LoGicAl: Local ancestry and genotype calling uncertainty-aware ancestry-specific allele frequency estimation from admixed samples

Jiongming Wang[1,2] and Sebastian Zöllner[1,2,3*]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.
[2]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA.
[3]Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109, USA.

*Corresponding author(s). E-mail(s): szoellne@umich.edu;
Contributing authors: jomiwang@umich.edu;

Abstract

In admixed groups, it is of interest to estimate allele frequencies of their ancestral contributing populations. These ancestry-specific allele frequencies inform genetic drivers of disease etiologies, facilitate genome-wide association study interpretations, enhance polygenic risk prediction and portability, and provide insights into demographic history. Their estimation leverages inferred locus-specific ancestry background, i.e., local ancestry, generated by tools like RFMix. However, existing estimation methods lose accuracy and incur biases by failing to model uncertainty from upstream local ancestry inference and genotype calling. Here, we introduce LoGicAl, a novel likelihood-based method for estimating ancestry-specific allele frequencies from admixed samples, simultaneously accommodating uncertainty from ancestry calling, genotyping, and statistical phasing. We demonstrate that modeling these uncertainties substantially reduces estimation errors, resulting in superior accuracy for both sequence-based and array-based genotyping with different levels of local ancestry inference quality. By integrating an accelerated fixed-point algorithm, LoGicAl achieves high scalability and enhanced computational efficiency compared with existing approaches. Applying LoGicAl to admixed cohorts in the 1000 Genomes Project, we illustrate the benefits of local-ancestry-based allele frequency estimates. Together, LoGicAl contributes to genomic analyses of admixed samples by providing precise and rapid ancestry-specific allele frequency estimates, and to constructing the spatial landscape and dynamics of genetic variations in admixed populations at a finer scale.

Keywords: Allele frequency; admixture; population structure; local ancestry; whole-genome sequencing; EM algorithm.

1 Introduction

Accurate allele frequency estimation is crucial to genetic analyses, providing insights into population demographic histories[1], informing population-enriched genetic drivers of disease etiologies[2, 3], facilitating genome-wide association study (GWAS) interpretations and replication cohort designs[4], as well as portability of polygenic risk scores (PRS)[5, 6] in both single-ancestry populations and admixed populations.

Interpretation of allele frequency estimates of admixed populations is complicated by their mosaic structure of genetic ancestral backgrounds[7]. Disregarding this structure when estimating allele frequencies treats genetic variation in current admixed population sample as the mean of the ancestral population allele frequencies weighted by their ancestral contributions. This creates challenges for interpreting admixed populations: since admixed groups can differ by ancestral contributions, allele frequency estimates are not transferable from one group to another. Additionally, it is often of interest to study the pre-admixture ancestral populations. To overcome these challenges, one can estimate the allele frequencies of the ancestral populations from the admixed sample that they contributed to, referred to as ancestry-specific allele frequencies[8].

This approach has been broadly used in genetic analyses of admixed populations. Ancestry-specific allele frequencies have been leveraged to identify ancestry-differentiated risk alleles for complex traits[9], such as asthma[10], blood pressure[11], and mild cognitive impairment[12]. In GWAS design, integrating information from ancestry-specific allele frequencies and matching them across corresponding ancestry cohorts can guide replication designs for validating discovered ancestry-enriched risk variants and improve polygenic prediction accuracy for admixed populations[13]. Moreover, when inferring population structure from admixed samples, incorporating reliable reference ancestry-specific allele frequencies can benefit statistical and computational performance of inference tools[14]. The performance of all these applications depends on the precision of the estimated ancestry-specific allele frequencies.

Several methods have been developed to deconstruct allele frequencies of the ancestral populations for admixed samples. Some methods primarily aim to infer population structure, characterized by admixture proportions (i.e., global ancestry)[15–20]. These methods usually apply mixture models to inform admixture proportion estimates while treating ancestry-specific allele frequencies as nuisance parameters. Another type of methods are explicitly designed to estimate ancestry-specific allele frequencies[21, 22]. These methods integrate information from locus-specific ancestry background (i.e., local ancestry). In heterozygous individuals who are also heterozygous for local ancestry, the ancestry–allele phase is unknown, inducing ambiguity about the ancestral background of each allele. ASAFE[21] estimates allele frequencies by resolving this phase ambiguity through an expectation-maximization (EM) framework; AFA[22] provides allele frequency estimates by modeling the weighted average of ancestry-specific allele frequencies given inferred local (or global) ancestry proportions. These "local-ancestry-aware" methods leverage called (i.e., "best-guess") local ancestries and genotypes from upstream analyses, ignoring the inherent uncertainty in the ancestry and genotype calling steps. With error-free ancestry calling and genotyping, ASAFE and AFA provide precise estimates. However, both ancestry calling and genotyping can suffer from high uncertainty (or high error rates). For example, continent-level local ancestry estimates have miscall rates of $\sim 5\%$, rising over 10% in more challenging scenarios[23], and low-coverage sequencing can generate genotype calls with high variance[24]. Directly using such error-prone upstream outputs in existing ancestry-specific allele frequency estimation tools violates their model assumptions and may introduce estimation error and bias.

In this article, we develop a novel ancestry-specific allele frequency estimation method, LoGicAl (Local ancestry and Genotype calling uncertainty-aware ancestry-specific Allele frequency estimation). LoGicAl is an accelerated EM algorithm robust to uncertainty from upstream local ancestry calling and genotype calling. It informs ancestry-specific allele frequency estimates by iteratively calculating sample-wide allele dosages integrated by probabilistic representations of local ancestry and genotype likelihoods over allele frequency spectrum. Variations of LoGicAl are derived to accommodate more flexible input formats. In particular, we provide rapid implementations for phased genotype inputs. We demonstrate the accuracy and efficiency of LoGicAl and benchmark its performance against other state-of-art estimation methods through extensive simulation studies and real data applications.

2 Methods

2.1 Overview

Here, we present LoGicAl, designed to estimate ancestry-specific allele frequencies from admixed samples by leveraging uncertainty measurement from ancestry calling and genotyping to fine-tune the estimates (Figure 1). LoGicAl is a fully parametric likelihood-based model that uses a hierarchical mixture structure to incorporate the uncertainty from the upstream procedures. We decompose the complete-data likelihood, motivated by this hierarchical mixture model, into conditionally independent components that correspond to different sources of uncertainty. LoGicAl calculates the maximum likelihood estimates using an accelerated EM algorithm. LoGicAl can accommodate either probabilistic representations or best-guess calls of local ancestries and genotypes. Alternative versions of this method further leverage phasing information to generate rapid estimates.

