OXFORD GENETICS

Low accuracy of complex admixture graph inference from *f*-statistics

Lauren E. Frankel[1] and Cécile Ané[2,*]

[1]Department of Botany, University of Wisconsin - Madison, Madison, 53706, USA
[2]Department of Statistics, University of Wisconsin - Madison, Madison, 53706, USA

*Corresponding author: Department of Statistics, University of Wisconsin-Madison, University Avenue, Madison WI 53706. Email: cecile.ane@wisc.edu

Abstract

F-statistics are commonly used to assess hybridization, admixture or introgression between populations or deeper evolutionary lineages. Their fast calculation from allele frequencies allows for rapid downstream admixture graph inference. One frequently overlooked assumption of the f_4-test is a constant substitution rate. This assumption is typically questionable when comparing distantly-related lineages. Using simulations, we find that rate variation across lineages decreases the accuracy of the f_4-test to detect the presence of reticulations in large data sets or with high average mutation rate. But when f-statistics are combined to infer an admixture graph, rate variation across lineages has a small effect on accuracy. Network inference was accurate on a simple network with 1 reticulation only, but extremely inaccurate to infer a complex network with 4 reticulations, even from large data sets and without rate variation. Yet in both cases, the network's major tree was inferred reliably. Rate variation significantly increased the distance between the true and closest estimated network, the score gap between the true and best-scoring network, and the rate of incorrectly rejecting 1 reticulation as adequate, under our simple network. We propose that identifiability, or lack thereof, is underlying the contrasting results between our simple and complex networks. Our findings suggest that the major tree is one feature that might be identifiable from f-statistics. In practice, we recommend evaluating a large set of top-scoring networks inferred from f-statistics, and even so, using caution in assuming that the true network is part of this set when inferred networks are complex. The extent of rate variation should be assessed in the system under study, especially at deeper time scales, in systems with rapid molecular evolution or with fast-evolving loci.

Keywords: rate variation, hybridization, introgression, gene flow, reticulation, phylogenetic network, identifiability, ABBA-BABA, D-statistic

Introduction

The f_4 statistic is very popular to identify histories of introgression or admixture and is the basis for many high-profile studies (Fig. 1). It uses allele frequencies from 4 populations and measures the correlation of allele frequency differences between pairs of populations (Peter 2016; Soraggi and Wiuf 2019; Lipson 2020; Peter 2022). It extends the D-statistic (Patterson *et al.* 2012; Green *et al.* 2010) to several individuals per population, thus shifting it more towards the realm of population genetics. F-statistics form the basis of many discoveries about archaic human population dynamics (Lipson *et al.* 2018; Wang *et al.* 2021; Flegontov *et al.* 2019), especially as ancient DNA extraction and sequencing techniques have advanced in recent years. f-statistics have also been used to study a variety of other organisms (for example, in domestication of horses, Librado *et al.* (2021); mammoths, van der Valk *et al.* (2021); wild grapes, Zecca *et al.* (2020)). The f_4 statistic is also called the f_4-ratio and interpreted as the "admixture proportion" between two populations. From f_4 statistics, admixture graphs can be readily and rapidly fitted to data and inferred, using tools such as `qpgraph` (Patterson *et al.* 2012), `qpWave` (Haak *et al.* 2015), `qpAdm` (Haak *et al.* 2015; Harney *et al.* 2021) and `find_graphs` (Maier *et al.* 2023) all implemented in the R package `admixtools` (Patterson *et al.* 2012; Maier *et al.* 2023), as well as `poolfstat` (Gautier *et al.* 2022) and `AdmixtureBayes` (Nielsen *et al.* 2023).

One underlying assumption of the f_4 test of admixture, and of other hybridization tests (for example, the D-statistic, Durand *et al.* (2011); Patterson *et al.* (2012); D_3, Hahn and Hibbins (2019); HyDe, Blischak *et al.* (2018)), is a constant substitution rate across genes and/or across lineages. Lineage-specific substitution rate variation occurs when genome-wide rates differ between population lineages, such as due to shifts in physiological or life-history traits between populations. Factors influencing lineage-specific variation include changes in generation time (Wang and Obbard 2023; Thomas *et al.* 2010), body size (Lanfear *et al.* 2013; Martin and Palumbi 1993), and other interacting components. When these tests are applied at a macroevolutionary scale where substitution rates are likely to vary, this could present a problem in the accuracy and interpretation of admixture detection and admixture histories. Even at short evolutionary timescales substitution rates can vary significantly, due to variation in effective population size or traits such as generation interval (Petit and Barbadilla 2009; Abu-Elmakarem *et al.* 2024; Gossmann *et al.* 2012), which have been documented to vary over different timescales between archaic and modern humans (Wang *et al.* 2023; Coll Macià *et al.* 2021; Ragsdale *et al.* 2023).

In simulations, Frankel and Ané (2023) found that substi-

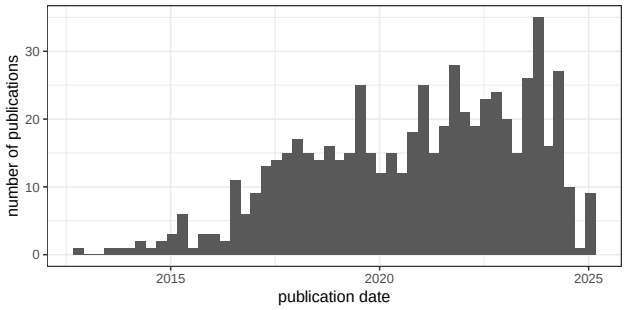

Figure 1 Number of publications in the OpenAlex catalog (Priem *et al.* 2022) with full text including 'admixtools', 'admixtools2' or 'f4 statistics'. Of these 616 publications, 110 were published in *Science* or in one of the following Nature journals: *Scientific Reports*, *Nature*, *Nature Communications*, *Nature Genetics*, *Nature Ecology & Evolution* and garnered a total of 14,770 citations. Data were retrieved using the R package `openalexR` (Massimo *et al.* 2024) on 2025-02-08.

tution rate variation across lineages (rather than across genes or lineage-by-genes) can be problematic for several methods based on 3-taxon and 4-taxon summary statistics. When certain lineages evolve faster or slower than others on average (across the genome), then hybridization summary statistics often erroneously indicate there is hybridization when there is none. Notably, the D-statistic, which the f_4 statistic extends, has a type-1 error rate around 20% when substitution rates varied across lineages at the level simulated by Frankel and Ané (2023).

Frankel and Ané (2023) also found that admixture graph complexity poses a problem to detect hybridization with summary statistics. Specifically, having 2 or more hybridization events in a 10-taxon graph lowered power, even in the absence of rate variation. Compared to D, f_4 has the advantage of using multiple individuals per population. But if it behaves similarly, evolutionary histories with 2 or more hybridization events may not be accurately estimated by `qpgraph` or `find_graphs`.

In this work, we conducted an extensive simulation study to quantify the performance of f_4 tests when rates vary across lineage, and the level of rate variation at which the f_4 test's type-1 error becomes consequential. Further, we quantify the performance of f_4-based admixture graph inference using `find_graphs`, across a range of rate-variation across lineages, and a range of complexity of the true admixture graph.

In this study, to simplify terminology, we use 'graph' and 'network' interchangeably, and we denote by 'reticulation' any biological process that results in a populations inheriting genetic material from multiple parent populations, including admixture, gene flow, and hybridization. We also denote a node with multiple parent populations (e.g. an admixed population) as a 'hybrid' node, regardless of the biological process. Similarly, edges leading to a hybrid node are call hybrid edges, also known as admixture edges when the process is known to be population admixture.

Methods

Simulation

Networks and simulation parameters We simulated data under one of two fixed "true" networks, chosen for being realistic networks of relevance to empirical data (Fig. 2). The most complex network came from Flegontov *et al.* (2023) Fig. 3A, with edge parameters (number of generations, population size, admixture weights) detailed in in their S13 Table. This network includes a chimpanzee, Denisovan, two Neanderthal, two African, and two non-African populations. It also includes a bottleneck event and four hybridization events, so we call it `g4` here. To assess the effect of network complexity on inference, we modified this network `g4` by pruning the 3 minor hybrid edges with lowest admixture weights $\gamma = 0.06$, 0.19 and 0.32). This left just one reticulation event (with minor $\gamma = 0.48$) ancestral to the Denisovan population – so we call this network `g1`.

Simulating gene trees From each network, `g1` and `g4`, we simulated sequence data for 1, 2, or 10 haploid individuals per population, and for various values of mutation rate and level of rate variation (see below). For each combination of parameters, we simulated either 11,800 gene trees or 118,000 under the network multispecies coalescent (NMSC, Degnan 2018) using `PhyloCoalSimulations` v0.1.2 (Fogg *et al.* 2023; Fogg and Ané 2022), written in `Julia` (Bezanson *et al.* 2017). Simulating 11,800 or 118,000 gene trees allowed for a "buffer" of 1,800 or 18,000 gene trees respectively, that resulted in multiallelic sites (instead of biallelic) that cannot be used in f_4 calculations. Gene trees had degree-2 nodes to mark each speciation and reticulation event, which allowed mapping each edge in a gene tree to the population lineage in the species network, that the gene edge evolved in.

Simulating rate variation across population lineages We used five different levels of rate variation across population lineages. The distribution of rates across lineages was set to a log-normal distribution with mean 1 and one of 5 standard deviation σ on the log-scale: $\sigma = 0$ (no rate variation), 0.15, 0.3, 0.5, or 0.7. The 0.1 and 0.9 quantiles of each distribution are $(0.82, 1.2)$ for $\sigma = 0.15$; $(0.65, 1.4)$ for $\sigma = 0.3$; $(0.46, 1.67)$ for $\sigma = 0.5$; $(0.32, 1.92)$ for $\sigma = 0.7$ (see Fig. S1).

For every lineage (population) l in the network (`g1` or `g4`), we assigned a rate r_l drawn from the log-normal distribution with a fixed σ. This rate r_l was applied to this lineage across all gene trees in the replicate data set, as follows.

To convert number of generations to number of substitutions per site in a gene tree edge e, we multiplied its length t_e in generations (from the coalescent simulation) by the average mutation rate μ, either $\mu = 1.25 \times 10^{-7}$ or $\mu = 1.25 \times 10^{-8}$ mutations per site per generation. We then multiplied by the relative mutation rate r_l of the population lineage l that the branch evolved in, simulated as described above. Therefore the final edge length $\ell(e)$ of a gene tree branch e evolving in population l, in substitutions per site, was

$$\ell(e) = t_e \times \mu \times r_l \,. \tag{1}$$

Simulating sequences Once we had 11,800 or 118,000 gene trees with lineage-specific branch lengths scaled relative to their amount of substitution rate variation, we then simulated sequences from these gene trees using `Seq-Gen` v.1.3.4 (Rambaut and Grass 1997). We used much of the same parameters as (Frankel and Ané 2023), including using the HKY model (different rate of transitions and transversions as well as unequal base frequencies of the four nucleotides) with starting base frequencies of 0.2 A, 0.3 C, 0.3 G and 0.2 T and a transition/transversion ratio of 3. Additionally, we simulated rate variation across sites, using a discretized gamma distribution with shape $\alpha = 0.35$ and 10 rate categories for the discretization. To simulate a data set of

