 these non-optimal models (with average ST distance greater or approximately equal to the
1633 landscape radius) is possible only for very specific combinations of landscape sparsity
1634 (influencing mostly power = 1 – FNR) and model feasibility criteria (influencing both FPR and
1635 FNR). Furthermore, this optimal combination of sparsity ("10^{-3} to 10^{-2}" landscapes) and
1636 feasibility criteria we found is very uncommon in practice: restricting both "right" and "left"
1637 deme sets to a sub-continental region, using a low *p*-value threshold such as $\geq$0.001, and
1638 requiring that EAF $\pm$ 2 SE $\in$ (0, 1) are very rare approaches, to our knowledge. For example,
1639 in Speidel *et al.* 2025, a study we re-visited, the former condition was satisfied, but the latter
1640 two were not. The "10^{-5} to 10^{-2}" landscapes resemble real *qpAdm* setups much more in
1641 distributions of F_{ST} (**Fig. S14**) and in distributions of models in the max|*EAF*–*EF*| vs. *p*-value
1642 coordinates (**Fig. S38**), and they demonstrated significantly poorer *qpAdm* performance in
1643 this respect (**Figs. 7, S24, Table S11**).

1644 Moving from interpreting individual models to the level of experiments which are usually
1645 stopped as soon as a simplest fitting model is found (see, e.g., Narasimhan *et al.* 2019;
1646 Lazaridis *et al.* 2022; Zeng *et al.* 2023), we see that *qpAdm* protocols generally find either
1647 accurate, but simplistic models (one- to three-way models with sources close to the target
1648 and approximately opposite each other), or complex models which are often inaccurate (**Figs.**
1649 **8, S27, S30**), as illustrated in the FDR-centered discussion above (**Figs. 7, S24**). The range of
1650 model complexities at which experiments end depends on details of *qpAdm* protocols
1651 (rotating vs. non-rotating), SSL sparsity, *p*-value thresholds and conditions on EAF. Systematic
1652 sampling of proxy source and "right" demes in circles around the target deme (**Fig. S16**)
1653 demonstrates similar problems at the level of experiments: when the vicinity of the target is
1654 poorly sampled *qpAdm* tends to report very complex inaccurate models, but when the vicinity
1655 of the target or even the whole landscape is well-sampled, the *qpAdm* experiment approach
1656 tends to find overly simple two- or three-way models (**Figs. 10, S36**). We found a common
1657 explanation both for this counter-intuitive behavior of the particular systematic setup we
1658 tested and for the significant differences in performance of the randomized rotating and non-
1659 rotating experiments (**Figs. 8-10, S24, S27-S29**). That is the influence of average spatial
1660 distance between "left" and "right" deme sets on the probability of encountering genetic drift
1661 shared by proxy and "right" demes to the exclusion of the target or shared by the target and
1662 "right" demes to the exclusion of proxy sources (**Fig. 1, Box 2**) and, consequently, on *qpAdm*
1663 *p*-values. The rotating approach, in contrast to the non-rotating one, allows average "left-
1664 right" distance to be maximized in a single experiment (**Fig. 29c,d**), generating higher *p*-values
1665 on average (**Fig. 29c,d**) and causing randomized rotating experiments to stop at simpler
1666 models (**Figs. 8, S27, S28**). We note that performance of the randomized rotating and non-
1667 rotating protocols is very similar when inspected through the lens of max|*EAF*–*EF*| or *p*-value
1668 correlations with model optimality metrics, and there are no statistically significant
1669 differences between these protocols in FDR (**Figs. S24-S26, Table S8**), but the differences we
1670 discuss at the level of experiments are statistically significant (**Figs. 8, S27, S28**). These
1671 contrasting patterns are seen because for an experiment to end at more complex models it
1672 is enough for spatial proximity of "left" and "right" sets to decrease *p*-values in just few
1673 optimal models available in that experiment (**Fig. S29d**), with other models having very
1674 diverse "left" and "right" arrangements and being rejected for various reasons.

As mentioned above, the accuracy of fitting two-way *qpAdm* models on SSL is much higher than that of more complex models such as three- and four-way (**Figs. 5-7, 11**), but the latter could have been more informative about complex gene-flow topologies of SSL (see two case studies in **Fig. S30**). And stopping the analysis at simple models may (**Figs. 8, 10, S27, S36**) favor thinking about history in terms of simple graphs instead of SSL. Moreover, under the conditions (SSL sparsity level and model feasibility criteria) producing most accurate three- and four-way fitting models (**Figs. 6, 7**), *qpAdm* experiments overwhelmingly end at one- and two-way models (**Figs. 8, S27, S28**). 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. This is especially true since both pre-study odds and SSL sparsity are inherently hard to estimate, and choice of optimal feasibility criteria and optimal model complexity range on SSL depends on these factors in a profound way (**Figs. 6, 7, 9, S24, 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, 2018; Lazaridis *et al.* 2016, 2022; 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; Patterson *et al.* 2022; Taylor *et al.* 2023; Zeng *et al.* 2023) likely falls in the category of sparsely sampled continent-wide landscapes (**Fig. S14**) 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 (**Figs. 6, 7, 9, S24, Table S11**).

Our geographical analysis of two large sets of *qpAdm* models tested on real human archaeogenetic data (Zeng *et al.* 2023; Speidel *et al.* 2024) revealed patterns strikingly similar to those observed on SSL (**Figs. 11, S38**), namely poor enrichment of sets of feasible *qpAdm* models for models optimal in the SSL context (that is, high FDR) in the case of sparse landscape sampling in Paleolithic to Bronze Age North, Central, and East Eurasia (Zeng *et al.* 2023), and much better enrichment (low FDR) in the case of dense landscape sampling in Iron Age to Medieval Europe (Speidel *et al.* 2024) with appropriate composite feasibility criteria. This lack of enrichment for optimal models in the former case 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-

1706 distance landscape with barriers to gene flow. However, our exploration of simulated SSL
1707 suggests another alternative: we are dealing with a sparsely sampled landscape which is not
1708 very suitable for application of high-throughput *qpAdm* protocols with their low pre-study
1709 odds (**Table S11**).

1710 Considering all the problems highlighted above, and the inherent uncertainty about types of
1711 genetic history we are dealing with, we recommend that it is safest to use the *qpAdm* method
1712 in controlled conditions when relationships among populations are understood well enough
1713 to exclude violations of the topological assumptions (**Fig. 1**; these are mostly "proxy-to-right"
1714 and "right-to-proxy" gene flows bypassing the target or "target-to-right" gene flows
1715 bypassing the sources), when radiocarbon or context dates of ancient populations are reliable
1716 and allow accurate temporal stratification, or when sets of potential proxy sources are well-
1717 constrained based on archaeological or historical scholarship (which means that pre-study
1718 odds are high): see, for instance, the earliest publications where the *qpAdm* method was
1719 employed (Haak *et al.* 2015; Mathieson *et al.* 2015) and some recent studies (e.g., Marcus *et*
1720 *al.* 2020; Papac *et al.* 2021; Yaka *et al.* 2021; Changmai *et al.* 2022a, 2022b; Patterson *et al.*
1721 2022). Obviously, the amount of new information that the *qpAdm* method provides in these
1722 conditions is unique for each study and may be low. In contrast, high-throughput *qpAdm*
1723 screens aiming to find optimal models (e.g., Narasimhan *et al.* 2019; Bergström *et al.* 2022;
1724 Lazaridis *et al.* 2022; Zeng *et al.* 2023; Speidel *et al.* 2025) are not recommended, especially
1725 in sparsely sampled regions/periods, such as the Paleolithic in the case of human population
1726 history (e.g., Posth *et al.* 2023; Zeng *et al.* 2023; Villalba-Mouco *et al.* 2023; Allentoft *et al.*
1727 2024), Paleolithic to Neolithic Siberia (Sikora *et al.* 2019; Zeng *et al.* 2023; Gill *et al.* 2024),
1728 ancient South (Narasimhan *et al.* 2019; Kerdoncuff *et al.* 2024) and Southeast Asia (Carlhoff
1729 *et al.* 2021, 2023; Oliveira *et al.* 2022; *qpAdm* played a central role in some of those studies).