2.2 Estimation algorithm

Suppose we have N admixed individuals genotyped at J bi-allelic sites. Each individual's genome can originate from K ancestral population groups. LoGicAl aims to estimate ancestry-specific allele frequencies $\mathbf{p} = (p_1, \ldots, p_K)$ for these K ancestral populations at each site of interest (subscript j is omitted for simplicity). Local ancestries are assumed to be estimated using existing inference tools, such as RFMix[25] and FLARE[26]. These local ancestry inference methods provide the most likely ancestry assignments and the corresponding probabilities of assigning each ancestral group. These assignment probabilities can be leveraged to calculate the marginal probabilities of ancestral groups for each individual i's two alleles, denoted by $\boldsymbol{\pi}_i = (\pi_{i,1}^{(1)}, \ldots, \pi_{i,K}^{(1)}; \pi_{i,1}^{(2)}, \ldots, \pi_{i,K}^{(2)})$. Assuming each individual is unrelated, the observed-data log-likelihood of ancestry-specific allele frequencies is

$$l(\mathbf{p}_{K\times 1}) = \sum_{i=1}^{N} \log \sum_{g=0}^{2} \Pr(X_i = x_i | G_i = g) \times \Pr(G_i = g | \boldsymbol{\pi}_i; \mathbf{p}_{K\times 1}) \tag{1}$$

where X_i is observed genotype call (or sequencing reads) and G_i is the true genotype for individual i.

LoGicAl maximizes the equation (1) to obtain the maximum likelihood estimates of ancestry-specific allele frequencies. It introduces latent variables $(Z_i^{(1)}, Z_i^{(2)})$ and $(H_i^{(1)}, H_i^{(2)})$, representing the true local ancestry and genotype per chromosome, respectively. Assuming Hardy-Weinberg equilibrium within each ancestral group, LoGicAl estimates ancestry-specific allele frequencies through the following EM algorithm[27] (derivations in Note S1):

Algorithm 1 LoGicAl Algorithm for Ancestry-specific Allele Frequency Estimation

Input: Observed data $\{x_i\}_{i=1,\ldots,N}$, local ancestry assignment probabilities $\{\pi^{(1)}_{i,k}, \pi^{(2)}_{i,k}\}_{i=1,\ldots,N;k=1,\ldots,K}$

1. Initialize ancestry-specific allele frequencies $\{p^{(0)}_k\}_{k=1,\ldots,K}$

repeat

2a. (E-step) For each individual i, compute individual-level probabilities of each possible combination of ancestry-allele phasing given the observed data and ancestry assignment probabilities

$$w_i(z^{(1)}, z^{(2)}, h^{(1)}, h^{(2)}) = \frac{\mathbb{P}\mathrm{r}(X_i = x_i | H^{(1)}_i = h^{(1)}, H^{(2)}_i = h^{(2)}) p^{h^{(1)}}_{z^{(1)}} (1 - p_{z^{(1)}})^{1-h^{(1)}} p^{h^{(2)}}_{z^{(2)}} (1 - p_{z^{(2)}})^{1-h^{(2)}} \pi^{(1)}_{i,z^{(1)}} \pi^{(2)}_{i,z^{(2)}}}{\sum_{u_1}\sum_{u_2}\sum_{v_1}\sum_{v_2} \mathbb{P}\mathrm{r}(X_i = x_i | H^{(1)}_i = u_1, H^{(2)}_i = u_2) p^{u_1}_{v_1} (1 - p_{v_1})^{1-u_1} p^{u_2}_{v_2} (1 - p_{v_2})^{1-u_2} \pi^{(1)}_{i,v_1} \pi^{(2)}_{i,v_2}}$$

2b. (M-step) Update ancestry-specific allele frequency estimates as sample-wide proportions of ancestry-alternate-allele phases among the corresponding ancestry-indexed phases

$$\hat{p}^{(t+1)}_k = \frac{\sum_{i=1}^{N} \left[\sum_{h^{(1)}=0}^{1} \sum_{z^{(1)}=1}^{K} w^{(t)}_i (z^{(1)}, k, h^{(1)}, 1) + \sum_{h^{(2)}=0}^{1} \sum_{z^{(2)}=1}^{K} w^{(t)}_i (k, z^{(2)}, 1, h^{(2)}) \right]}{\sum_{i=1}^{N} \sum_{h^{(1)}=0}^{1} \sum_{h^{(2)}=0}^{1} \left[\sum_{z^{(1)}=1}^{K} w^{(t)}_i (z^{(1)}, k, h^{(1)}, h^{(2)}) + \sum_{z^{(2)}=1}^{K} w^{(t)}_i (k, z^{(2)}, h^{(1)}, h^{(2)}) \right]}$$

until converge

Return: Estimated ancestry-specific allele frequencies $\{p^*_k\}_{k=1,\ldots,K}$

The classical EM algorithm can suffer from slow convergence[28], especially when sample size is large or the uncertainty structure is complicated, thereby limiting its applications to biobank-scale genomic data of admixed samples. We incorporate the squared iterative accelerating technique (SQUAREM)[29, 30], which approximates Newton's method to tackle the computational burden of the fixed-point algorithms[31], in LoGicAl to enhance its scalability without compromising model preciseness and stability.

2.3 Simulation study

We designed a simulation study to investigate estimation accuracy and computation time of LoGicAl and its variations across wide-range settings, and to benchmark with other state-of-art estimation tools (Figure S1). Briefly, we simulated ancestry-specific allele frequencies based on the Balding-Nichols model[32] and specified the probability distribution of underlying local ancestries in each individual. Given the simulated ancestries and ancestry-specific allele frequencies, we sampled the true genotypes and generated observed genotypes from the underlying truth after considering various genotyping error models (details in Note S2). Specifically, we quantified local ancestry resolvability by α and fixed $\alpha = 0.1$ across simulation based on its estimates from the 1000 Genomes Project (Table S1). Further, we varied $r \in \{0.2, 0.4, 0.6, 0.8, 1\}$ to control ancestral population differentiation: larger r approximates allele frequency differentiation observed between populations on different continents (e.g., between San, Han, and Maya populations), whereas smaller r reflects differentiation between populations on the same continent[33].

We implemented LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag (p: local ancestry assignment probability, a: best-guess local ancestry; l: genotype likelihood, g: best-guess genotype call), ASAFE, and AFA to the simulated unphased genotype inputs. We also applied LoGicAl-pg, LoGicAl-ag, and their phasing-aware variations as well as ASAFE and AFA to the phased genotype inputs. In particular, we implemented ASAFE and AFA using the functions `algorithm_1snp()` and `estimate_frequencies_dynamic_boundary()` with their default parameter settings, except for two modifications in `estimate_frequencies_dynamic_boundary()`: (1) setting `frequency_boundary_grid=c(0.00001,0.01)`, consistent with the AFA Common Workflow Language (CWL) implementation[22], and (2) setting `mac_filter=0` to include rare variants for benchmarking purpose.