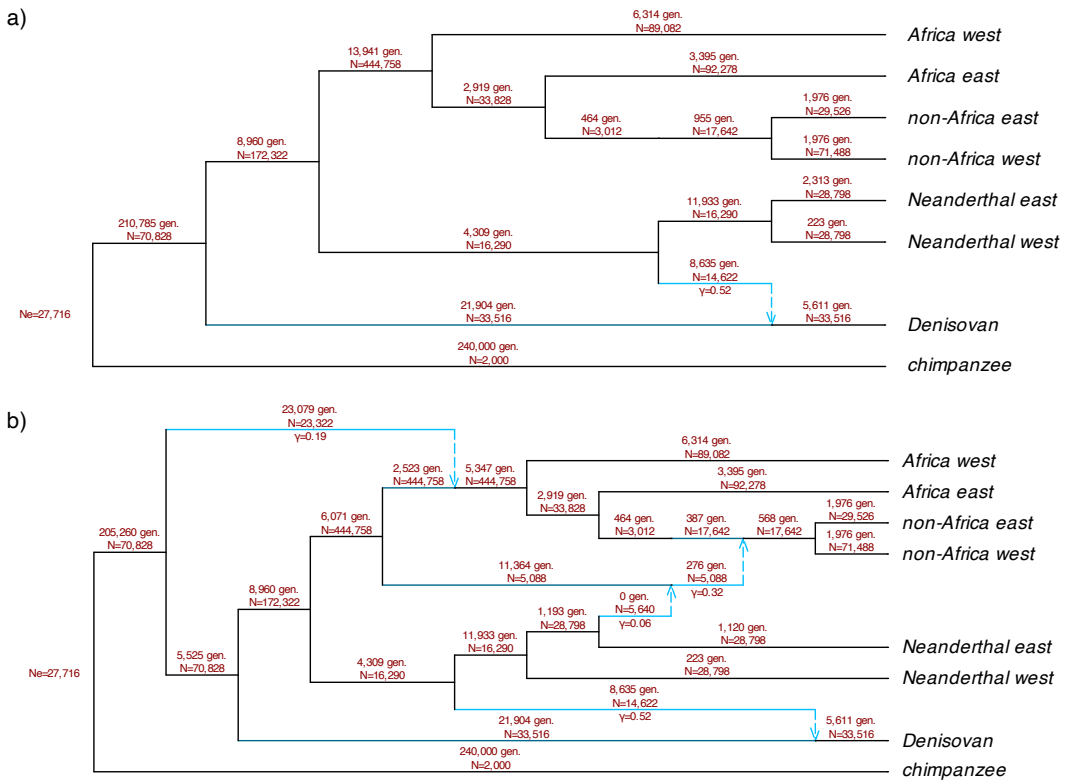

Figure 2 Networks used in simulations. Annotations above each edge indicate number of generations, effective haploid population size N and inheritance probability γ, also called admixture weight, for hybrid edges. a) Network `g1` with $h = 1$ reticulation of large effect (inheritance $\gamma = 0.48$ and 0.52 from its parent populations). b) Network `g4` with $h = 4$ reticulations (from Flegontov *et al.* 2023, see Methods). `g1` is a subnetwork of `g4`, and was obtained be removing the 3 reticulations of lesser effect.

unlinked sites, we aimed to extract a single variable site per gene tree. For each gene tree, we simulated an alignment of 500 base pairs. We took the first biallelic or multiallelic site in each alignment and wrote them to a `VCF` file. If the alignment yielded no biallelic or multiallelic sites, we re-simulated the alignment up to 20 times. We simulated alignments until we reached $n = 10,000$ or $n = 100,000$ biallelic sites.

Once we had a `VCF` file with n (10,000 or 100,000) biallelic sites, we generated a file mapping individual names to population names in `.ind` `EIGENSTRAT` format. From there, we converted the `VCF` file to `EIGENSTRAT` format using the `vcf2eigenstrat.py` Python script from Iain Mathieson, available at https://github.com/mathii/gdc/blob/master/vcf2eigenstrat.py, using our individual mapping file to specify populations. These `EIGENSTRAT` files served as input for `admixtools` analyses, detailed below.

Analysis of the simulated data

Tree-tests from f_4 statistics First, we calculated block-specific f_2 values for each replicate with the `admixtools` function `f2_from_geno` using the `EIGENSTRAT` files as input, using default settings except for specifying blocks of size 1 site. One such default setting was `adjust_pseudohaploid = TRUE`, which ensures that samples are treated as haploid. Using f_2 values, the `f4` function with default settings and `f4mode = TRUE`, we calculated f_4 statistics their standard errors and performed the f_4 tests (of $f_4 = 0$ expected under no reticulation). For each set of 4 populations, 3 f_4-tests can be performed, one for each of the 3 unrooted topologies on 4 taxa. We performed the test corresponding to the split $p_1p_2|p_3p_4$ where $\{p_1, p_2\}$ were the 2 populations that received most support by the data as being sisters. This was determined as the pair having the smallest f_2 statistic, (which measures genetic divergence) among the 6 pairs from the selected 4 populations.

Admixture graph inference We estimated admixture graphs using `find_graphs` from `admixtools`, with chimpanzee as outgroup. Search settings were chosen to match or exceed those in Maier *et al.* (2023)'s simulations, so as to be sure we explored graph space thoroughly, yet in a feasible timeframe (see Supplementary Material, Table S1). Specifically, settings increased from default values were as follows: number of generated and evaluated graphs per generation (`numgraphs`) to 25, total number of generations after which to stop generating graphs (`stop_gen`) to 10,000, number of generations without improvement after which to stop generating graphs (`stop_gen2`) to 60, number of generations without best score improvement before a new graph search strategy is attempted (`plusminus_generations`) to 100, and number of random initializations of starting weights (`numgraphs`) to 100.

We ran `find_graphs` for each replicate with various maximum allowed number reticulations, called "search h" (changing `max_admix`). For data simulated under `g1`, the search h was set to 0 (to estimate a tree), 1 (the true h in `g1`) or 2. For data simulated under `g4`, the search h was set to 0, 1 or 4 (the true h in `g4`). For each replicate data set and each search h, we ran 50 independent runs of `find_graphs`, each started with its own seed. In

all, `find_graph` was run 150 times for each replicate data set (3 search h values $\times$ 50 number of independent runs).

For each data set and each search h, we combined graphs across all 50 independent runs. We then filtered this large ensemble of graphs to keep a set of "top graphs", as follows. We identified the best graph with the highest log-likelihood score, $\hat{S}$, and ranked all graphs by their scores. The first 5 graphs (with the 5 best scores) were kept in the set of top graphs, as well as any graph with a score $S > \hat{S} - 10$.

Keeping multiple top graphs and looking for stable topological features (see below) is in line with best practices recommended by Maier *et al.* (2023). Also note that a drop threshold of 10 in log-likelihood is conservative, leading to a very large set of top graphs. For example, a threshold drop of 3 in log-likelihood, less conservative, could be argued for. It corresponds to an increase of $2 \times 3 = 6$ in the Bayesian information criterion (BIC) for graphs of the same h and same number of free edge parameters – a change traditionally considered as strong evidence to reject the graph with the worse score (Raftery 1995).

For each replicate data set, we also calculated the log-likelihood score of the true graph topology (`g1` or `g4`) using `qpgraph`.

Measures of accuracy

We first calculated the network inference accuracy as the probability that the true network is part of the set of top graphs. Equality (isomorphism) between a top graph and the true network was evaluated based on equal "hash" output by `admixtools`. Given our conservative set of top graphs, this accuracy measure is quite "generous": it does *not* require that the true network has the best log-likelihood. Rather, it requires that the true network has one of the best 5 scores, or has a score within 10 units of the best score found.

As the true `g4` topology was almost never in the set of top graphs, we looked for network features that might be correctly inferred in top graphs.

First, we checked if hardwired clusters were recovered. Each edge in a network is associated with a hardwired cluster: the set of taxa (populations) that contain genetic material inherited from that edge. For example, `g1` contains the cluster {Denisovan, Neanderthal east, Neanderthal west} from its hybrid edge with $\gamma = 0.52$. The *hardwired-distance* $d(N_1, N_2)$ between networks N_1 and N_2 is the number of clusters found in one but not the other network (Huson *et al.* 2010). This is the Robinson-Foulds distance when N_1 and N_2 are trees. More generally, d is not a proper distance, because distinct networks may share the same clusters. Fig. S3 shows networks $N \neq$ `g4` with an identical list of clusters, such that $d(N, \texttt{g4}) = 0$.

Second, we considered trees displayed by the inferred and true networks. The trees *displayed* in a network N are all the trees that can be obtained by keeping exactly one hybrid parent edge for each hybrid node, and deleting all other hybrid edges. One of these trees is the *major tree* of N: obtained by deleting from N any hybrid edge with inheritance weight $\gamma < 0.5$, thus keeping hybrid edges that contribute a majority $\gamma > 0.5$ of genetic material. Both `g1` and `g4` have the same major tree, with Denisovan sister to the Neanderthal populations. For each data set, we determined whether at least one top graph displayed the true major tree (of the true network used to generate the data).

For both `g1` and `g4`, the major tree is obtained by keeping a hybrid edge of weight $\gamma = 0.52$ (into Denisovan), almost tied with the alternative hybrid edge with $\gamma = 0.48$. So we considered the 2 displayed trees obtained by keeping either one of these 2 hybrid edges, and otherwise deleting any minor hybrid edge with $\gamma < 0.48$. These 2 *near-major* trees are the 2 trees displayed by `g1`, with Denisovan either sister to the Neanderthal populations, or sister to all Neanderthal and human populations. Compared to the other trees displayed by `g4`, the 2 near-major trees represent a greater proportion of the genome, so we expected that inferred graphs could display them more often than the other displayed trees. We calculated the number of true near-major trees displayed by each top graph (0, 1 or 2), then took the maximum over top graphs, for each data set.

To check that our settings lead to a sufficiently thorough search of the graph space, for each data set we calculated the score difference $\Delta S = S(\texttt{g}) - \hat{S}$, where $S(\texttt{g})$ is the score of the true graph topology (`g1` or `g4`) and $\hat{S}$ is the best score found during the search. When the search h is equal to the true h, the true topology is part of the space being searched, so a sufficiently thorough search should lead to $\Delta S \geq 0$ (as `admixtools` uses a score equal to the negative likelihood, up to a constant).

To determine the impact of rate variation on model selection on the number of reticulations, we calculated the "worst residual" (WR) for each top graph. This is obtained by comparing the f_4 statistic observed from the data with that expected from the network (residual), calculating a z-score for each and then taking the worst residual z-score across all subsets of 4 populations. In most empirical studies, a graph with an absolute WR below 3 is interpreted as fitting the data adequately. Therefore, we performed the following model selection procedure for each replicate data set. We considered all top graphs obtained under a search h equal to the true h (1 for data generated under `g1`, 4 under `g4`). If at least one of these top graphs had $|\text{WR}| \leq 3$ then we concluded that h (or fewer) reticulations was adequate to explain the data. Otherwise, if all these top graphs had $|\text{WR}| > 3$ then we concluded that this h could be rejected, and more reticulations were necessary to explain the data, resulting in a type-1 error. This threshold of 3 is widely used (e.g. Maier *et al.* 2023; Gutaker *et al.* 2020; Zhao *et al.* 2023; Ge *et al.* 2023). For data generated under `g1`, we also calculated the WR of top graphs obtained under a search h of 2, to determine if $h = 2$ was accepted as adequate in case $h = 1$ was rejected.

Results

Proportion of f_4-tests that incorrectly infer hybridization

In taxon quartets with no reticulation f_4 performed well across mutation rates and number of biallelic sites in the absence of rate variation (Fig. 3, $\sigma = 0$ in pink) staying around the desired 5% rejection rate. Also, type-1 error stayed around 5% for 10,000 biallelic sites and a slow mutation rate (1.25×10^{-8}) regardless of rate variation.

Type-1 error was moderately elevated (as high as $\sim$ 9%) at higher levels of rate variation ($\sigma \geq 0.5$, blue and purple) for 10,000 biallelic sites and a fast mutation rate (1.25×10^{-7}) or 100,000 biallelic sites and a slow mutation rate (1.25×10^{-8}). Type-1 error was at its worst for data with 100,000 biallelic sites and a fast mutation rate, nearing about $\sim$ 25% under our highest level of rate variation.

In scenarios with elevated type-1 error, increasing the number of individuals per population mitigated the problem. There was no clear trend of f_4 performing better or worse depending on network (`g1` or `g4`).