1730 In other words, high-throughput *qpAdm* protocols on SSL generate results that very often do
1731 not look suspicious but are either overly simple or inaccurate (favoring very distant sources
1732 over available spatially and genetically close sources). For instance, finding a two-way fitting
1733 model and rejecting all alternative models of the same complexity can be interpreted in
1734 several ways: 1) as a true or (much less likely for distal protocols) a false model in the context
1735 of a relatively simple AGS history; 2) as a suboptimal model (with large ST distances and/or
1736 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 *qpAdm* 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 evidence for long-distance migrations in exploratory studies.

Methods

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). We have not explored *qpAdm* performance over a range of simulated admixture graph complexities, over a range of model feasibility criteria (except for those in **Table S3**), for models more complex than two-way, and have estimated FDR instead of FPR due to an important technical limitation: the process of model classification into true and false ones is difficult to automate fully since it requires careful interpretation of the simulated topology and simulated admixture proportions (this is illustrated by the case studies in **Fig. S6**; note that we attempted to estimate FPR and FNR for *qpAdm* results on AGS histories indirectly, see the end of *section 3*).

Due to this limitation, we did not attempt to classify universal rejections of two-way models by *qpAdm* or other methods into false rejections due to violations of the topological assumptions of *qpAdm* (**Fig. 1**) and true rejections when the true admixture history of the target does not fit a two-way model. Since archaeogenetic sampling is often very sparse, we also did not interpret fitting, but overly simple, models including correct sources as false in the AGS context (see the section "True positive models" in **Suppl. text 2**), but we highlight this problem in the SSL context where the simulated history is even more complex. We started from a collection of random simulated AGS histories of uniform complexity, tested all possible models of chosen complexity classes (only one- and two-way), considered two-way

1767 models supported by a *qpAdm* protocol, classified them manually into false and true positives
1768 according to a set of topological rules, and subjected them to further screening by the PCA
1769 and/or *ADMIXTURE* methods (**Suppl. text 4**).

1770 Another important caveat is that complexity of genetic history in a region and period of
1771 interest often remains unknown, and it is difficult to judge if a particular admixture graph
1772 complexity is adequate for simulating the real history (and if admixture graphs are
1773 appropriate models at all, see below).

1774

1775 ***Limitations of the SSL-based analyses in this study***

1776 While complex admixture models and a wide range of feasibility criteria were not explored
1777 on AGS histories above, the following issues were, conversely, not explored on SSL histories:
1778 effects of differing amounts and quality of data (all the simulations were based on high-
1779 quality diploid data from 500-Mbp-sized genomes with mutation and recombination rates
1780 typical for humans), combining *qpAdm* with other methods (PCA, *ADMIXTURE*, f_3-statistics),
1781 and model competition *qpAdm* protocols. Some of these differences in research coverage
1782 stem from the fact that it is impossible to strictly delineate false and true admixture models
1783 on SSL since at the timescales we explore (up to 300 generations) all demes are connected by
1784 gene flows, directly or indirectly. Instead, it is possible to rank proxy sources by genetic or
1785 spatial proximity to the target and symmetry with respect to the target. Since hard-to-
1786 automate classification of admixture models into false and true ones can be avoided, SSL
1787 histories are more tractable analytically than complex AGS histories.

1788 Since SSL simulations are computationally demanding, our analysis was restricted to relatively
1789 small landscapes composed of 64 demes, small samples per deme (three individuals), and
1790 just three sampling points in time. For the same reason, sampling was not randomized over
1791 >>3 points in time, and different time intervals between sampling points were not tested.
1792 More complex landscapes with rapid long-distance migrations, extinctions, repopulations,
1793 and changes in effective population size were also not explored for reasons of computational
1794 complexity. Thus, our SSL are still simple as compared to the real population history of
1795 humans and other species as we understand them. Ultimately, true isolation-by-distance

1796 landscapes should be explored instead of quantized deme-based landscapes, and the former
1797 can also be simulated by the *slendr* package (Petr *et al.* 2023) we used.

1798 Another notable limitation discussed in *section 5* is our reliance on spatial source-target
1799 distances for measuring model optimality. F_{ST} (**Fig. S14a**) and haplotype-sharing statistics
1800 depend on both source-to-target and target-to-source gene-flows, and for that reason they
1801 are also not ideal as metrics for ranking proxy sources and were not used as such. Ideally,
1802 density clouds of ancestors (Grundler *et al.* 2024) should be inferred for each target and then
1803 used for ranking proxy sources or models or for prediction of EAF. However, this analysis is
1804 also computationally challenging in our setup aiming to explore a wide range of *qpAdm*
1805 models, protocols and landscape properties, and most prominent features of *qpAdm*
1806 performance on SSL such as effects of pre-study odds, landscape sparsity, and average "left-
1807 right" distance can be explored even through the lens of our noisy spatial approximation for
1808 true ancestry density (see **Fig. S14d,e** and its discussion above). Moreover, our spatial
1809 analytical framework based on distances and angles can be directly applied to rich sets of
1810 *qpAdm* models from the literature (**Figs. 11, S38**) without inference of clouds of ancestry
1811 density which is even more complicated for real ancient DNA data.

1812

1813 ***AGS histories: simulations***

1814 For simulating genetic data, we used *msprime v.1.1.1* which allows accurate simulation of
1815 recombination and of multi-chromosome diploid genomes relying on the Wright-Fisher
1816 model (Nelson *et al.* 2020; Baumdicker *et al.* 2022). We simulated 3, 10, or 30 diploid
1817 chromosomes (each 100 Mbp long) by specifying a flat recombination rate (2×10^{-8} per nt per
1818 generation) along the chromosome and a much higher rate at the chromosome boundaries
1819 ($\log_e 2$ or ~0.693 per nt per generation, see
1820 https://tskit.dev/msprime/docs/stable/ancestry.html#multiple-chromosomes). A flat
1821 mutation rate, 1.25×10^{-8} per nt per generation (Scally and Durbin 2012), and the binary
1822 mutation model were used. To maintain the correct correlation between chromosomes, the
1823 discrete time Wright-Fischer model was used for 25 generations into the past, and deeper in
1824 the past the standard coalescent simulation algorithm was used (as recommended by Nelson
1825 *et al.* 2020).

1826 AGS histories of random topology including 13 populations and 10 pulse-like admixture
1827 events were generated using the *random_admixturegraph* and *random_sim* functions from
1828 the *ADMIXTOOLS 2* package
1829 (https://uqrmaie1.github.io/admixtools/reference/random_sim.html), which produced
1830 scripts for running the *msprime v.1.1.1* simulator. Demographic events were separated by
1831 date intervals ranging randomly between 20 and 120 generations, with an upper bound on
1832 the graph depth at 800 generations (or ca. 23,000 years in the case of humans). In other sets
1833 of simulations, all the dates were scaled up 3.75 times, with an upper bound on the graph
1834 depth at 3,000 generations (or 87,000 years in the case of humans). To be more precise,
1835 demographic events were not placed in time entirely randomly, but were tied to one or few
1836 other events of the same "topological depth" within the graph, as illustrated by ten examples
1837 of simulated topologies in **Fig. S3**. The same principle was applied to sampling dates, which
1838 were tied to other demographic events such as divergence and admixture of other
1839 populations. These restrictions were used to ensure topological consistency of random
1840 graphs.

1841 Ten diploid individuals with no missing data were sampled from each population at "leaves"
1842 of the graph. Effective population sizes were constant along each edge and were picked
1843 randomly from the range of 1,000 – 10,000 diploid individuals. Admixture proportions for all
1844 admixture events varied randomly between 10% and 50%. This setup generates groups
1845 sampled at widely different dates in the past or, in other words, located at various genetic
1846 distances from the root. Alternatively, all terminal branches were extended to the "present"
1847 of the simulation and sampled at "present", keeping their respective effective population
1848 sizes and topological relationships unchanged. Thus, another set of simulations was
1849 generated for the same topologies, where groups were more drifted with respect to each
1850 other (see F_{ST} distributions in **Fig. S5**).