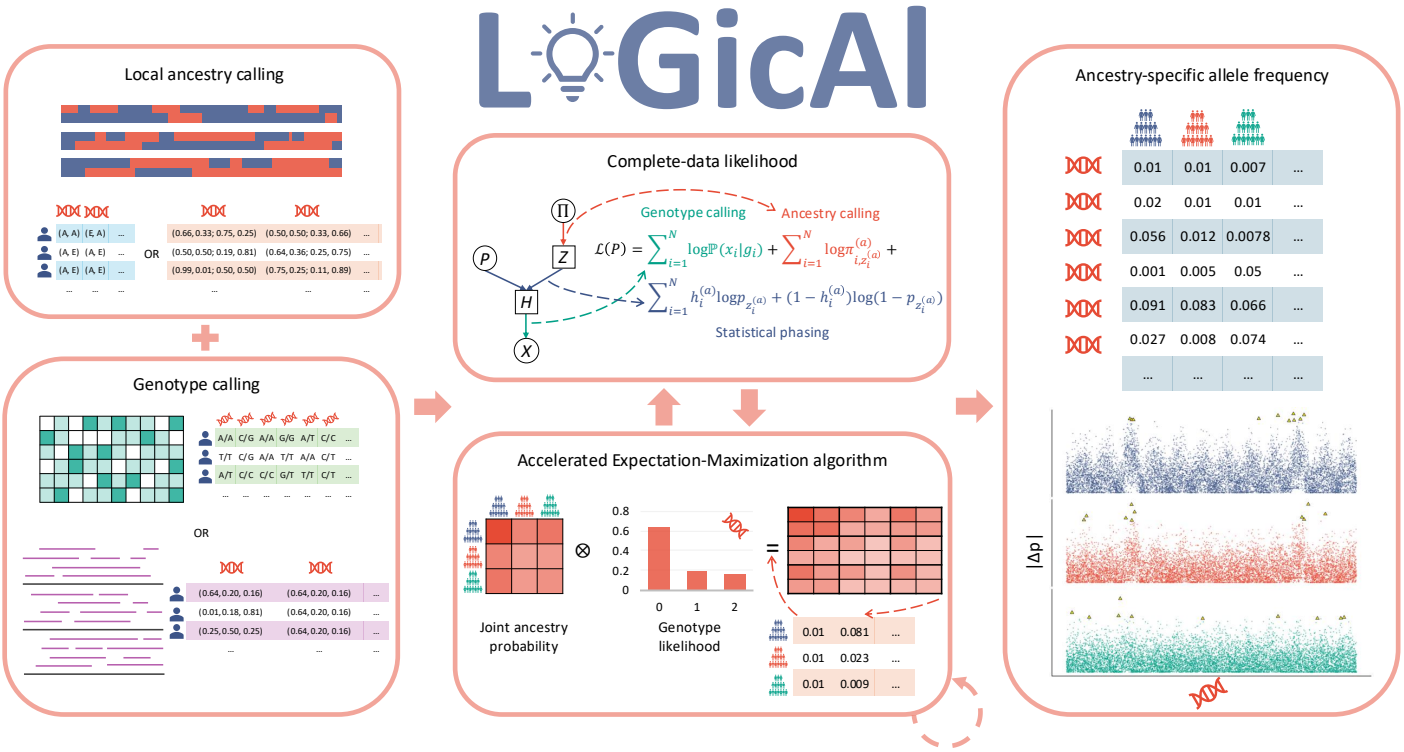

Figure 1: **Schematic overview of LoGicAl.** LoGicAl estimates ancestry-specific allele frequencies from admixed samples, simultaneously adjusting for uncertainty from upstream local ancestry calling and genotype calling. The inputs of LoGicAl are either best-guess calls or probabilistic representations of local ancestries (upper left) and genotypes (lower left). Specifically, LoGicAl-pl, LoGicAl-al, LoGicAl-pg, and LoGicAl-ag correspond to combinations of local ancestry assignment probabilities (p), or best-guess local ancestries (a), and genotype likelihoods (l), or best-guess genotype calls (g). LoGicAl leverages a hierarchical mixture framework, where the complete-data likelihood is decomposed into conditionally independent parts corresponding to different sources of uncertainty (upper middle). LoGicAl informs parameter estimates using an accelerated EM algorithm: it computes individual-level probabilities for all possible combinations of ancestry-allele phasing and updates allele frequency estimates as sample-wide proportions of ancestry-alternate-allele phases (lower middle). As output, LoGicAl provides ancestry-specific allele frequency estimates of each ancestral population at each genomic position of interest and quantifies pairwise allele frequency differences between ancestral populations across the genome (right). Outlier sites (highlighted in yellow) are defined as the top 0.1% most differentiated variants per pair.

2.4 Evaluation metrics

We evaluated and summarized the estimation accuracy of different methods using mean squared error (MSE) and relative efficiency (RE). MSE compares the true ancestry-specific allele frequencies $\mathbf{p} = (p_1, \ldots, p_K)$ and their estimates $\hat{\mathbf{p}}$ as

$$MSE(\mathbf{p}, \hat{\mathbf{p}}) = \frac{1}{K} \sum_{k=1}^{K} (p_k - \hat{p}_k)^2 \tag{2}$$

Relative efficiency quantifies the precision of an estimator T by comparing the MSE of the empirically-minimal-MSE estimator against that of the estimator T as

$$RE(T) = \frac{MSE(\mathbf{p}, \hat{\mathbf{p}}_M)}{MSE(\mathbf{p}, \hat{\mathbf{p}}_T)} \tag{3}$$

where the minimal-MSE estimates were empirically calculated as point estimates with fully known true local ancestry, genotype, and ancestry-genotype phasing information, thereby reflecting sampling variation only. Therefore, the minimal-MSE estimates correspond to a hypothetical genotyping-error-free sample that has the same sample sizes for each ancestral component as the original sample but consists of independent subsamples, and is thus unaffected by the uncertainty introduced by admixture.

2.5 Application to 1000 Genomes Project data

We estimated ancestry-specific allele frequencies from the 1000 Genomes Project (1kGP) cohort[34] using LoGicAl. This dataset contains a total number of $2,504$ unrelated individuals from 26 populations, sequenced at the New York Genome Center. With these deeply-sequenced and phased genotypes, we inferred local ancestries for each individual using RFMix[25] with the option `--node-size=5`. We used the Human Genome Diversity Panel (HGDP) dataset[35, 36] as reference panel for local ancestry calling after pre-processing this reference cohort in alignment with the quality control procedures described in [37]. The 54 populations in HGDP were condensed into seven superpopulations and we used all of them for local ancestry inference: (1) Sub-Saharan Africa, (2) Central & South Asia, (3) East Asia, (4) Europe, (5) Native America, (6) Oceania, and (7) Middle East. After obtaining local ancestry calling results, we applied LoGicAl to calculate ancestry-specific allele frequencies at biallelic single nucleotide polymorphism (SNP) sites from two admixed samples in 1kGP: (1) admixed African sample, including ACB and ASW populations, and (2) admixed American sample, including CLM, MXL, PEL, and PUR populations.