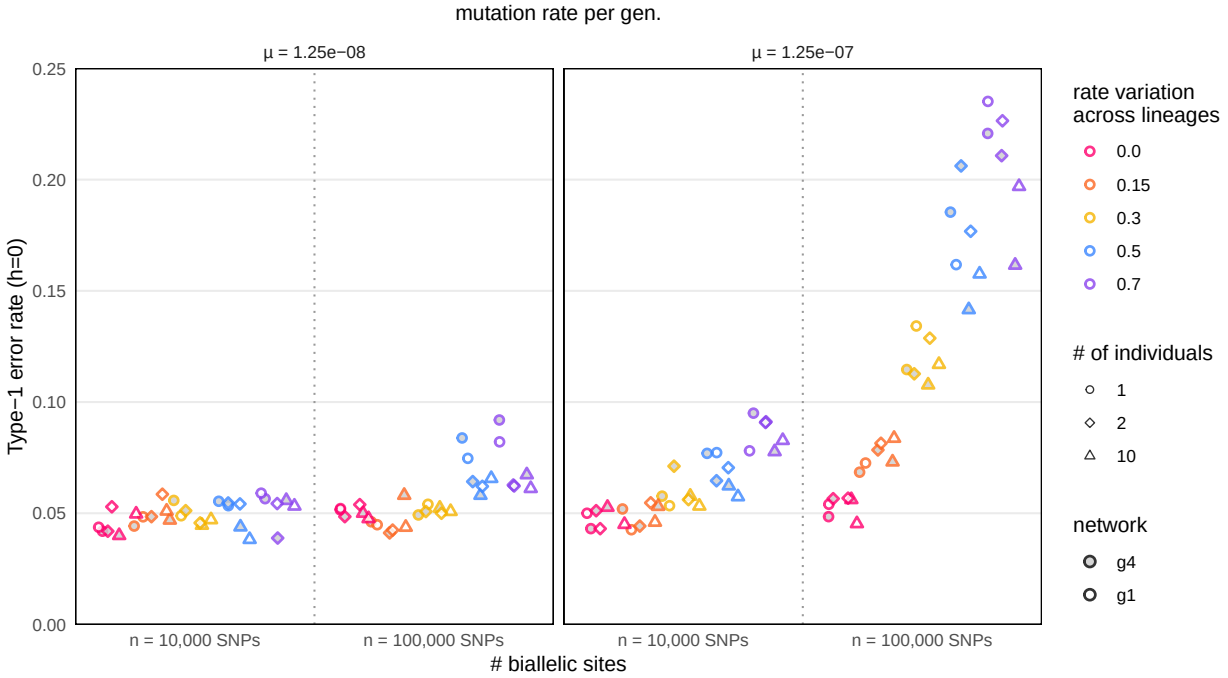

Figure 3 Type-1 error rate from f_4: proportion of quartets incorrectly inferred to have hybridization when none was present, depending on substitution rate ($\mu = 1.25 \times 10^{-8}$ or 1.25×10^{-7}) and data set size ($n = 10,000$ or $100,000$ biallelic sites). Each point represents the proportion of tests with a p-value below 0.05 across quartets from 100 replicate data sets, with 2,600 quartets per point from the g4 network, and 6,200 quartets per point from the g1 network. Shape color denotes the amount of rate variation (standard deviation σ, on a log-normal distribution with mean 1 as in Fig. S1), with pink representing no rate variation $\sigma = 0$, orange $\sigma = 0.15$, yellow $\sigma = 0.3$, blue $\sigma = 0.5$, and purple $\sigma = 0.7$. Shape represents number of individuals per population: 1 (circles), 2 (diamonds) or 10 (triangles). Shape fill indicates the network used to simulated data, either g1 (empty shapes) or g4 (filled shapes).

Power of f_4-tests to correctly infer hybridization

The power of f_4 to detect the presence of reticulation was most affected by the data set size, unsurprisingly. In all scenarios, mutation rate and rate variation across lineages had little effect. With $10,000$ biallelic sites, the presence of reticulation was detected between 5% (from 1 individual/population) to 25% of the time at most (with 10 individuals/population), regardless of the number of reticulations. Larger data sets ($100,000$ biallelic sites) had much more power to detect reticulation, especially with more individuals per population. However, power generally decreased with more reticulations. At 1 hybridization event within a 4-population subnetwork (quarnet), power averaged 87%. At 3 hybridization events within a quarnet, average power decreased to 56%.

Graph inference accuracy

We obtained a total of $40,538,966$ graphs, with sets of top graphs (1 set per replicate data set and search h value) ranging from 5 to $16,195$ distinct graphs.

As a primary measure of accuracy, we examined how many replicate data sets recovered the true network in their set of top graphs, when searching for the correct number of hybridizations (search $h = 1$ for g1, 4 for g4).

The less complex graph g1 was recovered for nearly all replicates, across rate variation levels and data set size when the mutation rate was slow ($\mu = 1.25 \times 10^{-8}$, Fig. 4). Under a faster mutation rate, rate variation across lineages started to affect accuracy. Particularly on large data sets ($100,000$ biallelic sites) and under a fast mutation rate ($\mu = 1.25 \times 10^{-7}$), there was a large gap in accuracy between no rate variation (100% of replicates recovered the true network in their set of top graphs) and high rate variation (49%-68% accuracy, depending on the number of individuals per population) for g1 (Fig. 4).

However, g4 was almost never recovered in the inferred set of top graphs, regardless of other factors (Fig. 4): accuracy was 0% in 56 of the 60 parameter combinations, and 1% otherwise.

We then measured accuracy by requiring at least one top graph at hardwired-distance 0 from the true network. For g1 this is equivalent to requiring that one top graph is equal to g1. But for g4, this is a relaxed measure of accuracy. (Fig. S3 shows example graphs that differ from g4 but are at hardwired-distance 0 from it). Despite this relaxed measure, accuracy was largely the same, with a minuscule improvement in inferring g4 (Fig. S4). At best, 5% of replicates had a top graph at hardwired-distance 0 from g4.

Given that almost none of the top graphs were at hardwired-distance 0 from g4, we looked at the distance of the closest top graph to the true network, as a more informative measure of accuracy (with lower distances corresponding to better accuracy). Unsurprisingly, accuracy was best when searching for the correct h (Fig. 5 for g4, Fig. S5 for g1). Importantly, there was strong evidence that rate variation across lineages decreases accuracy (likelihood ratio test $p < 10^{-15}$; see Supplemental section 5), particularly at levels $\sigma \geq 0.3$. In other words, more rate variation increased the distance of the closest top graph to the true network.

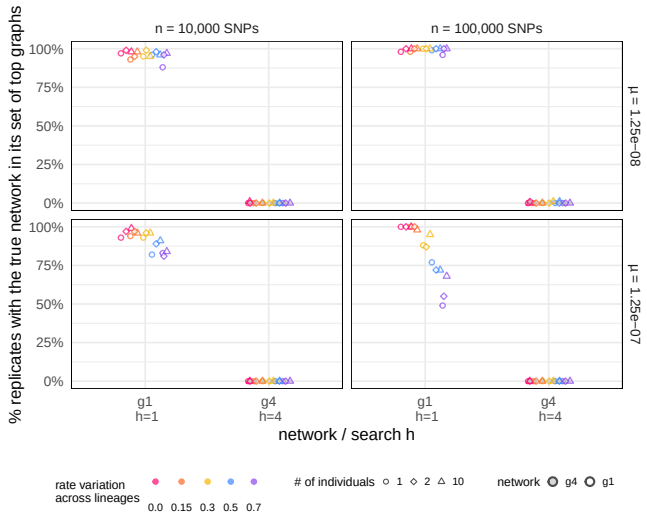

Figure 4 Percentage of replicate data sets whose set of top graphs includes the true network, when the search h is equal to the true number of reticulations (1 under g1, 4 under g4). Panels separate different substitution rates μ and number of biallelic sites n. Shape and color is as in Fig. 3.

Accuracy of graph features

As g4 was almost never correctly estimated, we assessed whether inferred top graphs contained graph features of the true network. We focused on two features: whether top graphs display the true network's major tree, and whether top graphs display any (or both) of the 2 near-major trees.

Display of the major tree Under g1, the major tree was recovered (displayed in at least one top graph) very frequently, generally (Fig. 6), even when searching for an underfitting $h = 0$ reticulations. When searching for trees only ($h = 0$), this metric essentially states whether the inferred tree was the major tree. When doing so, increasing the number of individuals per population had a major effect, increasing the ability to infer the major tree from as low as 25% (1 individual) to nearly 100% (10 individuals/population). Rate variation across lineages had a substantial effect on the larger data set (100,000 biallelic sites) with a fast mutation rate ($\mu = 1.25 \times 10^{-7}$), with increased lineage-rate variation decreasing the ability to display the major tree, especially at search $h \geq 1$.

Under the more complex network of g4, the major tree was almost always recovered when searching for the correct number of 4 reticulations (Fig. 7). This high accuracy measure is in stark contrast with the low accuracy to infer the full network. Even when searching for one reticulation, the major tree was recovered fairly frequently from data sets of 10,000 biallelic sites (51%-95%), but less frequently from data of 100,000 biallelic sites (0%-40%). Surprisingly, more individuals lead to a lower recovery of g4's major tree when searching for graphs with at most 1 reticulations (Fig. 7). This may be, in part, because data with more individuals tended to have fewer top graphs than data with few individuals (see Table S2), and keeping more top graphs increases the chance that one of them displays the major tree. This reduction in the number of top graphs can be the result of our filtering criterion (keep graphs of score $S > \hat{S} - 10$) and the tendency of larger data sets to have larger score differences between candidate networks. When searching for trees ($h = 0$), the true major tree was recovered 0%-54% of replicates, with more individuals per populations greatly improving performance (in this case, 5 top graphs were kept almost always).

Unlike for g1, the effect of rate variation was limited compared to the effect of other factors.

When the major tree is not recovered, we can look at the best hardwired-distance between the major tree and trees displayed by top graphs (Figs. S6 and S7). This distance is almost always at its minimum (2), showing that when the major tree is not recovered, a very similar tree is displayed by top graphs.

Display of the two near-major trees The accuracy of displaying both near-major trees showed the same pattern as that of displaying the major tree (Figs. S8 and S9). One exception was when constraining top graphs to be trees (search $h = 0$). In that case, a top graph may display (be equal to) the true major tree, but may never display both near-major trees, clearly. In short, both near-major trees were recovered with high accuracy under g1 and with good-to-high accuracy under g4 (depending on the search h); and rate variation across lineages decreased accuracy in several conditions.

With a search h equal to or above the true h, if one top graph displayed the major tree, then some top graph was very likely to actually display both near-major trees (Figs S10 and S11). With a search $h = 1$ and under g4, many data sets have their top graphs display at most 1 near-major tree, not both (Fig. S11).

Accuracy to select the number of reticulations

Based on the worst residual criterion to select the number of reticulations h, rate variation across lineages and number of biallelic sites both affected the ability to correctly choose $h = 1$ for g1. With smaller data sets (10,000 biallelic sites) and low levels of rate variation, "$h \leq 1$" was rejected in about 3% of data sets (Fig. 8, $\sigma \leq 0.3$ in pink, orange and yellow). For the smaller data sets, this type-1 error rate increased beyond the traditionally tolerated 5% when rate varied more across lineages and the overall mutation rate was fast (1.25×10^{-7})

In the absence of rate variation across lineages, the rate at which $h \leq 1$ was rejected was greater for bigger data sets (100,000 biallelic sites) than smaller data sets, between 5-10%. This type-1 error rate was strongly affected by rate variation across lineage for these bigger data sets, especially when the average mutation rate was fast. At its worst, $h = 1$ was rejected for almost $\sim$ 50% data sets. In other words, under strong rate variation across lineages and a fast mutation rate, empirical practitioners with large data sets would have a 50% chance of wrongly rejecting the true $h = 1$ in favor of $h \geq 2$, when using $|\text{WR}| \leq 3$ as a criterion of adequate fit. In all cases, $h \geq 2$ could be considered adequate, as there was always at least $h = 2$ graph with $|\text{WR}| \leq 3$ per replicate.

Under g4 and when allowing for up to 4 reticulations, one top graph had $|\text{WR}| \leq 3$ always, such that "$h \leq 4$" was correctly inferred to fit the data adequately. While this is excellent model selection accuracy, it is paired with very large sets of top graphs that almost never contain the true network, raising doubts about model overfitting within the class of 4-reticulation networks.

Score difference

The difference between the scores of the true network and best-scoring graph was always $\Delta S \geq 0$ when searching for graphs

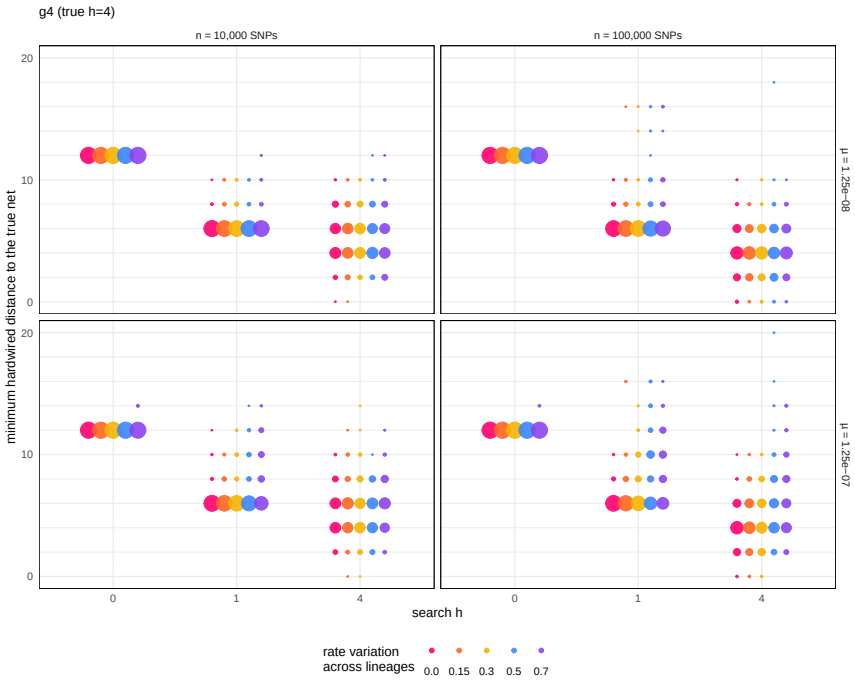

Figure 5 Distribution of the hardwired-distance between the true network and the closest top graph (that is, minimum hardwired-distance to the true network across top graphs), for data sets simulated under `g4`. The area of each circle is proportional to the percentage of replicate data sets, among the 300 data sets with a given combination of substitution rate μ and number of biallelic sites n. Color denotes the amount of rate variation as in Fig. 3.

with up to the true number of reticulations or more, suggesting a sufficiently thorough search of the graph space (Figs. S12 and S13).