1851 In summary, five sets of independent simulations ("simulation setups") differing by the
1852 amount of data generated and by population divergence metrics were performed for a set of
1853 40 random admixture graph topologies (see an overview in **Fig. S2**):

1854 1) three 100-Mbp-sized chromosomes; groups sampled at different points in time; maximal
1855 simulated history depth at 800 generations (10 simulation replicates, for median number of
1856 polymorphic sites see **Fig. S4**);

1857 2) ten 100-Mbp-sized chromosomes; groups sampled at different points in time; maximal
1858 simulated history depth at 800 generations (10 simulation replicates);

1859 3) thirty 100-Mbp-sized chromosomes; groups sampled at different points in time; maximal
1860 simulated history depth at 3,000 generations (10 simulation replicates);

1861 4) three 100-Mbp-sized chromosomes; groups sampled at different points in time; maximal
1862 simulated history depth at 3,000 generations (one simulation replicate);

1863 5) three 100-Mbp-sized chromosomes; all terminal branches extended to the "present" of the
1864 simulation and sampled at that point; maximal simulated history depth at 800 generations
1865 (one simulation replicate).

1866 To create more realistic datasets, we performed randomized subsampling of polymorphic
1867 sites and individuals (replicates no. 1 of the first, second, and third simulation setups were
1868 used for this, see the list above). First, we randomly sampled alleles at heterozygous sites,
1869 creating pseudo-haploid data. Then we introduced missing data by randomly selecting a
1870 missing rate between 5% and 95%, followed by randomly selecting sites according to the
1871 missing rate. This site subsampling was repeated for each individual independently. Lastly,
1872 we randomly sampled *n* (from 1 to 10) individuals from each population independently. The
1873 subsampling procedure described above was conditioned on the number of sites
1874 polymorphic in the set of 13 simulated populations and was repeated until a subsampling
1875 replicate with more than 20,000 (for 300-Mbp-sized genomes) or 66,000 such sites (for 1,000-
1876 and 3,000-Mbp-sized genomes) was obtained (**Fig. S4**). We generated 10 independent
1877 subsampled replicates for each topology and simulation setup (1,200 replicates in total).
1878 Polymorphism data in the *EIGENSTRAT* format were generated from the tree sequences using
1879 the *TreeSequence.genotype_matrix* function (https://tskit.dev/tskit/docs/stable/python-
1880 api.html#tskit.TreeSequence.genotype_matrix) and used for all subsequent analyses (*f*-
1881 statistics and *qpAdm*, PCA, *ADMIXTURE*).

1882

1883 ***AGS histories: calculating f-statistics***

1884 For all the work on *f*-statistics and *qpAdm*, the *ADMIXTOOLS 2* software package (Maier *et al.*
1885 2023) was used. For diploid SNP sets without missing data, we first calculated all possible f_2-
1886 statistics for 4-Mbp-sized genome blocks (with the "*maxmiss=0*",

1887 "*adjust_pseudohaploid=TRUE*", and "*minac2=FALSE*" settings) and then used them for
1888 calculating f_3- and f_4-statistics as linear combinations of f_2-statistics and for testing *qpAdm*
1889 models using the *qpadm* function in *ADMIXTOOLS 2*
1890 (https://uqrmaie1.github.io/admixtools/) under default settings. Harney *et al.* (2021) showed
1891 that *qpAdm* is robust to changes in genome block size except for extreme values: 0.0001 and
1892 1 Morgans (0.05 Morgans is a recommended default value, which corresponds to ~5 Mbp in
1893 the case of humans). Inferred admixture proportions in our analysis were not constrained
1894 between 0 and 1 (see Harney *et al.* 2021 for a justification of this unusual property of the
1895 default *qpAdm* algorithm). For pseudo-haploid SNP sets with missing data and uneven group
1896 sizes, the *qpadm* function was applied to genotype files directly, with the "*allsnps=TRUE*"
1897 setting. In other words, f_4-statistics utilized by *qpAdm* and f_3-statistics were calculated off the
1898 genotype files without intermediate f_2-statistics, and removal of missing data was done for
1899 each population quadruplet or triplet separately. This setup is often used in the literature in
1900 the presence of missing data (e.g., Harney *et al.* 2018, 2019; Narasimhan *et al.* 2019; Lazaridis
1901 *et al.* 2022).

1902

1903 ***AGS histories: qpAdm protocols***

1904 *QpWave* tests were performed on sets of 13 groups divided randomly into 2 "left" and 11
1905 "right" groups, testing all possible bisections of this form. *QpAdm* was applied to the same
1906 sets of 13 groups divided randomly into 3 "left" and 10 "right" groups, testing all possible
1907 bisections of this form for all possible target groups in "left" sets (858 such models per
1908 simulated history). This proximal rotating protocol was applied to all simulation setups.
1909 Subsequent work was focused only on feasible *qpAdm* models defined as follows: 1) *p*-values
1910 calculated by *qpWave* for one-way models "target = proxy source$_1$", "target = proxy source$_2$",
1911 and "proxy source$_1$ = proxy source$_2$" are all below 0.01; 2) in the case of the two-way model
1912 "target = proxy source$_1$ + proxy source$_2$", EAF $\pm$ 2 standard errors are between 0 and 1; 3) the
1913 *p*-value calculated by *qpAdm* for the two-way model $\geq$0.01. For exploring performance of the
1914 distal rotating protocol, feasible two-way *qpAdm* models were simply filtered according to
1915 sampling dates of target groups and proxy sources. If target group's sampling date was equal
1916 to or smaller than sampling dates of both proxy sources, such a model was considered distal.

1917 In the non-rotating protocol, for each simulated admixture graph six oldest sampled groups
1918 were selected as a fixed "right" set (ties in sampling dates were resolved in alphabetical order;
1919 these "right" sets remained unchanged for a given topology across all independent
1920 simulations), and for the remaining seven groups all possible one-way and two-way
1921 admixture models were tested (105 models), applying the same composite feasibility
1922 criterion that was used above for the rotating protocol. This is the proximal non-rotating
1923 protocol, and alternatively we focused on distal admixture models only (distal non-rotating
1924 protocol).

1925 In the proximal model competition protocol, subsequent analysis was focused on targets for
1926 whom two or more alternative *qpAdm* models emerged as feasible at the first step. For each
1927 target, alternative proxy sources were pooled and rotated between the "left" and "right" sets,
1928 testing only the models that emerged as feasible at the first step and applying the composite
1929 feasibility criterion (p-value $\geq$0.01, EAF ± 2 SE are between 0 and 1). Rotation of alternative
1930 proxy sources was performed in two alternative ways: "whatever is not on the left is on the
1931 right", or placement of alternative sources in the "right" set one by one. In the latter case
1932 several "right" sets were tested for each model, and the model was considered supported by
1933 the model competition protocol only if it was not rejected under any of these "right" sets (the
1934 latter protocol follows Maróti *et al.* 2022). If only one model was feasible for a target, such a
1935 model was evaluated as passing the model competition procedure. A distal model
1936 competition protocol was not tested in this study.

1937 For testing statistical significance of differences in FDR between *qpAdm* protocols, the
1938 following approach was used. FDR was calculated either on low-quality data for 10 random
1939 site/individual subsampling replicates derived from simulation replicate no. 1 (simulation
1940 setups no. 1 to 3) or on high-quality data for 10 independent simulation replicates (simulation
1941 setups no. 1 to 3). Comparisons of six *qpAdm* protocols (rotating and non-rotating;
1942 considering exclusively proximal, exclusively distal models, or both) were performed on these
1943 six sets of replicates using the two-sided paired Wilcoxon test (implemented as the
1944 *pairwise_wilcox_test* function in the *rstatix* R package v. 0.7.2). Comparisons of the same
1945 *qpAdm* protocols on simulations with different amounts or quality of data were performed
1946 using the two-sided non-paired Wilcoxon test since simulation replicates were independent
1947 unlike alternative *qpAdm* protocols applied to the same data (**Fig. 3**). In all applications of the

1948 Wilcoxon test for various purposes (**Figs. 3, 10, S37, Table S1b**), *p*-values adjusted for multiple
1949 testing with the Holm method are presented.