3 Results

3.1 Estimation of ancestry-specific allele frequencies with uncertainty in ancestry calling and genotyping

To assess the impact of ancestry calling uncertainty and genotyping uncertainty on different estimation methods, we vary the ancestry calling resolution from deterministic, i.e., no uncertainty in local ancestry inference, to probabilistic ($\alpha = 0.1$), and consider genotypes called from hypothetical scenario with no genotyping uncertainty, low-coverage sequencing ($1\times$), and from high-coverage sequencing ($30\times$). We generated the ancestry and genotype information of $100,000$ individuals to approximate the sample size of biobank-scale cohorts, and calculated estimation errors of LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag, ASAFE and AFA.

We first examine a scenario with no uncertainty in either ancestry inference or genotype calling (Figure 2, Column 1 and Figures S2-S3). All LoGicAl variations and ASAFE show similar estimates across allele frequency bins. Their MSEs increase monotonously with increasing allele frequency while efficiencies decrease. For the smallest frequency bin, $(0.001, 0.005)$, the MSE is 4.84×10^{-8} and the efficiency is 0.915, while for the largest frequency bin, $(0.1, 0.5)$, the MSE is 4.04×10^{-6} and the efficiency is 0.769. AFA shows comparable accuracy to LoGicAl and ASAFE for common variants (MAF$>$ 0.05), but its accuracy decreases for rare variants.

To evaluate the impact of ancestry calling uncertainty, we consider ancestry inference resolvability $\alpha = 0.1$ while assuming no genotyping uncertainty (Figure 2, Column 2 and Figures S2-S3). Parameter choice of α

corresponds to the local ancestry posterior distribution estimated from the 1000 Genomes Project (Table S1). Under this scenario, methods that explicitly model ancestry calling uncertainty (LoGicAl-pl/pg) achieve MSEs similar to the scenario without ancestry calling uncertainty, ranging from 5.73×10^{-8} for the smallest frequency bin to 5.26×10^{-6} for the largest frequency bin. The efficiencies of these methods are slightly lower than under the hypothetical scenario, ranging from 0.841 to 0.680. Methods that do not model ancestry calling uncertainty show $\sim$ 10-fold higher MSEs: LoGicAl-al/ag and ASAFE range from 5.12×10^{-7} to 4.23×10^{-5}, while AFA ranges from 8.14×10^{-6} to 6.73×10^{-5}. The relative efficiencies of LoGicAl-al/ag and ASAFE drop by 65.6% to $0.324 - 0.242$.

To evaluate the impact of genotype calling uncertainty, we consider sequencing depths from low-coverage ($1\times$) to high-coverage ($30\times$) in the absence of local ancestry inference uncertainty (Figure 2, Column 3 and Figures S2-S3). At $1\times$ depth, methods that explicitly model genotyping uncertainty (LoGicAl-pl/al) achieve MSEs about three times the hypothetical scenario, ranging from 1.82×10^{-7} to 1.24×10^{-5} across allele frequency bins. The efficiencies of these methods are 43.1% lower, ranging from 0.546 to 0.438. This reduction reflects the uncertainty generated by both low-coverage sequencing and the estimation procedure. Methods that do not model genotyping uncertainty show $\sim$ 850-fold higher MSEs compared to the scenario without genotyping uncertainty: LoGicAl-pg/ag and ASAFE range from 1.24×10^{-5} to 0.0137, while AFA ranges from 7.67×10^{-5} to 0.0138. The relative efficiencies of LoGicAl-pg/ag and ASAFE drop by 96.7% to $0.0628 - 0.0138$. At $30\times$ depth, estimation accuracy of all methods approaches that under the hypothetical scenario with no uncertainty in either genotyping or ancestry calling (Figures S4-S5).

In a scenario with both uncertain local ancestry calling and low-coverage sequencing (Figure 2, Column 4 and Figures S2-S3), most methods perform similarly to the scenario with genotyping uncertainty only. LoGicAl-pl, which models both types of uncertainty, achieves the lowest MSEs of $2.37 \times 10^{-7} - 1.48 \times 10^{-5}$ across allele frequency bins. LoGicAl-ag and ASAFE range from 1.37×10^{-5} to 0.0138. LoGicAl-pg, modeling ancestry calling uncertainty but not genotyping uncertainty, ranges from 1.22×10^{-5} to 0.0137; LoGicAl-al, modeling genotyping uncertainty but not ancestry calling uncertainty, ranges from 5.45×10^{-7} to 4.95×10^{-5}. AFA shows MSEs ranging from 7.77×10^{-5} to 0.0139. These performance patterns remain consistent across varying numbers of ancestral populations, from two-way to five-way admixture (Figures S6-S7).