Rate variation across lineages significantly increased this score gap ΔS. This may negatively impact model selection for methods based on likelihood scores such as AIC or BIC, as best-scoring graphs may be wrongly interpreted as fitting the data significantly better than the true network. Future work could investigate this further.

Discussion

Identifiability of complex graphs from f-statistics

One striking result from our study is the gap in accuracy between the network we considered. Our network `g1` with 1 reticulation was estimated with high accuracy from f_4 data, even in the presence of rate variation across lineages despite the sensitivity of individual f_4 tests to detect admixture. In contrast, the network `g4` with 4 reticulations was recovered in the set of top graphs with close to 0% accuracy, across all conditions. We propose that this gap is due, at least in part, to differences in theoretical identifiability.

We are unaware of any theoretical study of topology identifiability from f-statistics. Maier *et al.* (2023) proposed a numerical tool to determine the number of free edge parameters on a given network topology, and stated that the network root is not identifiable. Consequently, inference from f statistics should aim to estimate a *semidirected network*, rather than a rooted network (Solís-Lemus and Ané 2016; Xu and Ané 2023). Yet, there is a lack of graph theory to characterize outgroup constraints on the rooted network that do not overconstrain the semidirected network.

However, there is growing knowledge about network identifiability from other data types (Solís-Lemus and Ané 2016; Baños 2019; Gross *et al.* 2021; Allman *et al.* 2022, 2023; Xu and Ané 2023; Allman *et al.* 2024; Rhodes *et al.* 2025). Most similar to f-statistics are average genetic distances, because f_2 statistics match distances on trees (and f_3, f_4 values are linear combinations of f_2 statistics). As from f-statistics, the lengths of the 3 edges incident to a hybrid node are not identifiable from each other, based on average distances and under a model without incomplete lineage sorting (Xu and Ané 2023). This local lack of edge length identifiability is the reason to constrain hybrid edge lengths to 0 in `qpgraph`. Xu and Ané (2023) proved that many local topological features (subgraphs) are not distinguishable from one another from average distances. For example, a reticulation in a (undirected) cycle of 4 edges can perfectly reproduce the expected distance data that are expected from various subgraphs in which the cycle is modified with additional reticulations. If such features also lack identifiability from f_4 statistics, then one would expect an exponential number of graphs to fit a given data set, all with a similar likelihood score, when searching for top-scoring graphs with an increasing allowed number of reticulations. This is indeed what we observed in our study.

On the positive side, studies about network identifiability found that the network's *tree of blobs* is identifiable from many data types (Allman *et al.* 2023; Xu and Ané 2023; Rhodes *et al.* 2025). A *blob* is a maximal subgraph such that deleting any edge in it does not disconnect its nodes (when considering edges as undirected) (Gambette *et al.* 2012; Xu and Ané 2023). In `g1` for example (Fig. 2, the cycle of 4 edges created by the reticulation forms a blob. In `g4`, all reticulations are part of the same blob. The tree of blobs is then obtained from the network by contracting each blob into a single node. Doing so leads to a tree, whose

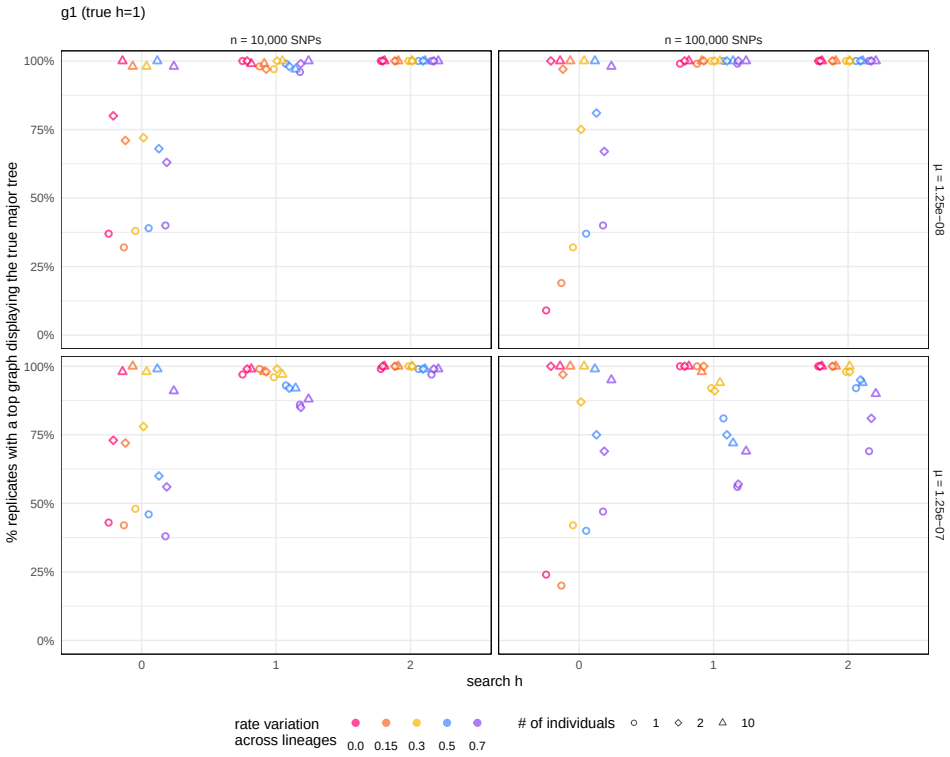

Figure 6 Percentage of 100 replicate data sets simulated under `g1` that have at least one top graph displaying the true major tree, depending on substitution rate μ and data set size. Top graphs were obtained by searching for high-likelihood graphs with at most "search h" reticulations. Shape and color is as in Fig. 3.

edges are the edges in the original network that, when removed, disconnect the graph. For `g4` for example, the tree of blobs is almost reduced to a star tree, except for a non-African clade. In contrast, `g1`'s tree of blobs retains the modern human subtree and the Neanderthal clade.

Also identifiable from many data types are networks of level 1 without "small" cycles (with "small" depending on the data type, e.g. with at most 3 edges) (Baños 2019; Gross *et al.* 2021; Xu and Ané 2023; Allman *et al.* 2024). The *level* of a network is the maximum number of reticulations within a blob. In a level-1 network distinct reticulations are well separated, in cycles that do not overlap. Any network with 1 reticulation, like `g1`, is of level 1. If level-1 networks are identifiable from f statistics as well, this may explain the accuracy of estimating `g1` in our study.

In `g4`, all 4 reticulations belong in the same blob, so `g4` is of level 4. Given that even level-2 networks are generally not identifiable from many data types Xu and Ané (2023); Rhodes *et al.* (2025), we suspect that identifiablity from f_4 statistics does not hold either. When multiple networks are not distinguishable from one another on the basis of f_4 data, they form a "terrace" of networks with the same score for a given data set, and cause an increase in the number of similarly high-scoring top graphs — as observed in our study.

Our findings suggest that the major tree is one graph feature that might be identifiable from f-statistics. However, we are unaware of any study exploring the identifiability of trees displayed by the network, regardless of data type and evolutionary model. The notion of the major tree requires that one of the two (or more) parents of each admixed population be designated as major and the other one(s) as minor, based on which parental population contributed most to the admixed population's genome. Identifiability of the major tree may be facilitated by an imbalance between parental contributions, to make the major parent easier to detect.

Theoretical work on identifiability from f-statistics is needed, and could borrow ideas and tools developed for other data types. Empirical analyses will benefit from knowledge of which network classes guarantee identifiability and which network features are identifiable. These features should be summarized across top-scoring graphs, as they are likely to be conserved and receive high support.

Recommendations for best practices

Our complex graph `g4` was almost never recovered from `find_graphs`: neither as the best-scoring graph nor in the set of top graphs, despite a quite lenient threshold for retaining graphs. With our threshold of keeping graphs within 10 log-likelihood points of the best-scoring graph, we kept a median of $1,868 - 6,760$ graphs per replicate when searching for 4 hybridization events. Thus, the recommendation from Maier *et al.* (2023) to review a set of best scoring graphs is necessary when using `find_graphs` to infer complex topologies, but seems insufficient with h as low as 4. The adequate number of top-scoring graphs seems to be low when searching for trees ($h = 0$) and to increase drastically as more reticulations are allowed, but may depend on the complexity of true underlying network.

Many studies published graphs as or more complex than `g4`, inferred using `admixtools` or `AdmixtureBayes` (for example, Fig. 3 in Lazaridis *et al.* 2014, $h = 4$, level 4; Fig. 3 in Moreno-Mayar *et al.* 2018, $h = 6$, level 6; Fig. 5 in Narasimhan *et al.* 2019,

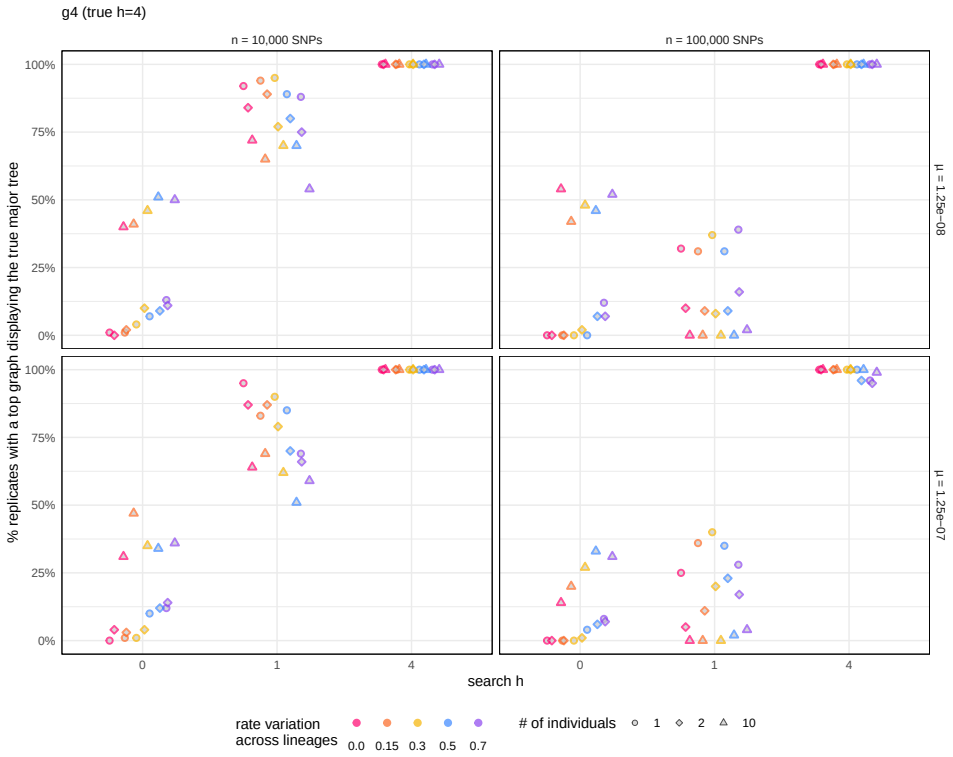

Figure 7 Percentage of 100 replicate data sets simulated under g4 that have at least one top graph displaying the true major tree, depending on substitution rate μ and data set size. Top graphs were obtained by searching for high-likelihood graphs with at most "search h" reticulations. Shape and color is as in Fig. 3.

$h = 12$, level 12). Our results should discourage the assumption that keeping a large set (e.g. thousands) of graphs ensures that the true network is included, especially if inferred graphs are of high level. Though our study did not assess the accuracy of inferring networks with a true $h = 2$ or $h = 3$, even $h \geq 2$ may present a problem in recovering the true network, depending on the size of its cycles and if / how they intersect. Like Maier *et al.* (2023), we recommend summarizing features shared by top-scoring graphs, and report features that are stable across graphs.