1950 For comparing numerical results of *ADMIXTOOLS 2* v. 2.0.6 (released in 2024) with those of
1951 classic *ADMIXTOOLS* v. 7.0 (*qpAdm* v. 1201, released in 2020), we used the most realistic AGS
1952 simulations we generated, that is low-quality pseudohaploid data for 3,000-Mbp-sized
1953 genomes: 10 AGS topologies, with one simulation and one data subsampling replicate per
1954 topology. All possible two-way models were tested with the rotating protocol. *ADMIXTOOLS*
1955 *2* was run with the "*fudge_twice=TRUE*" argument which makes the results more similar
1956 numerically to those of classic *ADMIXTOOLS*. Since the data was of low quality
1957 (pseudohaploid, with high missing rates, and with some populations composed on one
1958 individual), we used the "*allsnps=TRUE*" setting in *ADMIXTOOLS 2* and the identical
1959 "*useallsnps: YES*" setting in classic *ADMIXTOOLS* (all the other settings were default ones). If
1960 "*useallsnps*" is set to "*YES*", the latest *ADMIXTOOLS* v. 8.0.1 (*qpAdm* v. 2050, released in 2024)
1961 automatically calls *qpfstats*
1962 (https://github.com/DReichLab/AdmixTools/blob/master/README.qpfstats), an algorithm
1963 for handling low-quality data that is substantially different from that of the "*allsnps=TRUE*"
1964 setting in *ADMIXTOOLS 2* (https://uqrmaie1.github.io/admixtools/reference/qpfstats.html).
1965 To perform a fair comparison, we used *ADMIXTOOLS* v. 7.0 instead of the latest version v.
1966 8.0.1.

1967 The results produced by these two packages are not identical, but highly correlated across
1968 8,580 unique models we tested on low-quality data (**Table S15**). Pearson's correlation
1969 coefficient is 0.99 for *p*-values (this analysis prioritizes high *p*-values), 0.91 for $\log_{10}(|EAF_1|)$,
1970 0.73 for $\log_{10}$(*p*-values), and 0.79 for $\log_{10}$(SE). Spearman's coefficients are slightly higher for
1971 the logarithmic values (all coefficients are statistically significant with two-sided *p*-values =
1972 0). The weaker correlation seen for logarithmic *p*-values is in line with the high variability in
1973 logarithmic *p*-values calculated by *ADMIXTOOLS 2* on simulation replicates (**Fig. S9**).

1974

1975 ***AGS histories: classifying two-way admixture models into false and true positives***

1976 Since the simulated admixture graph topologies were complex and random, target groups
1977 modelled with *qpAdm* had complex admixture history in some cases, being a part of gene

1978 flow networks. In this context it is hard to draw a strict boundary between true and false
1979 admixture models composed of a target and only two proxy sources. Two-way admixture
1980 models were considered false only if at least one of the following criteria was satisfied
1981 (considering only graph topologies and admixture proportions):

1982 1. The target and at least one of the proxy sources are simulated as strictly cladal (**Figs. 2,**
1983 **S6a**). In this case the target may either be unadmixed, or it may have experienced gene
1984 flows earlier in its history that do not break its cladality with one of the proxy sources.

1985 2. A proxy source lineage is a recipient of gene flow from the target lineage (after the last
1986 admixture event in a target's history), possibly mediated by other lineages (**Figs. 2, S6a,b**).
1987 In other words, the incorrect proxy source is a descendant of the target lineage, i.e., the
1988 expected gene flow direction is reversed.

1989 3. A proxy source does not represent any true source. In other words, it is symmetrically
1990 related to all true sources of ancestry in the target (**Fig. S6c,d**). Alternatively, both proxy
1991 sources represent the same true source, and are symmetrically related to all the other
1992 true sources (**Fig. S6e**).

1993 4. A proxy source shares genetic drift with the corresponding true source that is not shared
1994 by the second proxy source (and the same is true for the other proxy source and another
1995 true source, i.e., condition no. 3 above is not satisfied), however less than 40% of its
1996 ancestry is derived from the true source (**Fig. S6a,d**).

1997 5. The target gets a gene flow from a deep-branching source not represented by any
1998 sampled population, and an inappropriate proxy source is included in the fitting model
1999 (**Fig. S6f,g**).

2000 We illustrate these topological rules with eight case studies of FP and feasible *qpAdm* models
2001 in **Figs. 2** and **S6a-g**. Two-way models for targets whose population history is best
2002 approximated with three-way and more complex models were considered as TP if they
2003 included source proxies (that do *not* satisfy the criteria above) for at least two of three or
2004 more true ancestry sources (see three case studies in **Fig. S6k-m**).

2005

2006 ***AGS histories: principal component analysis***

2007 PCA was performed for one simulation replicate per simulation setup. Only high-quality data
2008 was used, and all individuals sampled from a simulation were co-analyzed. Prior to the
2009 analysis, sites with allelic states in statistical association were pruned with *PLINK v.2.00a3LM*
2010 (Chang *et al.* 2015) using the following settings: window size, 2000 SNPs; window step, 100
2011 SNPs; r^2 threshold = 0.5 (argument "--indep-pairwise 2000 100 0.5"). This kind of linkage
2012 disequilibrium (LD) pruning is common in the population genetic literature since complex LD
2013 structure may confound PCA, especially higher PCs (Privé et al. 2020). PCA was also
2014 performed using *PLINK v.2.00a3LM* under default settings, calculating 10 PCs. Interactive
2015 three-dimensional plots visualizing PC1, PC2, and PC3 were made using the *plotly* R package.
2016 A two-way admixture model was considered supported by PCA if:

2017 1. the target group (the center of the cluster of target individuals, to be precise) lay between
2018 the clusters of proxy source individuals on a straight line in the three-dimensional PC
2019 space;

2020 2. or if it was located at a distance of no more than three target cluster diameters from that
2021 straight line connecting the proxy source clusters.

2022 The second pattern was more common among both TP and FP two-way admixture models:
2023 1.5 and 1.3 times, respectively (across all non-subsampled simulated datasets). This situation
2024 is expected since many targets represent three-way and more complex mixtures, and due to
2025 post-admixture genetic drift (Peter 2022).

2026

2027 ***AGS histories: ADMIXTURE analysis***

2028 *ADMIXTURE* analysis was performed for one simulation replicate per simulation setup. Only
2029 high-quality data was used, and all individuals sampled from a simulation were co-analyzed.
2030 Prior to the analysis, sites with allelic states in statistical association were pruned with *PLINK*
2031 *v.2.00a3LM* (Chang *et al.* 2015) using the following settings: window size, 2000 SNPs; window
2032 step, 100 SNPs; r^2 threshold = 0.5 (argument "--indep-pairwise 2000 100 0.5"). This kind of LD
2033 pruning relying on r^2 is recommended since the underlying model assumes sites in linkage
2034 equilibrium (Alexander et al. 2009). *ADMIXTURE v.1.3* (Alexander *et al.* 2009) was used in the
2035 unsupervised mode under the default settings. The algorithm was run on each SNP dataset
2036 only once, with the number of hypothetical ancestral populations (*K*) ranging from 3 to 10.

2037 This range was selected since the total number of populations in each simulated history was
2038 13. A two-way admixture model was considered supported by *ADMIXTURE* analysis if:

2039 1. for at least one *K*, at least 5 of 10 target individuals were modelled as a mixture of at least
2040 two ancestry components, with a minor ancestry component exceeding 2%;

2041 2. typically, ancestry component *A* in the target group was shared with at least 5 individuals
2042 in proxy source 1, but not in proxy source 2, and ancestry component *B* was shared with
2043 at least 5 individuals in proxy source 2, but not in proxy source 1 (see examples in **Figs. 2**
2044 and **S6**); in some cases, both components *A* and *B* were found in the proxy sources, but in
2045 different proportions;

2046 3. if only one ancestry component in the target was shared with the two proxy sources, the
2047 model was considered unsupported;

2048 4. ancestry components in the target that are absent in any of the sources were ignored
2049 since three-way and more complex admixture histories are common in the set of random
2050 admixture graphs explored here (see examples of such topologies in **Fig. S6k-m**);

2051 5. ancestry components in a proxy source that are absent in the target were also ignored
2052 since a proxy source may not be fully cladal with the real source (see examples of such
2053 topologies in **Fig. S6i,j,m**).