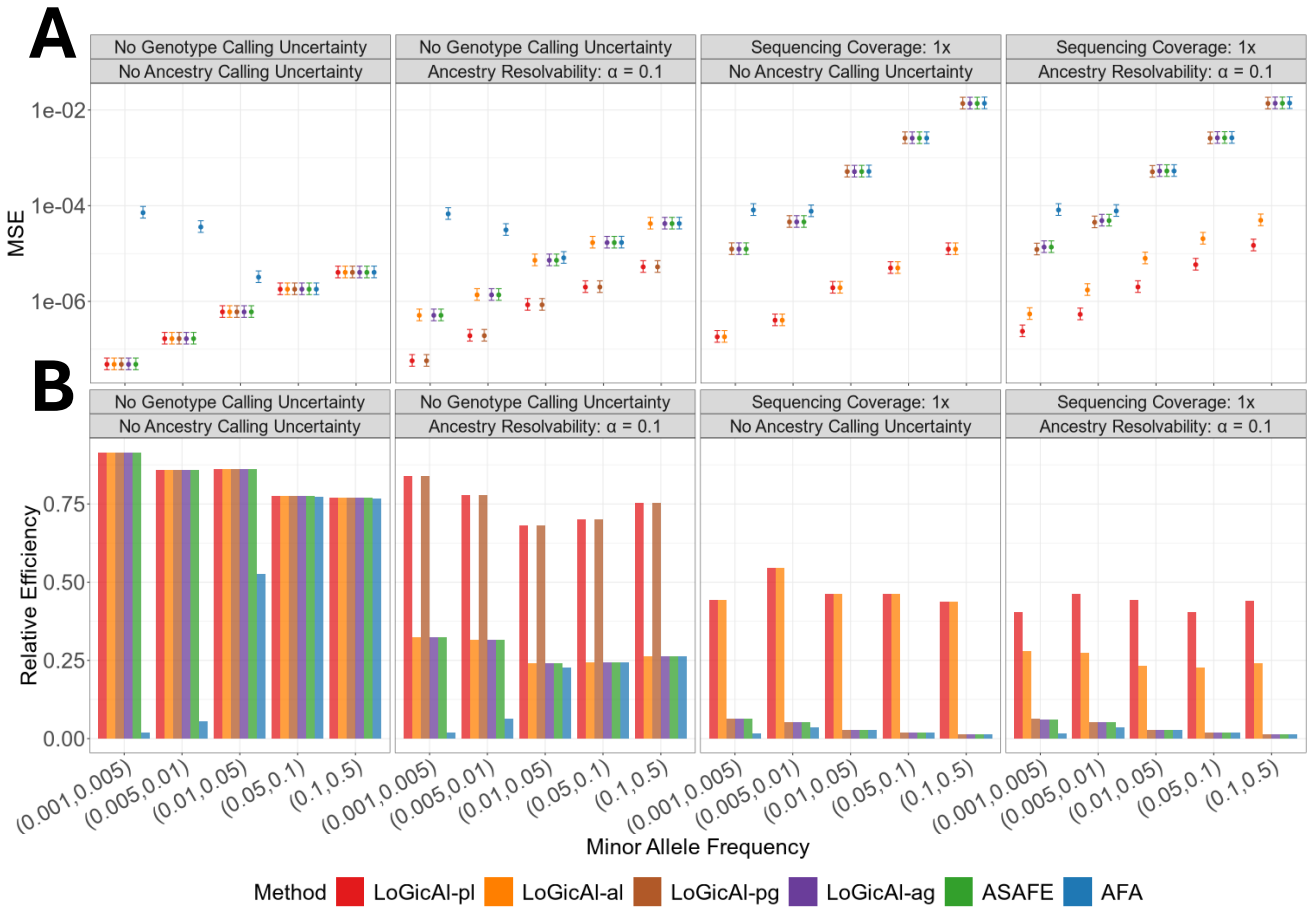

Figure 2: **Accuracy of ancestry-specific allele frequency estimation methods across uncertainty scenarios and allele frequency bins.** (A) Mean squared error (MSE) and (B) relative efficiency, each calculated over 100 simulation replicates. We simulated $100,000$ unphased individuals from a three-way admixture model with intra-continental ancestral population differentiation ($r = 0.2$). Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag), ASAFE, and AFA under four settings: no calling uncertainty, ancestry calling uncertainty only, genotype calling uncertainty only, and both ancestry and genotype calling uncertainty. In panel A, points denote replicate-averaged MSEs, and intervals indicate 95% confidence intervals for MSEs based on chi-squared approximation. In panel B, bars denote relative efficiencies relative to the optimal estimator.

3.2 Impact of ancestral population differentiation

We investigate the impact of genetic differentiation among ancestral populations on the precision of ancestry-specific allele frequency estimators in the interplay with uncertainty structure. We varied the differentiation between intra-continental differentiation ($r = 0.2$) and inter-continental differentiation ($r = 1$). Initially focusing on variants with reference allele frequency in $(0.01, 0.05)$, we observe that under the scenario with no ancestry calling or genotyping uncertainty (Figure 3, Column 1), all LoGicAl variations and ASAFE show MSEs ranging from 5.83×10^{-7} to 7.05×10^{-7}, with efficiencies of $0.915 - 0.801$. We observe no clear directional effect of increasing ancestral population differentiation. AFA exhibits larger MSE than LoGicAl and ASAFE at $r = 0.2$, and its MSE increases with increasing differentiation. This pattern for AFA is more pronounced for rare variants than for common variants. For common variants with reference allele frequency in $(0.1, 0.5)$, LoGicAl, ASAFE and AFA show similar accuracy across differentiation levels (Figures S2-S3).

In the presence of ancestry calling uncertainty (Figure 3, Column 2), MSEs for methods that explicitly model it (LoGicAl-pl/pg) increase slightly, ranging from 6.72×10^{-7} to 9.11×10^{-7} across differentiation levels, while efficiencies range from 0.776 to 0.680 without clear impact of increasing differentiation. Methods that do not model ancestry calling uncertainty have higher errors that grow with increasing differentiation: LoGicAl-al/ag and ASAFE increase 12-fold at $r = 0.2$ (7.23×10^{-6}) to 51-fold at $r = 1$ (3.57×10^{-5}). For these methods efficiencies range from 0.242 at $r = 0.2$ to 0.134 at $r = 1$.

In the presence of genotyping uncertainty (Figure 3, Column 3), genotyping-uncertainty-aware methods (LoGicAl-pl/al) achieve MSEs about three times as high as the hypothetical scenario ($1.55 \times 10^{-6} - 1.93 \times 10^{-6}$), while their efficiencies are reduced to $0.555 - 0.463$ without clear impact of increasing differentiation. Genotyping-uncertainty-unaware methods (LoGicAl-pg/ag, ASAFE, AFA) show at least 857 times increased MSEs from 5.16×10^{-4} at $r = 0.2$ to 1.15×10^{-3} at $r = 1$, with efficiencies below 0.03 for all levels of differentiation. Taken together, MSEs for uncertainty-aware methods remain relatively stable across ancestral population differentiation, while MSEs for uncertainty-unaware methods monotonically increase with greater ancestral differentiation. These performance patterns are consistent across allele frequency bins (Figures S2-S3).

3.3 Estimation with phased genotype inputs

To evaluate estimation accuracy when local ancestries and genotypes are jointly phased (Figure 4), we consider a scenario with probabilistic ancestry calling ($\alpha = 0.1$) and a realistic genotyping error rate ($\theta = 5 \times 10^{-4}$). We simulated ancestry and genotype data for $100,000$ individuals and varied the discordance rate from 0% to 10%. This discordance rate quantifies proportions of wrongly flipped ancestry assignments, which can arise from switch errors and double switch errors (i.e., flip errors) in statistical phasing[38] and inferred ancestry segment flips[39] (details in Note S2). We benchmarked LoGicAl-pg and LoGicAl-ag in both phasing-unaware (unphased) and phasing-aware (phased) implementations against ASAFE and AFA.

The phasing-unaware methods (LoGicAl-pg/ag, ASAFE, AFA) show no change in MSEs across discordance rates: 1.14×10^{-6} for LoGicAl-pg, 7.65×10^{-6} for LoGicAl-ag and ASAFE, and 8.21×10^{-6} for AFA. In contrast, the phasing-aware methods exhibit increasing errors with higher discordance, resulting in smaller MSEs than the phasing-unaware methods at low discordance ($\sim 0\%$) and larger MSEs at high discordance ($\sim 10\%$). Specifically, LoGicAl-pg (phased) increases from 7.83×10^{-7} at 0% discordance to 3.05×10^{-6} at 10% discordance, and LoGicAl-ag (phased) increases from 7.00×10^{-6} at 0% discordance to 1.51×10^{-5} at 10% discordance. These performance patterns are consistent across allele frequency bins and across levels of ancestral population differentiation (Figures S8-S9), as well as across numbers of ancestral populations (Figures S10-S11).