Individually, f_4-tests were found to be error-prone, with type-1 error exacerbated by a combination of lineage-rate variation and increased mutation rate. This is most likely due to homoplasies (Frankel and Ané 2023; Koppetsch *et al.* 2023). While graph inference is more robust to rate variation, each f_4-statistic is associated with one residual. Like individual f_4-tests, the worst residual (worst-fitting f_4) and associated selection of the appropriate number of hybridizations is strongly affected by rate variation. We recommend assessing the extent of rate variation across lineages in empirical data sets, and using caution in interpreting f_4-tests and worst residuals if rate variation is found to be strong.

In their protocol, Maier *et al.* (2023) recommend using top graphs identified by fitting f_4 data as candidate networks to be evaluated by methods that use other data types, and that might be more computationally intensive. Our study suggests using these candidate networks with caution, as the true network may still be excluded from this candidate list. Candidate networks could be used to seed the search made by other methods that explore the network space, for example.

Limitations and future directions

Our simulations only considered two fixed networks, leaving it unclear where identifiablity might be lost between g1 and g4. Simulated networks (as opposed to fixed) using the birth-death-hybridization model (Justison *et al.* 2022, 2023) could prove useful in generating networks, from which to simulate data. To distinguish the effects of reticulation number versus network level on inference accuracy (and identifiability), one could consider level-1 networks with multiple (separated) reticulations.

While this work and others (Frankel and Ané 2023; Koppetsch *et al.* 2023) raise concerns when the molecular clock is violated, methods to quantify rate variation across lineages and across loci are underdeveloped in the presence of reticulation (but see Snir *et al.* 2012; Duchêne *et al.* 2014, 2024, in a tree context).

Selecting the appropriate number of reticulations is particularly important. For this, the worst residual is commonly used in the admixture graph literature. Given the sensitivity of this criterion to rate variation, more robust methods are needed to select the graph complexity.

Finally, there is a lack of methods to summarize a set of networks. Such methods are needed to automatically extract features shared by a large number of high-scoring graphs, as recommended by Maier *et al.* (2023). Some concepts of consensus networks have been developed (Huber *et al.* 2023, for 1-nested networks), but more methods are needed to summarize various network features, especially those known to be identifiable and recoverable with high support.

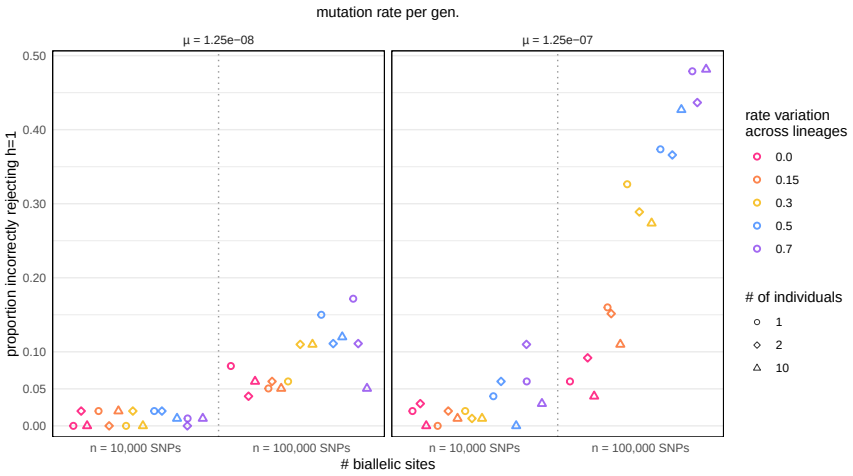

Figure 8 Type-1 error rate from worst residuals: proportion of 100 replicate data sets simulated under `g1` incorrectly rejecting $h \leq 1$. For each data set, top graphs were obtained by searching for high-likelihood graphs with at most 1 reticulation. If all these top graphs had a worst residual $|\text{WR}| > 3$ then $h \leq 1$ was rejected. Shape and color is as in Fig. 3.

Data availability

The code to reproduce all analyses and figures is available at https://github.com/laufran/f4-ratevar. The Supplemental material is currently included at the end of this document, to provide easier access for reviewers.

Funding

This work was supported by the National Science Foundation [DMS-1902892, 2023239 to C.A.] and [DGE 2137424 to L.E.F.].

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

Abu-Elmakarem H, MacLean OA, Venter F, Plenderleith LJ, Culleton RL, Hahn BH, Sharp PM. 2024. Remarkable evolutionary rate variations among lineages and among genome compartments in malaria parasites of mammals. Molecular Biology and Evolution. 41:msae243.

Allman ES, Baños H, Garrote-Lopez M, Rhodes JA. 2024. Identifiability of level-1 species networks from gene tree quartets. Bulletin of Mathematical Biology. 86.

Allman ES, Baños H, Mitchell JD, Rhodes JA. 2023. The tree of blobs of a species network: Identifiability under the coalescent. Journal of Mathematical Biology. 86:10.

Allman ES, Baños H, Rhodes JA. 2022. Identifiability of species network topologies from genomic sequences using the logdet distance. Journal of Mathematical Biology. 84:35.

Baños H. 2019. Identifying species network features from gene tree quartets under the coalescent model. Bulletin of Mathematical Biology. 81:494–534.

Bezanson J, Edelman A, Karpinski S, Shah VB. 2017. Julia: A fresh approach to numerical computing. SIAM Review. 59:65–98.

Blischak PD, Chifman J, Wolfe AD, Kubatko LS. 2018. HyDe: A python package for genome-scale hybridization detection. Systematic Biology. 67:821–829.

Coll Macià M, Skov L, Peter BM, Schierup MH. 2021. Different historical generation intervals in human populations inferred from neanderthal fragment lengths and mutation signatures. Nature Communications. 12:5317.

Degnan JH. 2018. Modeling hybridization under the network multispecies coalescent. Systematic Biology. 67:786–799.

Duchêne DA, Duchêne S, Stiller J, Heller R, Ho SYW. 2024. ClockstaRX: Testing molecular clock hypotheses with genomic data. Genome Biology and Evolution. 16:evae064.

Duchêne S, Molak M, Ho SYW. 2014. ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis. Bioinformatics. 30:1017–1019.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Molecular Biology and Evolution. 28:2239–2252.

Flegontov P, Altınışık NE, Changmai P, Rohland N, Mallick S, Adamski N, Bolnick DA, Broomandkhoshbacht N, Candilio F, Culleton BJ *et al*. 2019. Palaeo-eskimo genetic ancestry and the peopling of chukotka and north america. Nature. 570:236–240.

Flegontov P, Işıldak U, Maier R, Yüncü E, Changmai P, Reich D. 2023. Modeling of African population history using f-statistics can be highly biased and is not addressed by previously suggested SNP ascertainment schemes. bioRxiv. pp. 2023–01.

Fogg J, Allman ES, Ané C. 2023. PhyloCoalSimulations: a simulator for network multispecies coalescent models, including a new extension for the inheritance of gene flow. Systematic Biology. 72:1171–1179.

Fogg J, Ané C. 2022. PhyloCoalSimulations v0.1.0. https://github.com/cecileane/PhyloCoalSimulations.jl. Last accessed: 2023-01-17.

Frankel LE, Ané C. 2023. Summary Tests of Introgression Are Highly Sensitive to Rate Variation Across Lineages. Systematic Biology. p. syad056.

Gambette P, Berry V, Paul C. 2012. Quartets and unrooted phylogenetic networks. Journal of Bioinformatics and Computational Biology. 10:1250004.

Gautier M, Vitalis R, Flori L, Estoup A. 2022. f-statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package poolfstat. Molecular Ecology Resources. 22:1394–1416.

Ge X, Lu Y, Chen S, Gao Y, Ma L, Liu L, Liu J, Ma X, Kang L, Xu S. 2023. Genetic origins and adaptive evolution of the deng people on the tibetan plateau. Molecular Biology and Evolution. 40:msad205.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome biology and evolution. 4:658–667.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY *et al*. 2010. A draft sequence of the Neandertal genome. Science. 328:710–722.

Gross E, van Iersel L, Janssen R, Jones M, Long C, Murakami Y. 2021. Distinguishing level-1 phylogenetic networks on the basis of data generated by Markov processes. Journal of Mathematical Biology. 83:32.

Gutaker RM, Groen SC, Bellis ES, Choi JY, Pires IS, Bocinsky RK, Slayton ER, Wilkins O, Castillo CC, Negrão S *et al*. 2020. Genomic history and ecology of the geographic spread of rice. Nature Plants. 6:492–502.

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. 522:207–211.

Hahn MW, Hibbins MS. 2019. A three-sample test for introgression. Molecular Biology and Evolution. 36:2878–2882.

Harney É, Patterson N, Reich D, Wakeley J. 2021. Assessing the performance of qpadm: a statistical tool for studying population admixture. Genetics. 217:iyaa045.

Huber KT, Moulton V, Spillner A. 2023. Computing consensus networks for collections of 1-nested phylogenetic networks. Journal of Graph Algorithms and Applications. 27:541–563.

Huson DH, Rupp R, Scornavacca C. 2010. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press. Cambridge.

Justison J, Solís-Lemus C, Heath TA. 2022. *SiPhyNetwork: A Phylogenetic Simulator for Reticulate Evolution*. R package version 1.0.0.

Justison JA, Solís-Lemus C, Heath TA. 2023. SiPhyNetwork: An R package for simulating phylogenetic networks. Methods in Ecology and Evolution. .

Koppetsch T, Malinsky M, Matschiner M. 2023. Among-species rate variation produces false signals of introgression. bioRxiv. pp. 2023–05.

Lanfear R, Ho SY, Jonathan Davies T, Moles AT, Aarssen L, Swenson NG, Warman L, Zanne AE, Allen AP. 2013. Taller plants have lower rates of molecular evolution. Nature Communications. 4:1879.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M *et al*. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 513:409–413.

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. 598:634–640.

Lipson M. 2020. Applying f_4-statistics and admixture graphs: Theory and examples. Molecular Ecology Resources. 20:1658–1667.

Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, Pryce TO, Willis A, Matsumura H, Buckley H *et al*. 2018. Ancient genomes document multiple waves of migration in southeast asian prehistory. Science. 361:92–95.

Maier R, Flegontov P, Flegontova O, Isildak U, Changmai P, Reich D. 2023. On the limits of fitting complex models of population history to f-statistics. Elife. 12:e85492.

Martin AP, Palumbi SR. 1993. Body size, metabolic rate, generation time, and the molecular clock. Proceedings of the National Academy of Sciences. 90:4087–4091.

Massimo A, Le Trang, Corrado C, Alessandra B, June C. 2024. openalexr: An r-tool for collecting bibliometric data from openalex. The R Journal. 15:167–180.

Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas AS, Sikora M *et al*. 2018. Terminal pleistocene alaskan genome reveals first founding population of native americans. Nature. 553:203–207.

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:eaat7487.

Nielsen SV, Vaughn AH, Leppälä K, Landis MJ, Mailund T, Nielsen R. 2023. Bayesian inference of admixture graphs on native american and arctic populations. PLOS Genetics. 19:1–22.

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. 192:1065–1093.

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

Peter BM. 2022. A geometric relationship of f_2, f_3 and f_4-statistics with principal component analysis. Philosophical Transactions of the Royal Society B: Biological Sciences. 377:20200413.

Petit N, Barbadilla A. 2009. Selection efficiency and effective population size in drosophila species. Journal of Evolutionary Biology. 22:515–526.

Priem J, Piwowar H, Orr R. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts.

Raftery AE. 1995. Bayesian model selection in social research. Sociological Methodology. 25:111–163.

Ragsdale AP, Weaver TD, Atkinson EG, Hoal EG, Möller M, Henn BM, Gravel S. 2023. A weakly structured stem for human origins in africa. Nature. 617:755–763.

Rambaut A, Grass NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics. 13:235–238.

Rhodes JA, Baños H, Xu J, Ané C. 2025. Identifying circular orders for blobs in phylogenetic networks. Advances in Applied Mathematics. 163:102804.

Snir S, Wolf YI, Koonin EV. 2012. Universal pacemaker of genome evolution. PLOS Computational Biology. 8:1–9.

Solís-Lemus C, Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genetics. 12:e1005896.

Soraggi S, Wiuf C. 2019. General theory for stochastic admixture graphs and f-statistics. Theoretical Population Biology. 125:56–66.