2054 These rules were designed to reproduce typical reasoning of an archaeogeneticist
2055 interpreting *ADMIXTURE* results. Observing a pattern of ancestry components in the target
2056 group and proxy sources compatible with the admixture model "target = proxy source$_1$ +
2057 proxy source$_2$" for one *K* value was enough for declaring that the model is supported by the
2058 *ADMIXTURE* analysis. This condition was motivated by an observation that models supported
2059 at one *K* value only were equally common among FP and TP *qpAdm* models (10% and 13%,
2060 respectively, across simulation setups no. 1, 2, 4, and 5). Models supported at four or more *K*
2061 values were more common among TP *qpAdm* models (3.3% of FP and 12.6% of TP models
2062 across the four simulation setups).

2063

2064 ***SSL histories: simulations***

2065 The *slendr* v. 0.7.2.9000 R package (Petr *et al.* 2023) with the *msprime* v. 1.2.0 backend was
2066 used for the stepping-stone simulations. Sixty-four panmictic demes (with stable effective
2067 sizes of 1,000 diploid individuals) arose via multifurcation at generation 470, counted from
2068 the past. Starting at generation 471, three gene-flow eras followed each other: "pre-LGM"
2069 lasting 1,400 generations, "LGM" lasting 350 generations, and "post-LGM" lasting 770
2070 generations (LGM stands for the last glacial maximum). Three diploid individuals were
2071 sampled from each deme at the end of the simulation ("present"), at 100, and 300
2072 generations before "present". The demes were arranged on a hexagonal lattice (demes were
2073 placed at vertices of triangles tiling the plane; not to be confused with hexagonal or
2074 honeycomb tiling), forming an approximately circular landscape, where each deme was
2075 connected by bidirectional gene flows to 3 to 6 neighboring demes (**Fig. 4**). Gene-flow
2076 intensities were sampled randomly (in each direction independently) from either uniform or
2077 normal distributions, generating four types of landscapes with the following approximate
2078 per-generation gene-flow intensity ranges: 10^{-5} to 10^{-4}, 10^{-4} to 10^{-3}, 10^{-3} to 10^{-2}, 10^{-5} to 10^{-2}
2079 (the values stand for percentage of individuals in a generation coming from a given
2080 neighboring deme; see **Table S6** for details). In the case of the "10^{-5} to 10^{-2}" landscape, normal
2081 distributions centered at 0 were used (see **Table S6** for their parameters). Random sampling
2082 of the gene-flow intensity distributions was done at the beginning of each era, except for the
2083 "10^{-3} to 10^{-2}" landscape, where sampling was repeated five times per era, at equal intervals
2084 of time. Ten independent simulation replicates were generated for each landscape type, with
2085 gene-flow intensities re-sampled for each replicate. Genome size was 500 Mbp,
2086 recombination rate was simulated at 1×10^{-8} per nt per generation, and mutation rate was
2087 1.25×10^{-8} per nt per generation.

2088 There are so many ways of "unfolding" an initial lattice state via a serial-founder process
2089 represented as a bifurcating tree (Estavoyer and François 2022) that we decided to avoid this
2090 additional level of complexity in our study and initialized the lattice simulations with a
2091 multifurcation, but deep in the past, at 2,220 generations before the oldest sampling time
2092 point. In this way we focus on genetic structure generated by gene-flow networks on SSL
2093 instead of structure generated by phylogenetic processes. If gene-flow intensities on SSL are
2094 relatively high, traces of the initial multifurcation are lost by the beginning of the sampling
2095 epoch, but this is not true for simulations with the lowest per-generation gene-flow

2096 intensities (the "10^{-5} to 10^{-4}" landscapes, see below), as demonstrated by our analysis of one-
2097 way *qpAdm* models interpreted as cladality tests (**Suppl. text 5** and **Fig. S21**).

2098

2099 ***SSL histories: qpAdm protocols with randomized landscape sampling***

2100 Four *qpAdm* protocols were applied to SSL: distal and proximal non-rotating, distal and
2101 proximal rotating. For constructing these *qpAdm* experiments, demes were sampled
2102 randomly from the landscapes, keeping their original sample size of three diploid individuals.
2103 For each protocol, 36 composite feasibility criteria and two deme sampling densities were
2104 tested: 13 or 18 demes and a target per a rotating *qpAdm* experiment; 10 or 15 proxy sources,
2105 10 or 15 "right" demes, and a target per a non-rotating experiment. For brevity, we refer to
2106 these sampling densities as "13 demes" or "18 demes" for all types of protocols. Here we
2107 describe random deme sampling rules for the four protocols, and they are also illustrated in
2108 **Fig. S16**:

2109 1. Distal non-rotating. Target demes were sampled at "present", and for each target two
2110 alternative sets of "right" groups and proxy sources were generated: 10 (or 15, depending
2111 on the sampling density) "right" demes were from 300 generations before present (ca.
2112 8700 years in the case of humans), and 10 or 15 proxy source demes were from 100
2113 generations before present (ca. 2900 years in the case of humans).

2114 2. Proximal non-rotating. Ten or fifteen "right" demes were sampled at 300 generations
2115 before present, 10 or 15 proxy source demes were sampled at 100 generations and at
2116 "present" (demes from the two dates were combined before sampling), and a target
2117 deme was sampled from either the 100-generations or 300-generations sampling point.

2118 3. Distal rotating. Target demes were sampled at "present", and for each target two
2119 alternative rotating sets were generated: 13 or 18 "right"/source demes were sampled
2120 from 300 and 100 generations ago (demes from the two dates were combined before
2121 sampling).

2122 4. Proximal rotating. A rotating set of 14 or 19 demes was sampled from all the three time
2123 points combined, and a single target was sampled from that subset, but 300-generations-
2124 old targets were not allowed.

2125 The *qpAdm* experiments described above consisted of all possible one-, two-, three-, and
2126 four-way models for one target (one-, two-, and three-way models for proxy sources were
2127 also counted towards some feasibility criteria as described below). Before each *qpAdm* test,
2128 per-block f_2-statistics were calculated with *ADMIXTOOLS* v.2.0.0 (Maier *et al.* 2023) for the
2129 selected set of demes under the following settings: "*maxmiss=0"* (no missing data allowed at
2130 the group level), "*blgsize=4000000"* (genome block size in nt), "*minac2=FALSE"*,
2131 "*adjust_pseudohaploid=FALSE"* (the latter two settings have their default values intended for
2132 high-quality diploid data). *QpAdm* tests were performed in *ADMIXTOOLS* v.2.0.0 under the
2133 default settings, except for the "*fudge_twice=TRUE*" setting designed to make numerical
2134 results more similar to those of classic *ADMIXTOOLS* (Patterson *et al.* 2012). Here we describe
2135 36 composite feasibility criteria (arranged from more to less stringent) tested on SSL which
2136 are variations on the criterion used for the work on AGS histories (**Box 1**):

2137 1. Admixture fraction estimates (EAF) $\pm$ 2 SE $\in$ (0, 1); *p*-value threshold for *n*-way models is
2138 either 0.05 or 0.5; all "trailing" simpler models (**Box 1**) are rejected according to the 0.001
2139 *p*-value threshold;

2140 2. EAF $\pm$ 2 SE $\in$ (0, 1); *p*-value threshold for *n*-way models is either 0.001, 0.01 (this version
2141 of the criterion matches the main version applied to AGS histories), 0.05, 0.1, or 0.5; all
2142 "trailing" simpler models are rejected according to the same *p*-value thresholds;

2143 3. EAF $\in$ (0, 1); *p*-value threshold for *n*-way models is either 0.05 or 0.5; all "trailing" simpler
2144 models are rejected according to the 0.001 *p*-value threshold;

2145 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
2146 "trailing" simpler models are rejected according to the same *p*-value thresholds;

2147 5. EAF $\in$ (-0.3, 1.3); *p*-value threshold for *n*-way models is either 0.05 or 0.5; all "trailing"
2148 simpler models are rejected according to the 0.001 *p*-value threshold;

2149 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
2150 0.5; all "trailing" simpler models are rejected according to the same *p*-value thresholds;

2151 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"
2152 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. 4**). 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°, 26.6°, or 20° for two-, three-, and four-way models, respectively). 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.