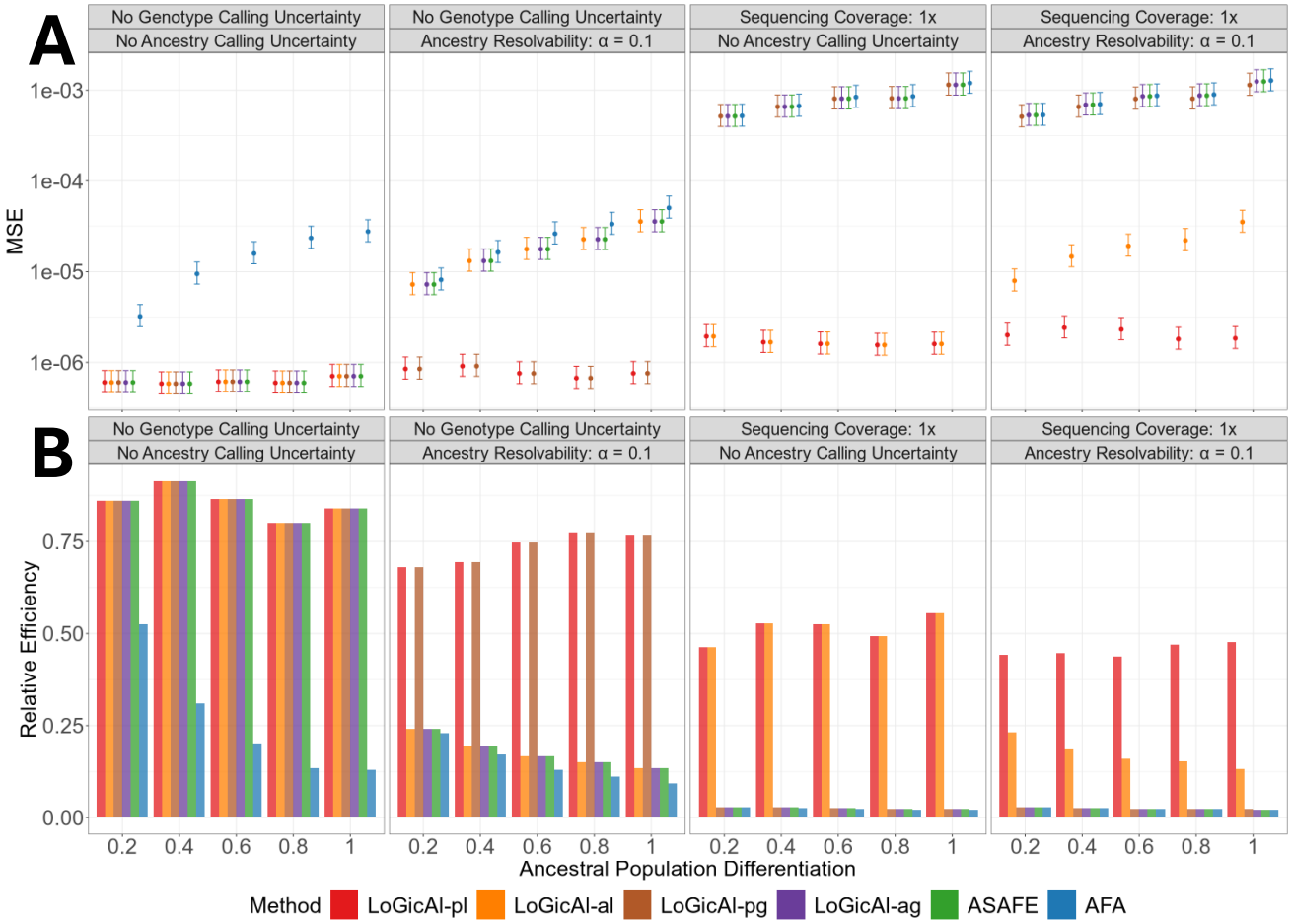

Figure 3: **Accuracy of ancestry-specific allele frequency estimation methods across uncertainty scenarios and levels of ancestral population differentiation.** (A) Mean squared error (MSE) and (B) relative efficiency, each calculated over 100 simulation replicates. We simulated $100,000$ unphased individuals from a three-way admixture model with reference allele frequency in $(0.01, 0.05)$. Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag), ASAFE, and AFA under four settings: no calling uncertainty, ancestry calling uncertainty only, genotype calling uncertainty only, and both ancestry and genotype calling uncertainty. Smaller value of ancestral population differentiation denotes intra-continental differentiation; larger value denotes inter-continental differentiation. In panel A, points denote replicate-averaged MSEs, and intervals indicate 95% confidence intervals for MSEs based on chi-squared approximation. In panel B, bars denote relative efficiencies relative to the optimal estimator.