Thomas JA, Welch JJ, Lanfear R, Bromham L. 2010. A generation time effect on the rate of molecular evolution in invertebrates. Molecular biology and evolution. 27:1173–1180.

van der Valk T, Pečnerová P, Díez-del Molino D, Bergström A, Oppenheimer J, Hartmann S, Xenikoudakis G, Thomas JA, Dehasque M, Sağlıcan E *et al*. 2021. Million-year-old DNA

sheds light on the genomic history of mammoths. Nature. 591:265–269.

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. 591:413–419.

Wang RJ, Al-Saffar SI, Rogers J, Hahn MW. 2023. Human generation times across the past 250,000 years. Science Advances. 9:eabm7047.

Wang Y, Obbard DJ. 2023. Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis. Evolution Letters. 7:216–226.

Xu J, Ané C. 2023. Identifiability of local and global features of phylogenetic networks from average distances. Journal of Mathematical Biology. 86:12.

Zecca G, Labra M, Grassi F. 2020. Untangling the evolution of american wild grapes: Admixed species and how to find them. Frontiers in Plant Science. 10:1814.

Zhao X, Guo Y, Kang L, Yin C, Bi A, Xu D, Zhang Z, Zhang J, Yang X, Xu J *et al*. 2023. Population genomics unravels the holocene history of bread wheat and its relatives. Nature Plants. 9:403–419.

Supplemental material

1. Distribution of rates across population lineages

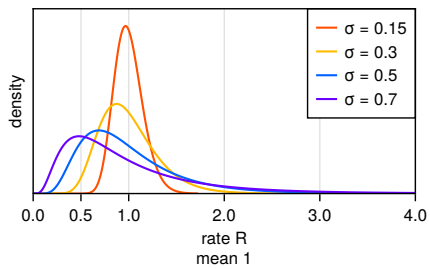

Figure S1 Density of the log-normal distribution used for simulating rates R, with standard deviation σ of 0.15, 0.3, 0.5 and 0.7 on the log-scale. To set the mean of R to 1, the mean of $\log(R)$ was set to $-\sigma^2/2$.

2. Power of the f_4-statistic to detect reticulations

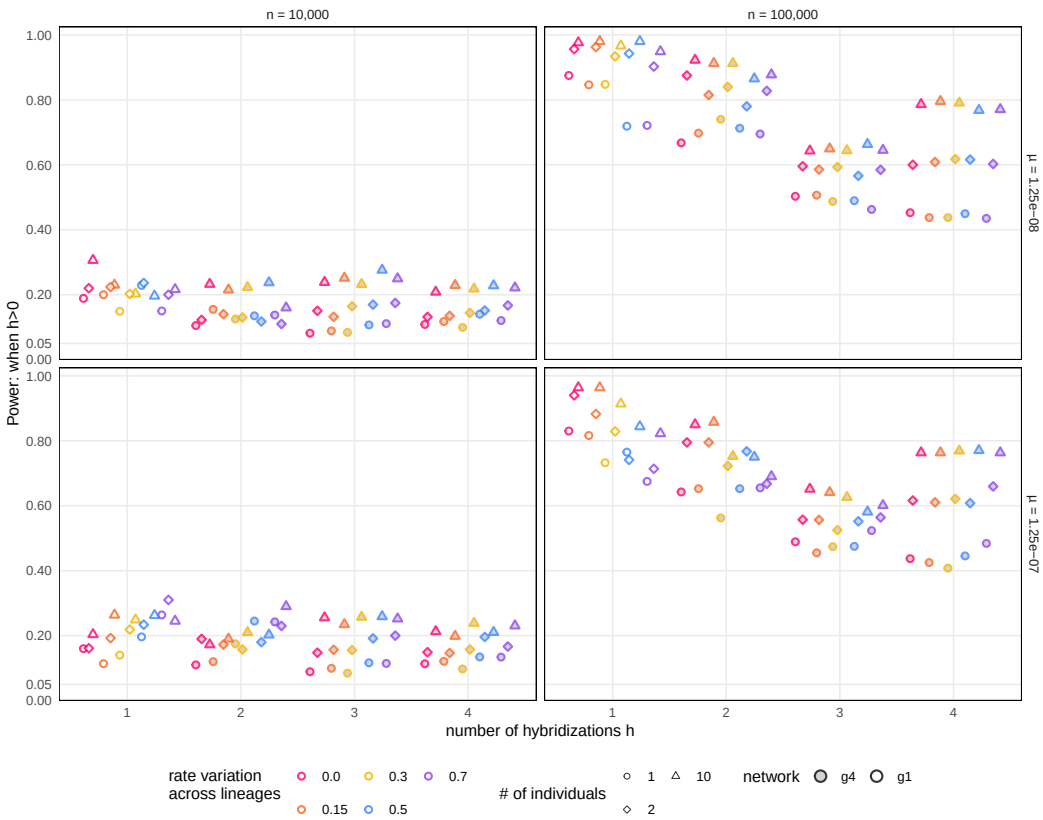

Figure S2 Power of the f_4 statistic to detect reticulations in a given quartet. depending on substitution rate μ and number of biallelic sites n. Shape color, type, and fill are as in Fig. 3. Note that subsets of 4 populations whose subnetwork retains a reticulation have $h = 1$ from g1 (which only has 1 reticulation in the entire network) and $h \geq 2$ from g4.

3. Complexity of graph search using `find_graphs`

We summarized the time needed to estimate admixture graphs using `find_graphs`, using the search parameters as specified in the Methods, for one combination of simulation parameters: across the 100 data sets simulated under `g4` (because searching for 4 admixture events is more time intensive than the maximum of 2 events with `g1`), 10 individuals per population, $n = 100,000$ biallelic sites, no rate variation across lineages ($\sigma = 0.0$), and a low substitution rate of 1.25×10^{-8}.

	elapsed time per run (minutes)			total for 1 parameter combination
search h	average	minimum	maximum	(100 replicates $\times$ 50 runs)
0	2.47	0.779	5.47	8.57 days
1	11.8	1.49	69.2	41.10 days
4	122	2.67	697	424.87 days

Table S1 Time needed to run `find_graphs` from one set of simulation parameters. Our analysis includes 50 independent runs for each of 100 data sets and each search h. Thus, each row summarizes 5,000 runs. The total time across all runs, search h values, and data sets was 474.54 days.

To ease the computational burden of running `find_graphs` across 60 combinations of simulation parameters, we ran these analyses on 13 clusters with a total of 652 threads available for parallelization. Despite this parallelization of computations, it took over 3 months (clock time) to run all these analyses. Of note, the computing time reported in Table S1 does not include the time required to calculate worst residuals, as those were calculated separately (using `qpgraph`, after determining the sets of top graphs using `find_graphs`.

A total of 40,538,966 graphs were saved across all `find_graphs` runs (1 set per replicate data set and search h value). The size of top graphs varied substantially across parameter settings, as shown below.

search	# ind	# biallelic	data generated under `g1`				data generated under `g4`			
h	/pop	sites	minimum	median	average	maximum	minimum	median	average	maximum
0	1	10,000	5	5	5.01	9	5	5	5.43	15
0	1	100,000	5	5	5	5	5	5	5	5
0	2	10,000	5	5	5	5	5	5	5.02	15
0	2	100,000	5	5	5	5	5	5	5	5
0	10	10,000	5	5	5	5	5	5	5	5
0	10	100,000	5	5	5	5	5	5	5	5
1	1	10,000	5	144	176	428	5	166	189	957
1	1	100,000	5	11	32.5	133	5	8	12.0	139
1	2	10,000	5	133	157	350	5	130	139	574
1	2	100,000	5	10	24.4	129	5	6	7.64	37
1	10	10,000	5	129	141	351	5	56.5	89.5	359
1	10	100,000	5	9	23.3	132	5	5	5.40	17
2 (`g1`) or 4 (`g4`)	1	10,000	22	2,112	1,991	4,433	183	6,760	6,755	16,195
2 (`g1`) or 4 (`g4`)	1	100,000	5	635	658	2,637	29	2,398	2,363	6,235
2 (`g1`) or 4 (`g4`)	2	100,00	16	2,120	1,948	4,517	245	5,442	5,565	11,756
2 (`g1`) or 4 (`g4`)	2	100,000	5	634	620	2,636	5	2,024	1,955	5,569
2 (`g1`) or 4 (`g4`)	10	10,000	30	2,158	2,007	4,679	728	5,436	5,459	12,757
2 (`g1`) or 4 (`g4`)	10	100,000	5	641	641	3,012	28	1,868	1,854	4,229

Table S2 Number of top graphs saved, per replicate data set and search h parameter, after pooling graphs from the 50 independent runs and filtering to keep at least 5 graphs, and all those with a score $S > \hat{S} - 10$, where $\hat{S}$ was the score found across the 50 runs. Each row summarizes the number of top graphs across 500 replicate data sets (100 for each level of rate variation across sites).

4. Network topologies with identical hardwired clusters as g4

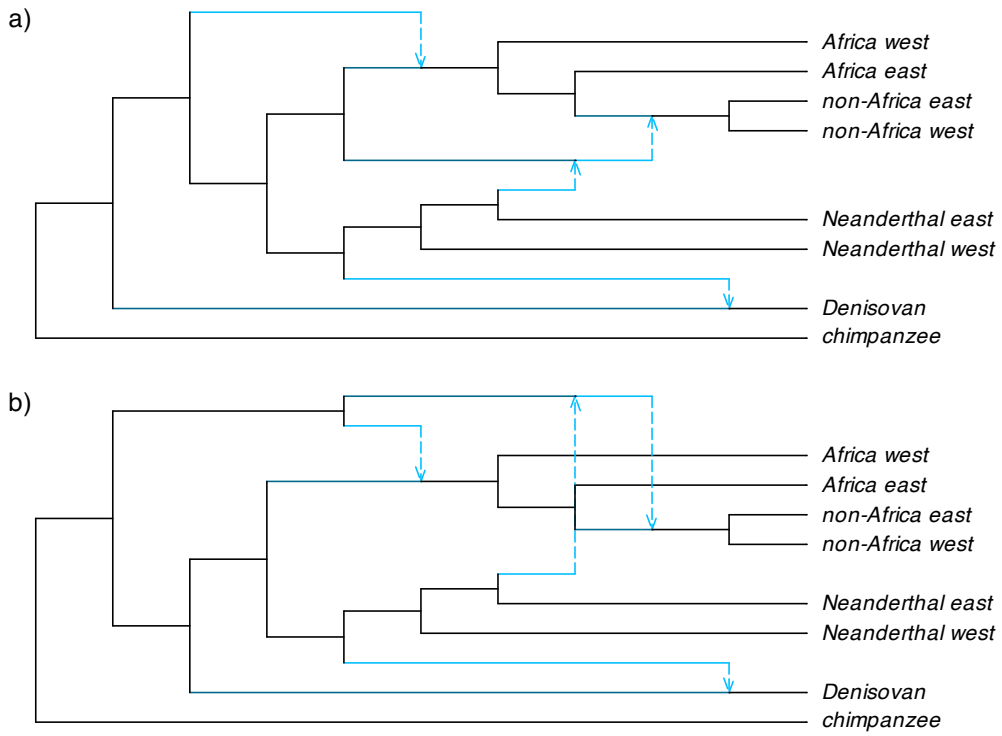

Figure S3 Example networks at hardwired-distance 0 from g4, yet different from g4. These examples illustrate why inferred networks may be at hardwired-distance 0 of g4 more frequently than they are identical to g4 (based on identical hashes). a) Network that differs from g4 by its order of the first two divergence events: the split leading to the Denisovan lineage (on the major tree) occurring before the split with the ghost population that contributed gene flow to the 4 modern human populations (instead of after in g4). b) Network that differs from g4 by the second ghost population (into nonafricans): stemming from the first ghost lineage instead of stemming from the human stem in g4.

5. How close are the top graphs from the true network?

We calculated the hardwired-distance $d(G, \mathtt{g})$ between each top graph G and the true network $\mathtt{g}$ (g1 or g4), and found the closest top graph, that is, the top graph minimizing this distance: $d_{\min} = \min\{d(G, \mathtt{g}) : G \text{ in top graph set}\}$.

Probability that the closest top graph is at distance 0 from the true network Fig. S4 shows the frequency with which the closest top graph was at distance 0 from the true network, given a search h equal to the true h. If G and $\mathtt{g}$ are trees or networks with 1 reticulation at most, then the hardwired-distance d is a proper distance: $G = \mathtt{g}$ exactly when $d(G, \mathtt{g}) = 0$, and this criterion captures accuracy of network inference exactly. But for networks with $h = 4$ reticulations, if is possible that $d(G, \mathtt{g}) = 0$ for $G \neq \mathtt{g}$ (see examples in Fig. S3). So for g4, the probability that at least top graph is at distance 0 from g4 is a relaxed (lenient) measure of accuracy. Still, even with this relaxation, g4 was recovered with extremely low accuracy (at most 5%).