2182 Forty randomized independent experiments progressing to four-way admixture models
2183 (irrespective of model feasibility assessments) were performed per one simulation replicate,
2184 *qpAdm* protocol (distal and proximal, rotating and non-rotating), and landscape sampling
2185 density (13 or 18 demes) (12,800 independent experiments in total). For the most realistic
2186 landscape type, that with per-generation gene-flow intensities ranging from ca. 10^{-5} to 10^{-2},
2187 160 independent experiments progressing to three-way models were performed per one
2188 simulation replicate, *qpAdm* protocol, and landscape sampling density (12,800 experiments
2189 in total). For the three other landscape types, 40 independent experiments progressing to
2190 three-way models were performed per one simulation replicate, *qpAdm* protocol, and
2191 landscape sampling density (9,600 experiments in total). The former set of experiments
2192 progressing to four-way models was used for all the analyses in the study, and the latter two
2193 sets were employed as additional data for comparing estimated and predicted admixture
2194 fractions only (**Figs. S39, S40**).

2195 Spearman's or Pearson's correlation coefficients and statistical significance of their difference
2196 from 0 were calculated in the spaces of EAF or SE vs. model *p*-values (**Tables S1c, S7, S8, S12**),
2197 or in the spaces of optimality metrics (**Table S11**), using the *cor.test* function from the R
2198 package *stats* v. 4.3.0. Two-sided significance tests were used: for Spearman's test, *p*-values
2199 were computed via the asymptotic *t* approximation and subjected to the Bonferroni
2200 correction. The same approaches were used for all applications of Spearman's (**Tables S1c,**
2201 **S7, S8, S11-S13**) or Pearson's (**Figs. S14, S39, S40, Table S15**) correlation coefficients in this
2202 study.

2203

2204 ***SSL histories: qpAdm protocols with systematic landscape sampling***

2205 For *qpAdm* tests based on systematic landscape sampling, we used the landscapes (10
2206 simulation iterations) with per-generation gene-flow intensities ranging between
2207 approximately 10^{-5} and 10^{-2}. On this landscape, we tested two protocols (**Fig. S16**):

2208 1. distal rotating (in each experiment a target was selected from demes sampled at
2209 "present", and 16 demes sampled 100 generations before "present" formed a rotating
2210 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. S16**). 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. 4, 9, 10, S16, S31, S32, S35, S37, S39,** 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. S29**); 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. S33, S34, S36, 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 and published 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

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.

Funding

O.F, P.C., and P.F. were supported by the Czech Science Foundation (project no. 21-27624S). U.I., J.K., and L.A.V. 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.

Conflict of Interest

The authors declare no conflict of interest.

References

Al-Asadi H, Petkova D, Stephens M, Novembre J. 2019. Estimating recent migration and population-size surfaces. PLoS Genet. 15(1):e1007908. doi: 10.1371/journal.pgen.1007908.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals, Genome Res. 19(9):1655–1664. doi: 10.1101/gr.094052.109.

Allentoft ME, Sikora M, Refoyo-Martínez A, Irving-Pease EK, Fischer A, Barrie W, Ingason A, Stenderup J, Sjögren KG, Pearson A, *et al.* 2024. Population genomics of post-glacial western Eurasia. Nature. 625(7994):301–311. doi: 10.1038/s41586-023-06865-0.

Antonio ML, Gao Z, Moots HM, Lucci M, Candilio F, Sawyer S, Oberreiter V, Calderon D, Devitofranceschi K, Aikens RC, *et al.* 2019. Ancient Rome: A genetic crossroads of Europe and the Mediterranean. Science. 366(6466):708–714. doi: 10.1126/science.aay6826.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, *et al.* 2022. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 220(3):iyab229. doi: 10.1093/genetics/iyab229.

Baxter K. 2004. Extrinsic factors that effect the preservation of bone. The Nebraska Anthropologist. 19:38–45.

Bergström A, Stanton DWG, Taron UH, Frantz L, Sinding MS, Ersmark E, Pfrengle S, Cassatt-Johnstone M, Lebrasseur O, Girdland-Flink L, *et al.* 2022. Grey wolf genomic history reveals a dual ancestry of dogs. Nature. 607(7918):313–320. doi: 10.1038/s41586-022-04824-9.

Brielle ES, Fleisher J, Wynne-Jones S, Sirak K, Broomandkhoshbacht N, Callan K, Curtis E, Iliev L,

Lawson AM, Oppenheimer J, *et al.* 2023. Entwined African and Asian genetic roots of medieval peoples of the Swahili coast. Nature. 615(7954):866–873. doi: 10.1038/s41586-023-05754-w.

Carlhoff S, Duli A, Nägele K, Nur M, Skov L, Sumantri I, Oktaviana AA, Hakim B, Burhan B, Syahdar FA, *et al.* 2021. Genome of a middle Holocene hunter-gatherer from Wallacea. Nature. 596(7873):543–547. doi: 10.1038/s41586-021-03823-6.

Carlhoff S, Kutanan W, Rohrlach AB, Posth C, Stoneking M, Nägele K, Shoocongdej R, Krause J. 2023. Genomic portrait and relatedness patterns of the Iron Age Log Coffin culture in northwestern Thailand. Nat Commun. 14(1):8527. doi: 10.1038/s41467-023-44328-2.

Cerezo-Román JI, Wessman A, Williams H (eds.). 2017. Cremation and the Archaeology of Death. Oxford: Oxford University Press.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 4:7. doi: 10.1186/s13742-015-0047-8.

Changmai P, Jaisamut K, Kampuansai J, Kutanan W, Altınışık NE, Flegontova O, Inta A, Yüncü E, Boonthai W, Pamjav H, *et al.* 2022a. Indian genetic heritage in Southeast Asian populations. PLoS Genet. 18(2):e1010036. doi: 10.1371/journal.pgen.1010036.

Changmai P, Pinhasi R, Pietrusewsky M, Stark MT, Ikehara-Quebral RM, Reich D, Flegontov P. 2022b. Ancient DNA from Protohistoric Period Cambodia indicates that South Asians admixed with local populations as early as 1st-3rd centuries CE. Sci Rep. 12(1):22507. doi: 10.1038/s41598-022-26799-3.

Duforet-Frebourg N, Slatkin M. 2016. Isolation-by-distance-and-time in a stepping-stone model. Theor Popul Biol. 108:24–35. doi: 10.1016/j.tpb.2015.11.003.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol. 28(8):2239–2252. doi: 10.1093/molbev/msr048.

Estavoyer M, François O. 2022. Theoretical analysis of principal components in an umbrella model of intraspecific evolution. Theor Popul Biol. 148:11–21. doi: 10.1016/j.tpb.2022.08.002.

Fernandes DM, Mittnik A, Olalde I, Lazaridis I, Cheronet O, Rohland N, Mallick S, Bernardos R, Broomandkhoshbacht N, Carlsson J, *et al.* 2020. The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean. Nat Ecol Evol. 4(3):334–345. doi: 10.1038/s41559-020-1102-0.

Fernandes DM, Sirak KA, Ringbauer H, Sedig J, Rohland N, Cheronet O, Mah M, Mallick S, Olalde I, Culleton BJ, *et al.* 2021. A genetic history of the pre-contact Caribbean. Nature. 590(7844):103–110. doi: 10.1038/s41586-020-03053-2.

Flegontov P, Işıldak U, Maier R, Yüncü E, Changmai P, Reich D. 2023. Modeling of African population history using *f*-statistics is biased when applying all previously proposed SNP ascertainment schemes. PLoS Genet. 19(9):e1010931. doi: 10.1371/journal.pgen.1010931.

Frichot E, Schoville S, Bouchard G, François O. 2012. Correcting principal component maps for effects of spatial autocorrelation in population genetic data. Front Genet. 3:254. doi: 10.3389/fgene.2012.00254.

Gill H, Lee J, Jeong C. 2024. Reconstructing the genetic relationship between ancient and present-day Siberian populations. Genome Biol Evol. 16(4):evae063. doi: 10.1093/gbe/evae063.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, *et al.* 2010. A draft sequence of the Neandertal genome. Science. 328(5979):710–722. doi: 10.1126/science.1188021.