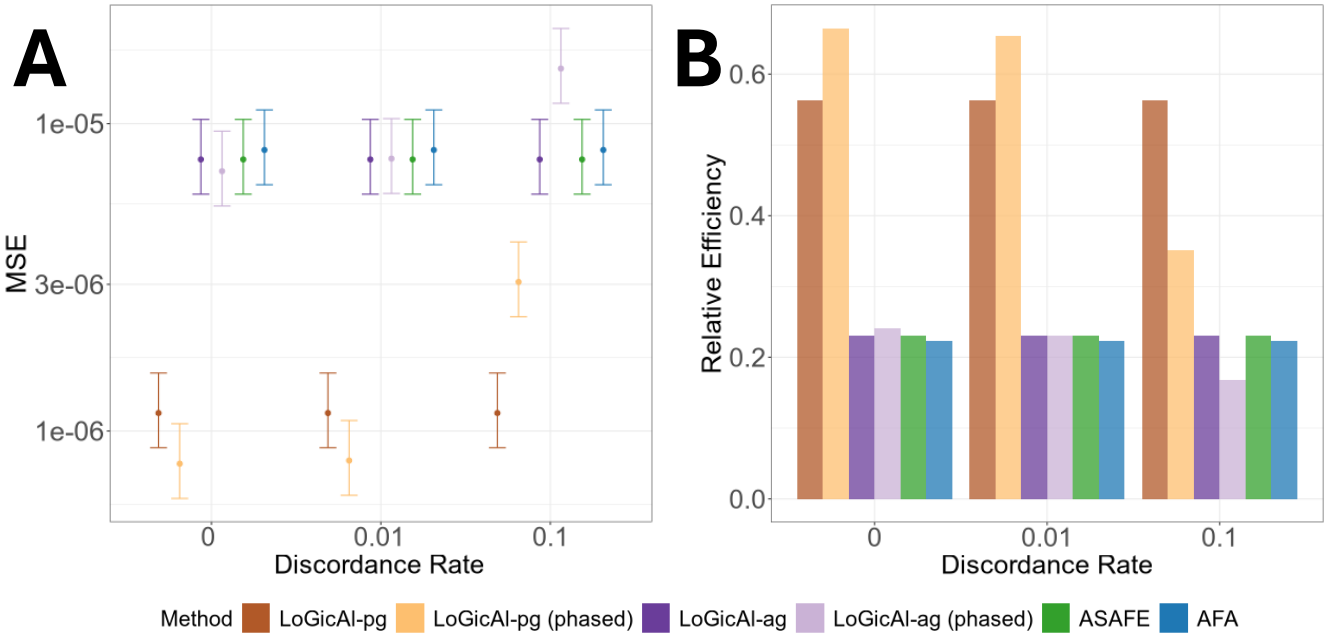

Figure 4: **Accuracy of ancestry-specific allele frequency estimation methods across discordance rates.** (A) Mean squared error (MSE) and (B) relative efficiency, each calculated over 100 simulation replicates. We simulated $100,000$ phased individuals from a three-way admixture model with reference allele frequency in $(0.01, 0.05)$ and intra-continental differentiation ($r = 0.2$). Methods compared include LoGicAl (LoGicAl-pg, LoGicAl-pg (phased), LoGicAl-ag, LoGicAl-ag (phased)), ASAFE, and AFA under a scenario with probabilistic ancestry calling ($\alpha = 0.1$) and a realistic genotyping error rate ($\theta = 5 \times 10^{-4}$). In panel A, points denote replicate-averaged MSEs, and intervals indicate 95% confidence intervals for MSEs based on chi-squared approximation. In panel B, bars denote relative efficiencies relative to the optimal estimator.

3.4 Computation time

We evaluate the computational performance of LoGicAl, ASAFE, and AFA in terms of runtime for unphased and phased genotype inputs. For unphased data (Table S2), the computation speed of LoGicAl depends on the types of uncertainty that are modeled: modeling both ancestry and genotype calling uncertainty (LoGicAl-pl) requires $\sim 3-11$ times the computation time of modeling neither (LoGicAl-ag), and modeling either ancestry calling uncertainty (LoGicAl-pg) or genotyping uncertainty (LoGicAl-al) requires $\sim 2-3$ times and $\sim 1-2$ times, respectively. Compared to AFA and ASAFE which do not model any uncertainty, LoGicAl is generally substantially faster: compared with AFA, all LoGicAl variants are at least 3 times and up to 56 times faster under the same model. When compared with ASAFE, LoGicAl is up to 6 times faster under the same model or when modeling either ancestry calling uncertainty or genotyping uncertainty. When modeling both, LoGicAl shows similar computation time to ASAFE, ranging from 49.5% faster than ASAFE to 63.4% slower, depending on the complexities of the uncertainty structure. These patterns are consistent across levels of ancestral differentiation (Figure S12).

For phased genotype data (Table S3), the phased version of LoGicAl is $\sim 2-5$ times faster than the unphased version, and thus uniformly faster than AFA (up to 206 fold) and ASAFE (up to 18 fold). These patterns are consistent across levels of ancestral differentiation (Figure S13).

3.5 Application to 1000 Genomes Project

We applied LoGicAl to estimate ancestry-specific allele frequencies from the admixed samples in the 1000 Genomes Project. Specifically, we estimated (1) African and European ancestral allele frequencies in the admixed African sample (ACB and ASW cohorts), and (2) African, European and Native American ancestral allele frequencies in the admixed American sample (CLM, MXL, PEL and PUR cohorts).

To validate LoGicAl estimates, we compared the ancestry-specific allele frequency estimates with allele frequencies of the corresponding non-admixed ancestral populations in HGDP. Correlations are high for ancestry-matched comparisons and lower for ancestry-mismatched pairs: from 1kGP admixed African cohorts, African ancestral estimates highly correlate with HGDP African cohorts ($r = 0.982$), and European ancestral estimates highly correlate with HGDP Europeans ($r = 0.975$). In contrast, cross-ancestry correlations are lower (African ancestral estimates versus HGDP Europeans $r = 0.792$; European ancestral estimates versus HGDP Africans $r = 0.772$). Similarly, from 1kGP admixed American cohorts, correlations of African ancestral allele frequency estimates with HGDP African, European and Native American cohorts are 0.965, 0.760, and 0.702, respectively; correlations of European ancestral estimates with HGDP African, European and Native American cohorts are 0.772, 0.994, and 0.848; correlations of Native American ancestral estimates with HGDP African, European and Native American cohorts are 0.705, 0.837, and 0.985.

To compare the impact of using local versus global ancestry on allele frequency estimation, we calculated correlations between LoGicAl estimates (local-ancestry-based) and allele frequency estimates derived from methods that assign a single global ancestry label per individual, as in the gnomAD v4.1.0 whole-genome sequencing dataset[40]. In gnomAD, both non-admixed and admixed African cohorts are labeled as "African", and admixed American cohorts are labeled as "American". Correlations between gnomAD African allele frequencies and LoGicAl African ancestral estimates are 0.989 for 1kGP admixed African cohorts and 0.970 for 1kGP admixed American cohorts; correlation between gnomAD American allele frequencies and LoGicAl Native American ancestral estimates is 0.923.

To illustrate how local- versus global-ancestry-based approaches differ at specific loci, we compared their estimates at LoGicAl-identified ancestry-differentiated variants on chromosome 1 (Figure S14). LoGicAl (local-ancestry-based) provides more distinct estimates of ancestral population differentiation compared to gnomAD (global-ancestry-based). For example, in 1kGP admixed African cohorts, $rs2814778$ (in $ACKR1$) is the most differentiated site with African ancestral estimate of 0.996 and European ancestral estimate of < 0.001, whereas gnomAD estimates are AFR = 0.822 and EUR = 0.004. Similar patterns are observed in estimates from 1kGP admixed American cohorts (Table S4).

4 Discussion

In this article, we develop LoGicAl, a novel accurate, flexible, and scalable tool for estimating ancestry-specific allele frequencies from admixed samples. LoGicAl applies likelihood-based estimation that resolves an

ancestry-genotype-integrated hierarchical mixture structure through an accelerated EM algorithm. LoGicAl provides faster implementations for phased genotype inputs when ancestry-genotype phasing is reliable. By effectively modeling uncertainty arising from ancestry calling and genotyping, as well as accounting for statistical phasing errors, LoGicAl achieves superior accuracy and scalability compared with state-of-art estimation methods in both realistic-parameter-driven simulations and analyses of 1000 Genomes Project data.

LoGicAl accommodates four important features of modern genetic analyses of admixed samples. First, it improves ancestry-specific allele frequency estimation accuracy by explicitly modeling uncertainty in local ancestry inference. Treating local ancestry as probability distributions rather than as point calls is particularly important when inferring local ancestry is challenging, for example, when ancestral populations are closely related[26]. Ignoring such uncertainty induces bias by shrinking ancestry-specific allele frequency estimates towards the unstratified allele frequency. The magnitude of this error grows with increasing ancestral population differentiation. Though local ancestry misspecifications are less common for more differentiated ancestral populations, even rare misspecifications can produce substantial errors if not appropriately modeled.