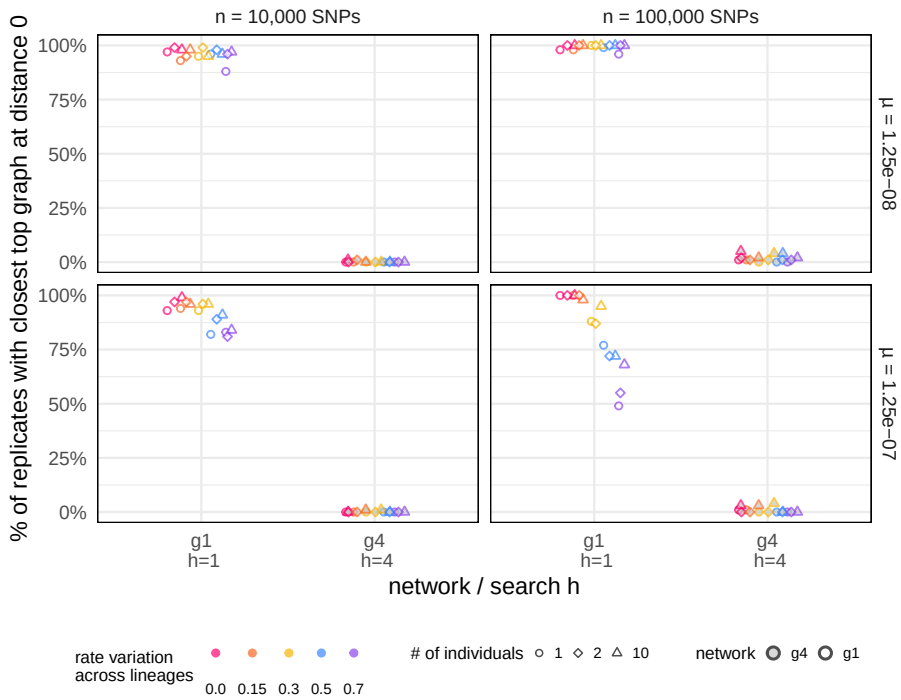

Figure S4 Percentage of 100 replicate data sets that have at least one top graph at hardwired-distance 0 from the true network (g1) or (g4), depending on substitution rate ($\mu = 1.25 \times 10^{-8}$ or 1.25×10^{-7}) and data set size ($n = 10,000$ or $100,000$ biallelic sites). Top graphs were obtained by searching for graphs with high log-likelihood scores with `find_graphs`, allowing for a maximum of "search h" reticulations. Shape and color is as in Fig. 3.

Distance of the closest top graph to the true network Fig. S5 (for data generated under g1) and Fig. 5 (for data generated under g4) show the distribution of the hardwired-distance of the closest top graph to the true network, $d_{\min} = \min\{d(G, \mathtt{g}) : G \text{ in top graph set}\}$ where g is the true network. To assess whether rate variation had an effect on this measure of accuracy, we conducted a Poisson regression on the transformed value $y = d_{\min}/2 - m$ where $m = 1$ for $\mathtt{g} = \mathtt{g1}$ if the search $h = 0$, $m = 6$ for $\mathtt{g} = \mathtt{g4}$ if the search $h = 0$, $m = 3$ for $\mathtt{g} = \mathtt{g4}$ if the search $h = 1$ and $m = 0$ otherwise (that is, if the search h is at least as large as the true h). This transformation was done to meet the Poisson distribution assumption for y, because $d_{\min}$ is a multiple of 2 (for resolved phylogenies), and because $d_{\min}/2$ is bounded by a minimum value when comparing phylogenies with different numbers of reticulation.

We used Poisson regression with our simulation parameters as predictors (all taken as factors with discrete levels): the true network, the search h (nested in network, as different h values were considered for g1 and g4), the mutation rate, the number of biallelic sites, the level of rate variation across lineages, and no interaction effects. The Poisson regression model fit the data adequately (using both the residual deviance test and the Pearson chi-square test). There was overwhelming evidence for an effect of rate variation (likelihood ratio test statistic $X^2 = 241$, $p < 2 \times 10^{-16}$). There was no discernible difference ($p = 0.10$) between no rate variation ($\sigma = 0$) and the lowest level of rate variation considered here ($\sigma = 0.15$) but strong evidence ($p = 0.001$) of an increase in the average distance between $\sigma = 0$ and $\sigma = 0.3$ and overwhelming evidence for $\sigma = 0.5$ and 0.7 ($p < 2 \times 10^{-16}$), with increasingly more rate variation leading to an increasing distance between the closest top graph and the true network.

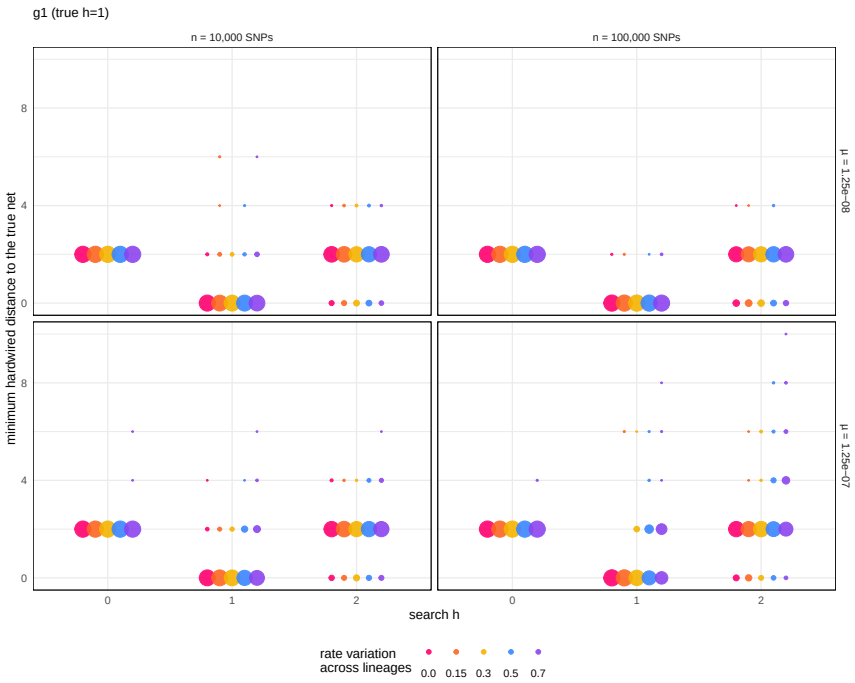

Figure S5 Distribution of the hardwired-distance between the true network and the closest top graph (that is, minimum hardwired-distance to the true network across top graphs), for data sets simulated under g1. The area of each circle is proportional to the percentage of replicate data sets, among data sets with a given combination of substitution rate μ, number of biallelic sites n, and maximum reticulation number h during graph search. Color denotes the amount of rate variation (the standard deviation, σ on a log-normal distribution with mean 1), with pink representing no rate variation $\sigma = 0$, orange $\sigma = 0.15$, yellow $\sigma = 0.3$, blue $\sigma = 0.5$, and purple $\sigma = 0.7$.

6. Do top graphs display the true major tree?

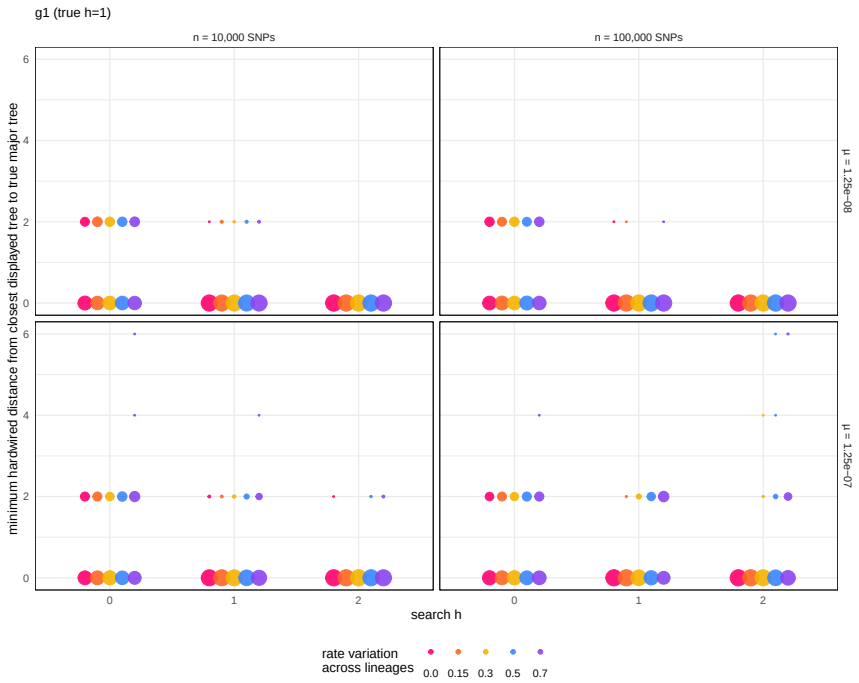

Figure S6 Distribution of the hardwired-distance between the true major tree and the closest top graphs' displayed tree (that is, minimum hardwired-distance to the true major tree across displayed trees from top graphs), for data sets simulated under g1. The area of each circle is proportional to the percentage of replicate data sets, among the 300 data sets with a given combination of substitution rate μ and number of biallelic sites n. Color denotes the amount of rate variation as in Fig. S5.

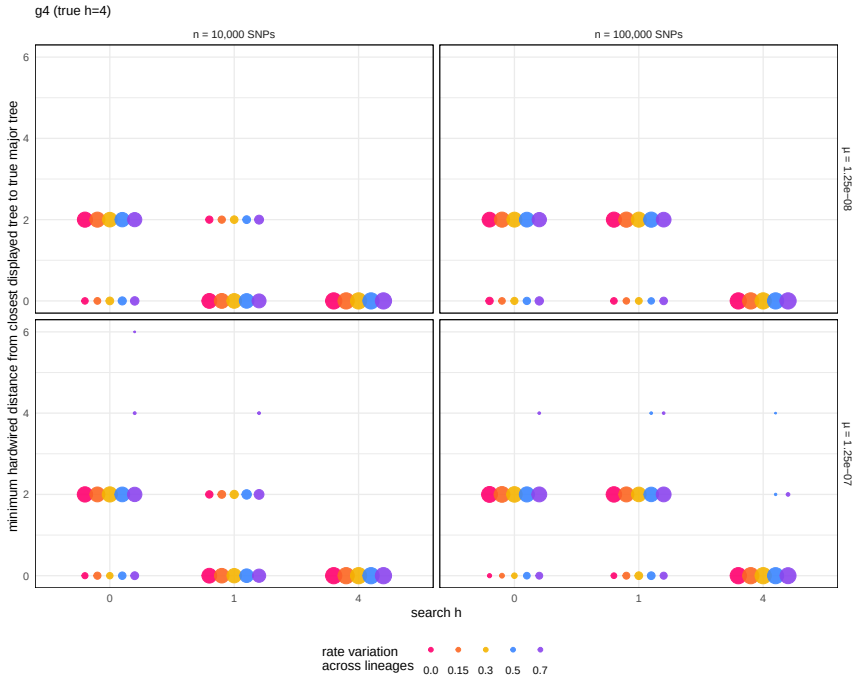

Figure S7 Hardwired-distance between the true major tree and the closest top graphs' displayed tree, as in Fig. S6, but for data sets simulated under g4.