Gnecchi-Ruscone GA, Szécsényi-Nagy A, Koncz I, Csiky G, Rácz Z, Rohrlach AB, Brandt G, Rohland N, Csáky V, Cheronet O, *et al.* 2022. Ancient genomes reveal origin and rapid trans-Eurasian migration of 7th century Avar elites. Cell. 185(8):1402–1413.e21. doi: 10.1016/j.cell.2022.03.007.

Grundler MC, Terhorst J, Bradburd GS. 2024. A geographic history of human genetic ancestry. doi: 10.1101/2024.03.27.586858. [accessed 2025 Jan 22]. https://www.biorxiv.org/content/10.1101/2024.03.27.586858v2.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E,

Stewardson K, *et al.* 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 522(7555):207–211. doi: 10.1038/nature14317.

Harney É, May H, Shalem D, Rohland N, Mallick S, Lazaridis I, Sarig R, Stewardson K, Nordenfelt S, Patterson N, *et al.* 2018. Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation. Nat Commun. 9(1):3336. doi: 10.1038/s41467-018-05649-9.

Harney É, Nayak A, Patterson N, Joglekar P, Mushrif-Tripathy V, Mallick S, Rohland N, Sedig J, Adamski N, Bernardos R, *et al.* 2019. Ancient DNA from the skeletons of Roopkund Lake reveals Mediterranean migrants in India. Nat Commun. 10(1):3670. doi: 10.1038/s41467-019-11357-9.

Harney É, Patterson N, Reich D, Wakeley J. 2021. Assessing the performance of qpAdm: a statistical tool for studying population admixture. Genetics. 217(4):iyaa045. doi: 10.1093/GENETICS/IYAA045.

Ioannidis JP. 2005. Why most published research findings are false. PLoS Med. 2(8):e124. doi: 10.1371/journal.pmed.0020124.

Kerdoncuff E, Skov L, Patterson N, Zhao W, Lueng YY, Schellenberg GD, Smith JA, Dey S, Ganna A, Dey AB, *et al.* 2024. 50,000 years of evolutionary history of India: Insights from ~2,700 whole genome sequences. doi: 10.1101/2024.02.15.580575. [accessed 2025 Jan 22]. https://www.biorxiv.org/content/10.1101/2024.02.15.580575v2.

Lawson DJ, van Dorp L, Falush D. 2018. A tutorial on how not to over-interpret *STRUCTURE* and *ADMIXTURE* bar plots. Nat Commun. 9(1):3258. doi: 10.1038/s41467-018-05257-7.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, *et al.* 2016. Genomic insights into the origin of farming in the ancient Near East. Nature. 536(7617):419–424. doi: 10.1038/nature19310.

Lazaridis I, Mittnik A, Patterson N, Mallick S, Rohland N, Pfrengle S, Furtwängler A, Peltzer A, Posth C, Vasilakis A, *et al.* 2017. Genetic origins of the Minoans and Mycenaeans. Nature. 548(7666):214–218. doi: 10.1038/nature23310.

Lazaridis I, Alpaslan-Roodenberg S, Acar A, Açıkkol A, Agelarakis A, Aghikyan L, Akyüz U, Andreeva D, Andrijašević G, Antonović D, *et al.* 2022. The genetic history of the Southern Arc: A bridge between West Asia and Europe. Science. 377(6609):eabm4247. doi: 10.1126/science.abm4247.

Lee J, Miller BK, Bayarsaikhan J, Johannesson E, Ventresca Miller A, Warinner C, Jeong C. 2023. Genetic population structure of the Xiongnu Empire at imperial and local scales. Sci Adv. 9(15):eadf3904. doi: 10.1126/sciadv.adf3904.

Librado P, Khan N, Fages A, Kusliy MA, Suchan T, Tonasso-Calvière L, Schiavinato S, Alioglu D, Fromentier A, Perdereau A, *et al.* 2021. The origins and spread of domestic horses from the Western Eurasian steppes. Nature. 598(7882):634–640. doi: 10.1038/s41586-021-04018-9.

Librado P, Tressières G, Chauvey L, Fages A, Khan N, Schiavinato S, Calvière-Tonasso L, Kusliy MA, Gaunitz C, Liu X, *et al.* 2024. Widespread horse-based mobility arose around 2200 BCE in Eurasia. Nature. 631(8022):819–825. doi: 10.1038/s41586-024-07597-5.

Lindig WH. 1964. Tree burial among the Seri Indians. Ethnology. 3(3):284–286. doi: 10.2307/3772884.

Lipson M. 2020. Applying f_4-statistics and admixture graphs: Theory and examples. Mol Ecol Resour. 20(6):1658–1667. doi: 10.1111/1755-0998.13230.

Maier R, Flegontov P, Flegontova O, Işıldak U, Changmai P, Reich D. 2023. On the limits of fitting complex models of population history to *f*-statistics. eLife. 12:e85492. doi: 10.7554/eLife.85492.

Marcus JH, Posth C, Ringbauer H, Lai L, Skeates R, Sidore C, Beckett J, Furtwängler A, Olivieri A, Chiang CWK, *et al.* 2020. Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia. Nat Commun. 11(1):939. doi: 10.1038/s41467-020-14523-6.

Maróti Z, Neparáczki E, Schütz O, Maár K, Varga GIB, Kovács B, Kalmár T, Nyerki E, Nagy I, Latinovics D, *et al.* 2022. The genetic origin of Huns, Avars, and conquering Hungarians. Curr Biol. 32(13):2858–2870.e7. doi: 10.1016/j.cub.2022.04.093.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol Biol Evol. 32(1):244–257. doi: 10.1093/molbev/msu269.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, *et al.* 2015. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 528(7583):499–503. doi: 10.1038/nature16152.

Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, Olalde I, Broomandkhoshbacht N, Candilio F, Cheronet O, *et al.* 2018. The genomic history of southeastern Europe. Nature. 555(7695):197–203. doi: 10.1038/nature25778.

McRae BH. 2006. Isolation-by-resistance. Evolution. 60(8):1551–1561. doi: 10.1554/05-321.1.

McVean G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5(10):e1000686. doi: 10.1371/journal.pgen.1000686.

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, *et al.* 2018. Early human dispersals within the Americas. Science. 362(6419):eaav2621. doi: 10.1126/science.aav2621.

Nakatsuka N, Holguin B, Sedig J, Langenwalter PE 2nd, Carpenter J, Culleton BJ, García-Moreno C, Harper TK, Martin D, Martínez-Ramírez J, *et al.* 2023. Genetic continuity and change among the Indigenous peoples of California. Nature. 624(7990):122–129. doi: 10.1038/s41586-023-06771-5.

Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M, *et al.* 2019. The formation of human populations in South and Central Asia. Science. 365(6457):eaat7487. doi: 10.1126/science.aat7487.

Nelson D, Kelleher J, Ragsdale AP, Moreau C, McVean G, Gravel S. 2020. Accounting for long-range correlations in genome-wide simulations of large cohorts. PLoS Genet. 16(5):e1008619. doi: 10.1371/journal.pgen.1008619.

Ning C, Li T, Wang K, Zhang F, Li T, Wu X, Gao S, Zhang Q, Zhang H, Hudson MJ, *et al.* 2020a. Ancient genomes from northern China suggest links between subsistence changes and human migration. Nat Commun. 11(1):2700. doi: 10.1038/s41467-020-16557-2.

Ning C, Fernandes D, Changmai P, Flegontova O, Yüncü E, Maier R, Altınışık NE, Kassian AS, Krause J., Lalueza-Fox C, *et al.* 2020b. The genomic formation of First American ancestors in East and Northeast Asia. doi: 10.1101/2020.10.12.336628. [accessed 2025 Jan 22]. https://www.biorxiv.org/content/10.1101/2020.10.12.336628v1.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 40(5):646–649. doi: 10.1038/ng.139.

Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, Dulias K, Edwards CJ, Gandini F, Pala M, *et al.* 2019. The genomic history of the Iberian Peninsula over the past 8000 years. Science. 363(6432):1230–1234. doi: 10.1126/science.aav4040.