7. Do top graphs display one or both near-major trees?

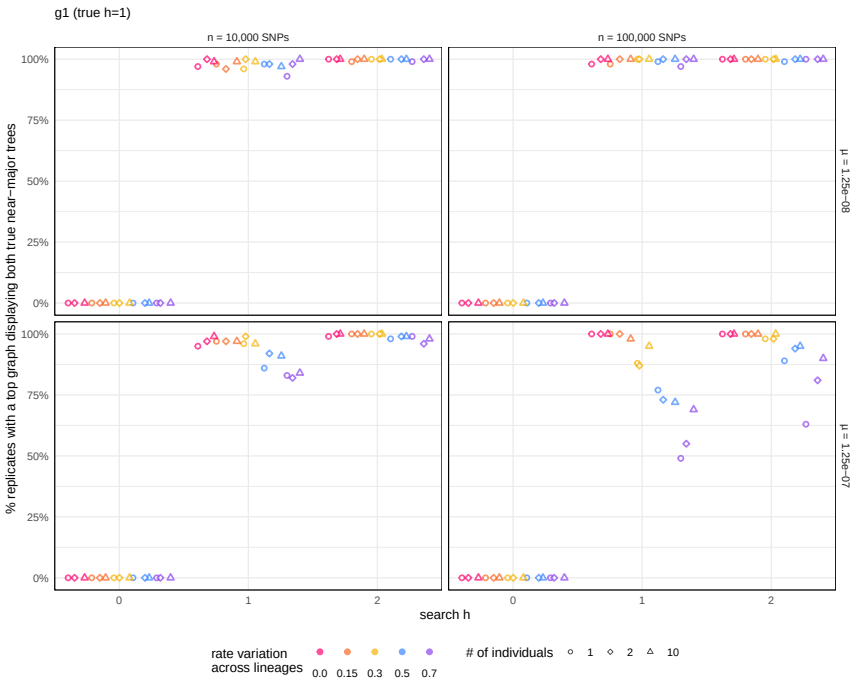

Figure S8 Percentage of 100 replicate data sets simulated under g1 that have at least one top graph displaying both true near-major trees, depending on substitution rates μ and data set size. Top graphs were obtained by searching for high-likelihood graphs with at most "search h" reticulations. Shape and color is as in Fig. 3.

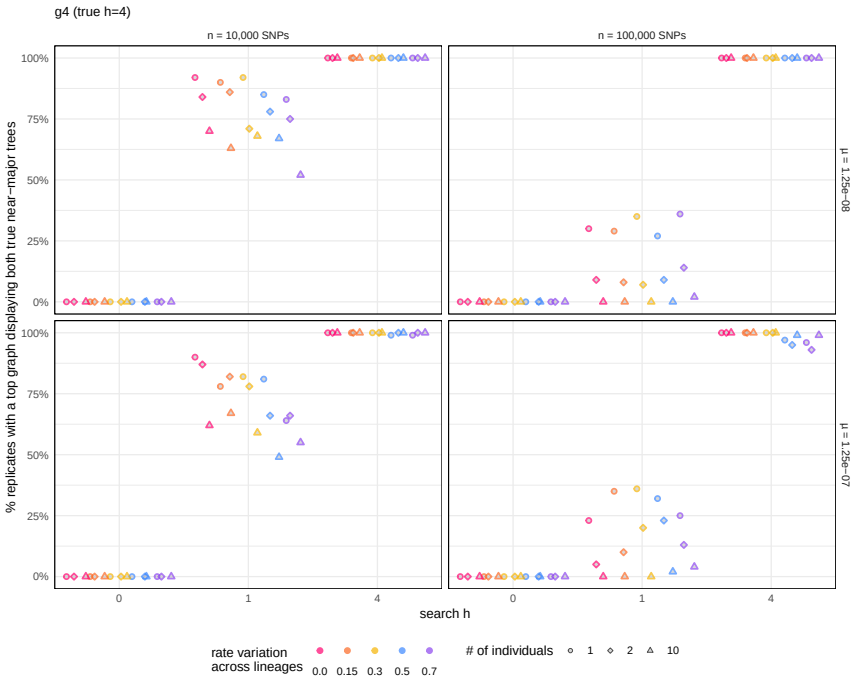

Figure S9 Percentage of data sets that have at least one top graph displaying both true near-major trees, as in Fig. S8, but for data sets simulated under g4.

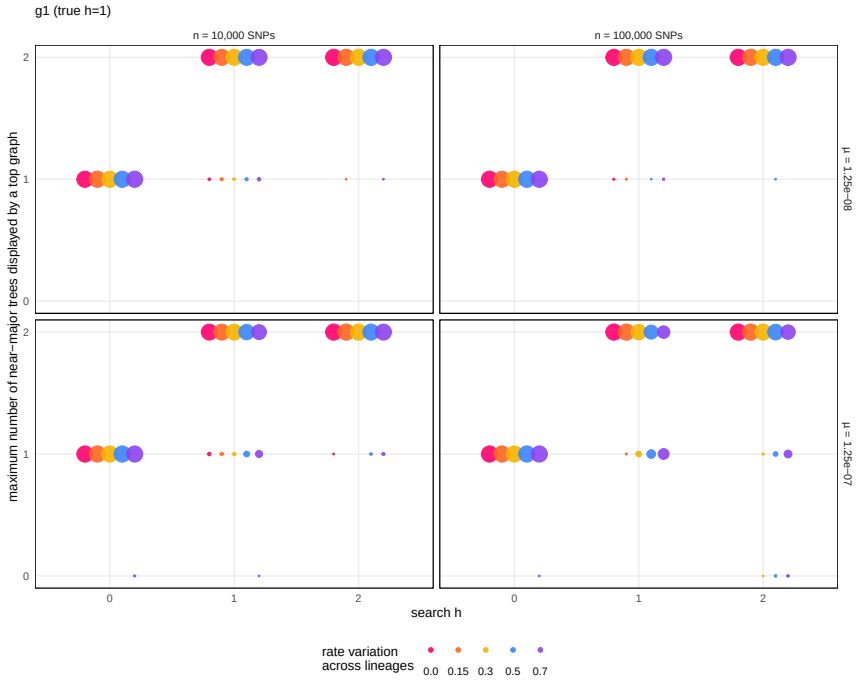

Figure S10 Maximum number of near-major trees displayed by a top graph (where the maximum is taken over all graphs in a replicate's set of top graphs), for data sets simulated under g1. The area of each circle is proportional to the percentage of replicate data sets, among data sets with a given combination of substitution rate μ, number of biallelic sites n, and maximum reticulation number h during graph search. Color is as in Fig. S5.

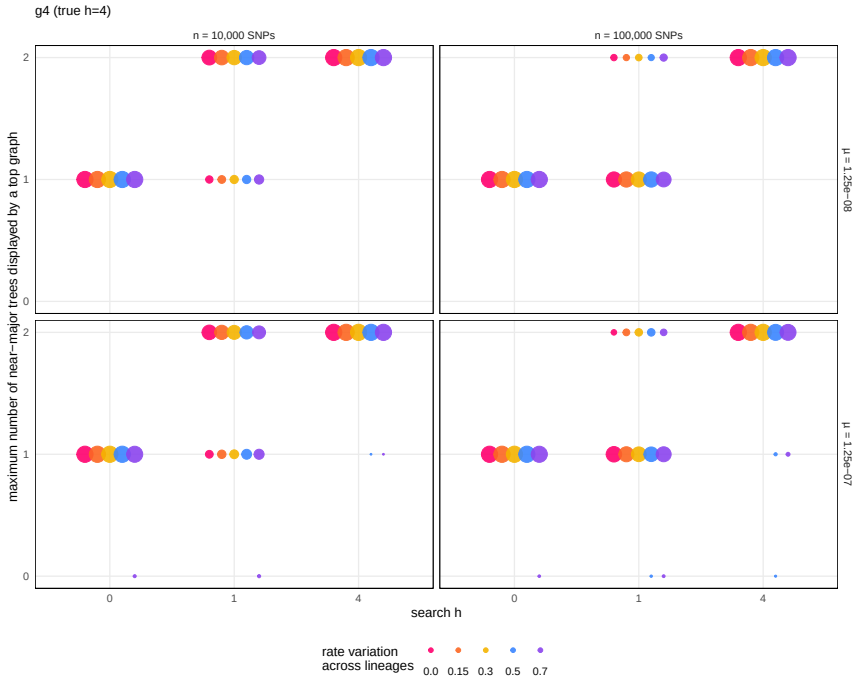

Figure S11 Maximum number of near-major trees displayed by a top graph (where the maximum is taken over all graphs in a replicate's set of top graphs) for data sets simulated under g4. Area and color are as in Fig. S10.

8. How thoroughly was the graph space explored?

As g4 was very rarely recovered among the set of top graphs, we might wonder if this was due to g4 having features that make it extremely difficult to recover (like many reticulations, some involving "ghost" lineages), or if this was due to poor exploration of the graph space. To ensure that our graph search was sufficiently thorough, we increased search parameters (see Methods) within limits of running time constraints (see Supplementary section 3 and Table S1), and we compared the score of the best-scoring top graph $\hat{S}$ with the score of the true network $S(\mathtt{g})$.

When searching for up to $h \geq 1$ reticulations under g1, or up to $h = 4$ reticulations under g4, we observed $\Delta S = S(\mathtt{g}) - \hat{S} \geq 0$ all the time, as expected theoretically from a complete search, suggesting a sufficiently thorough search of the graph space.

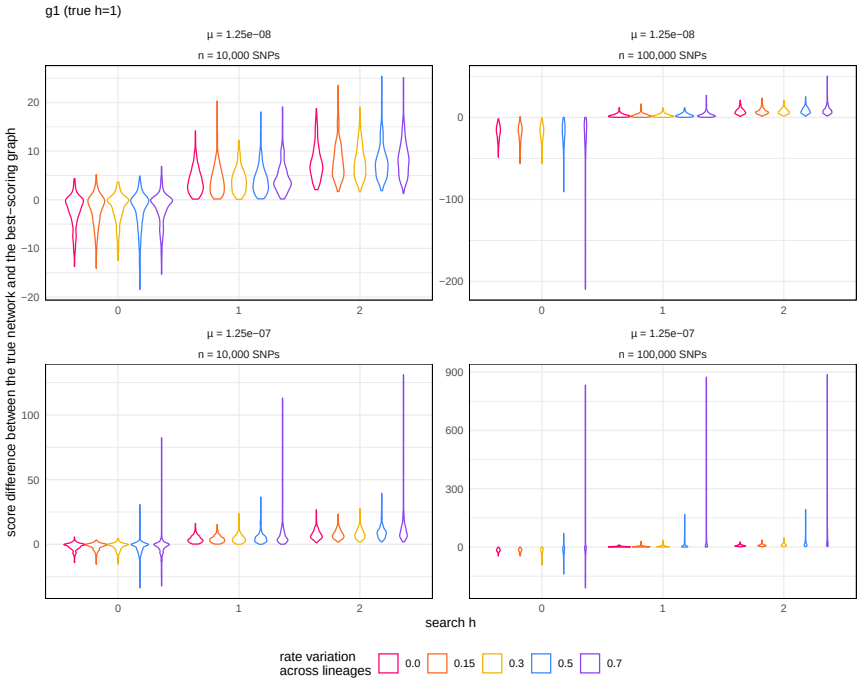

Figure S12 Distribution of the score difference between the true network and the best-scoring top graph, for data sets simulated under g1. Each violin represents the 300 replicate data sets with a given combination of substitution rate μ and number of biallelic sites n. Color denotes the amount of rate variation across lineages as in Fig. 3.

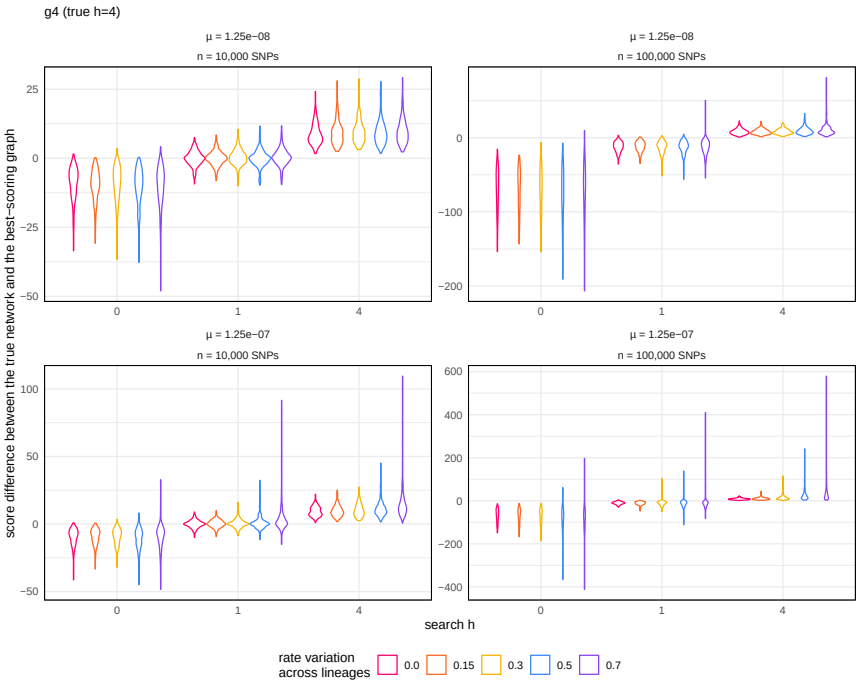

Figure S13 Score difference between the true network and the best-scoring top graph, as in Fig. S12, but for data sets simulated under g4.