Oliveira S, Nägele K, Carlhoff S, Pugach I, Koesbardiati T, Hübner A, Meyer M, Oktaviana AA, Takenaka M, Katagiri C, *et al.* 2022. Ancient genomes from the last three millennia support multiple human dispersals into Wallacea. Nat Ecol Evol. 6(7):1024–1034. doi: 10.1038/s41559-022-01775-2.

Papac L, Ernée M, Dobeš M, Langová M, Rohrlach AB, Aron F, Neumann GU, Spyrou MA, Rohland N, Velemínský P, *et al.* 2021. Dynamic changes in genomic and social structures in third millennium BCE central Europe. Sci Adv. 7(35):eabi6941. doi: 10.1126/sciadv.abi6941.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2(12):e190. doi: 10.1371/journal.pgen.0020190.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. Genetics. 192(3):1065–1093. doi: 10.1534/genetics.112.145037.

Patterson N, Isakov M, Booth T, Büster L, Fischer CE, Olalde I, Ringbauer H, Akbari A, Cheronet O, Bleasdale M, *et al.* 2022. Large-scale migration into Britain during the Middle to Late Bronze Age. Nature. 601(7894):588–594. doi: 10.1038/s41586-021-04287-4.

Pavlinskaya LR (ed.). 2009. Sibirskij sbornik-1: Pogrebalnyj obryad narodov Sibiri i sopredelnych territorij (book 2). St. Petersburg: Museum of Antrhopology and Ethnography of the Russian

Academy of Sciences.
Peter BM. 2016. Admixture, population structure, and *F*-statistics. Genetics. 202(4):1485–1501. doi: 10.1534/genetics.115.183913.
Peter BM. 2022. A geometric relationship of *F2*, *F3* and *F4*-statistics with principal component analysis. Philos Trans R Soc Lond B Biol Sci. 377(1852):20200413. doi: 10.1098/rstb.2020.0413.
Petkova D, Novembre J, Stephens M. 2016. Visualizing spatial population structure with estimated effective migration surfaces. Nat Genet. 48(1):94–100. doi: 10.1038/ng.3464.
Posth C, Yu H, Ghalichi A, Rougier H, Crevecoeur I, Huang Y, Ringbauer H, Rohrlach AB, Nägele K, Villalba-Mouco V, *et al.* 2023. Palaeogenomics of Upper Palaeolithic to Neolithic European hunter-gatherers. Nature. 615(7950):117–126. doi: 10.1038/s41586-023-05726-0.
Prendergast ME, Lipson M, Sawchuk EA, Olalde I, Ogola CA, Rohland N, Sirak KA, Adamski N, Bernardos R, Broomandkhoshbacht N, *et al.* 2019. Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. Science. 365(6448):eaaw6275. doi: 10.1126/science.aaw6275.
Privé F, Luu K, Blum MGB, McGrath JJ, Vilhjálmsson BJ. 2020. Efficient toolkit implementing best practices for principal component analysis of population genetic data. Bioinformatics. 36(16):4449–4457. doi: 10.1093/bioinformatics/btaa520.
Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature. 461(7263):489–494. doi: 10.1038/nature08365.
Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, *et al.* 2012. Reconstructing Native American population history. Nature. 488(7411):370–374. doi: 10.1038/nature11258.
Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 13(10):745–753. doi: 10.1038/nrg3295.
Sikora M, Pitulko VV, Sousa VC, Allentoft ME, Vinner L, Rasmussen S, Margaryan A, de Barros Damgaard P, de la Fuente C, Renaud G, *et al.* 2019. The population history of northeastern Siberia since the Pleistocene. Nature. 570(7760):182–188. doi: 10.1038/s41586-019-1279-z.
Sirak KA, Fernandes DM, Lipson M, Mallick S, Mah M, Olalde I, Ringbauer H, Rohland N, Hadden CS, Harney É, *et al.* 2021. Social stratification without genetic differentiation at the site of Kulubnarti in Christian Period Nubia. Nat Commun. 12(1):7283. doi: 10.1038/s41467-021-27356-8.
Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, Salie T, Rohland N, Mallick S, Peltzer A, *et al.* 2017. Reconstructing prehistoric African population structure. Cell. 171(1):59–71.e21. doi: 10.1016/j.cell.2017.08.049.
Soraggi S, Wiuf C. 2019. General theory for stochastic admixture graphs and *F*-statistics. Theor Popul Biol. 125:56–66. doi: 10.1016/j.tpb.2018.12.002.
Speidel L, Silva M, Booth T, Raffield B, Anastasiadou K, Barrington C, Götherström A, Heather P, Skoglund P. 2024. High-resolution genomic ancestry reveals mobility in early medieval Europe. doi: 10.1101/2024.03.15.585102. [accessed 2025 Jan 22]. https://www.biorxiv.org/content/10.1101/2024.03.15.585102v1.
Speidel L, Silva M, Booth T, Raffield B, Anastasiadou K, Barrington C, Götherström A, Heather P, Skoglund P. 2025. High-resolution genomic history of early medieval Europe. Nature. 637(8044):118–126. doi: 10.1038/s41586-024-08275-2.
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, *et al.* 2023. Early dispersal of domestic horses into the Great Plains and northern Rockies. Science. 379(6639):1316–1323. doi: 10.1126/science.adc9691.
Tricou T, Tannier E, de Vienne DM. 2022. Ghost lineages highly influence the interpretation of introgression tests. Syst Biol. 71(5):1147–1158. doi: 10.1093/sysbio/syac011.
van de Loosdrecht M, Bouzouggar A, Humphrey L, Posth C, Barton N, Aximu-Petri A, Nickel B, Nagel S, Talbi EH, El Hajraoui MA, *et al.* 2018. Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. Science. 360(6388):548–552. doi:

10.1126/science.aar8380.
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, *et al.* 2023. 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(4):597–609. doi: 10.1038/s41559-023-01987-0.
Wang K, Goldstein S, Bleasdale M, Clist B, Bostoen K, Bakwa-Lufu P, Buck LT, Crowther A, Dème A, McIntosh RJ, *et al.* 2020. Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa. Sci. Adv. 6(24):eaaz0183. doi: 10.1126/sciadv.aaz0183.
Wang CC, Yeh HY, Popov AN, Zhang HQ, Matsumura H, Sirak K, Cheronet O, Kovalev A, Rohland N, Kim AM, *et al.* 2021. Genomic insights into the formation of human populations in East Asia. Nature. 591(7850):413–419. doi: 10.1038/s41586-021-03336-2.
Williams MP, Flegontov P, Maier R, Huber CD. 2024. Testing times: Disentangling admixture histories in recent and complex demographies using ancient DNA. Genetics. 228(1):iyae110. doi: 10.1093/genetics/iyae110.
Yaka R, Mapelli I, Kaptan D, Doğu A, Chyleński M, Erdal ÖD, Koptekin D, Vural KB, Bayliss A, Mazzucato C, *et al.* 2021. Variable kinship patterns in Neolithic Anatolia revealed by ancient genomes. Curr Biol. 31(11):2455–2468.e18. doi: 10.1016/j.cub.2021.03.050.
Yang MA, Fan X, Sun B, Chen C, Lang J, Ko YC, Tsang CH, Chiu H, Wang T, Bao Q, *et al.* 2020. Ancient DNA indicates human population shifts and admixture in northern and southern China. Science. 369(6501):282–288. doi: 10.1126/science.aba0909.
Zeng TC, Vyazov LA, Kim A, Flegontov P, Sirak K, Maier R, Lazaridis I, Akbari A, Frachetti M, Tishkin AA, *et al.* 2023. Postglacial genomes from foragers across Northern Eurasia reveal prehistoric mobility associated with the spread of the Uralic and Yeniseian languages. doi: 10.1101/2023.10.01.560332. [accessed 2025 Jan 22]. https://www.biorxiv.org/content/10.1101/2023.10.01.560332v2.
Zhang F, Ning C, Scott A, Fu Q, Bjørn R, Li W, Wei D, Wang W, Fan L, Abuduresule I, *et al.* 2021. The genomic origins of the Bronze Age Tarim Basin mummies. Nature. 599(7884):256–261. doi: 10.1038/s41586-021-04052-7.