Second, LoGicAl provides accurate ancestry-specific allele frequency estimates from low-coverage whole-genome sequencing. Low-coverage sequencing has garnered broad interests for its ability to capture novel variants in underrepresented populations[24], increase GWAS power[41] and to improve PRS transferability[42]. Companies like Regeneron use low-pass genome plus high-coverage exome sequencing at scale as a cost-effective alternative to WGS or WES[43]. Our results demonstrate that ignoring genotyping uncertainty inherent to low-coverage sequencing induces bias in ancestry-specific allele frequency estimates. By explicitly modeling this uncertainty, LoGicAl substantially mitigates estimation errors and provides more precise and efficient estimates compared to existing methods across allele frequencies.

Third, phasing-aware LoGicAl reduces ancestry-specific allele frequency estimation errors by leveraging ancestry–genotype phase when phasing error is negligible. With increasing phasing error, phasing-aware estimators become less precise, exceeding MSEs of phasing-unaware estimators at reported phasing error rates in the literature[44]. Practically, the phased implementation is preferred when high-quality ancestry-genotype phasing is available (e.g., trio-based inference), whereas phasing-unaware approaches are preferred when phasing quality is uncertain or unknown.

Fourth, LoGicAl accelerates the implementation of the classical EM algorithm through the squared iterative method, without compromising accuracy or stability. When phased genotype inputs are available, they substantially reduce the parameter space within which the accelerated EM algorithm searches, thereby further reducing computational costs. Altogether, these computational improvements enable the utilities of LoGicAl for analyzing biobank-scale genomic data of admixed samples, such as the TOPMed program[45].

Our application to 1000 Genomes Project admixed samples shows that LoGicAl produces ancestry-specific allele frequency estimates that are well calibrated to non-admixed HGDP references, recovering allele frequencies in the corresponding ancestral populations. Comparing local- and global-ancestry strategies, we find that global labels reflect general ancestry patterns but obscure admixture-specific structure, resulting in shrunken allele frequency estimates. In contrast, the local-ancestry approach pinpoints locus-specific ancestral population differentiation masked with global labeling. We illustrate this by, for example, $rs2814778$ in the $ACKR1$ region, which codes atypical chemokine receptor 1 and can affect malaria susceptibility[46]. At this variant, African and European ancestral population differentiation is more pronounced using LoGicAl-derived estimates than with global-ancestry-based frequencies from gnomAD.

Several extensions could further enhance statistical and computational performance of LoGicAl. First, the current estimation framework assumes unrelated samples. While point estimators remain unbiased under relatedness, dependence can inflate precision metrics if not accounted for. Robust estimators or kinship-aware weighting could be leveraged to inform calibrated standard errors and confidence intervals. Second, LoGicAl provides marginal site-wise ancestry-specific allele frequency estimates and it does not explicitly leverage linkage disequilibrium (LD) structure across sites. Grouping variants by LD structure and local ancestry segments and performing block-wise analyses could benefit its computational performance[47].

To summarize, LoGicAl provides comprehensive and precise descriptions of genetic variation across genome and informs ancestry-specific allele frequency spectrum for admixed populations. LoGicAl further contributes to constructing a precise spatial landscape and dynamics of genetic variation across geographical regions, and promote the understandings in population structure and demographic history at a finer scale.

Data and code availability. The LoGicAl software is available at https://github.com/student-jw/LoGicAl.

Acknowledgments. This work was supported by the National Institutes of Health (NIH) grants R01 HG011031 and R01 HG005855. We acknowledge Gonçalo Abecasis for helpful discussion on phasing-aware approaches.

Declaration of interests. The authors declare no competing interests.

Supplemental figures

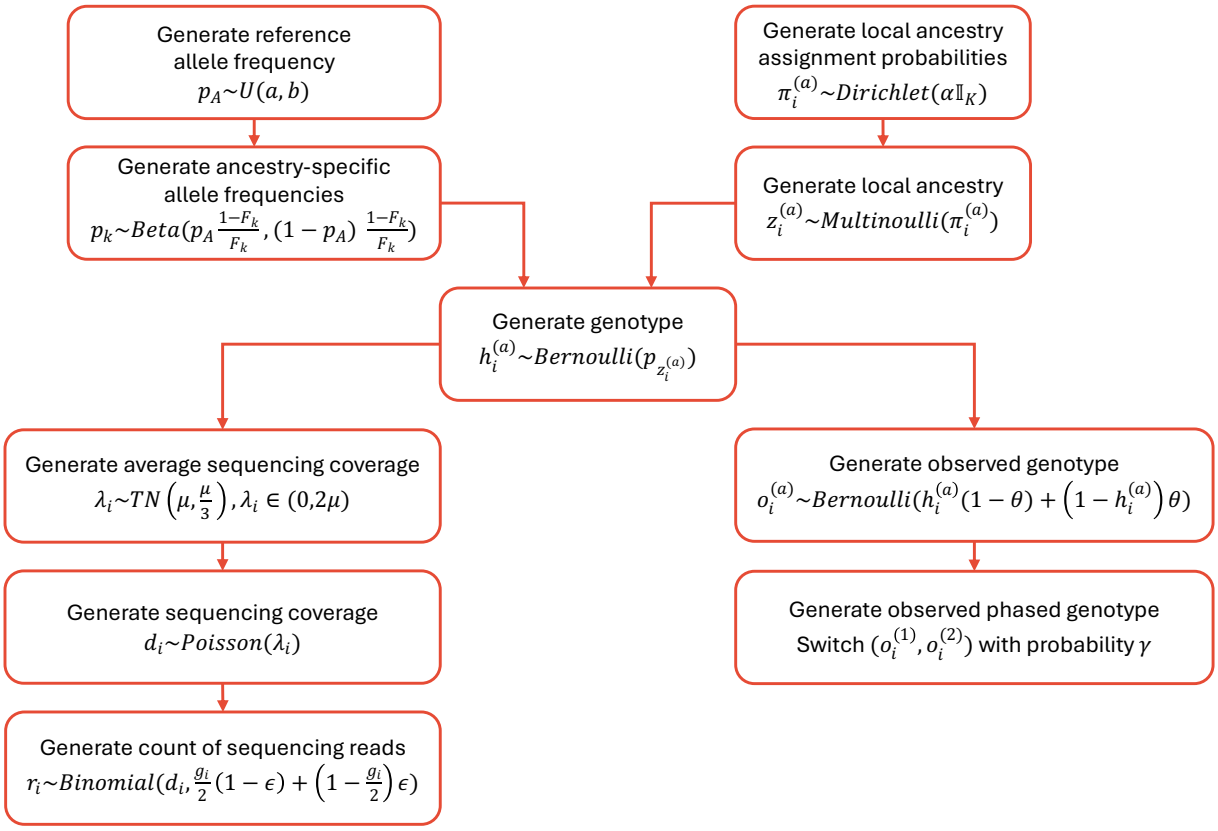

Figure S1: Simulation overview.

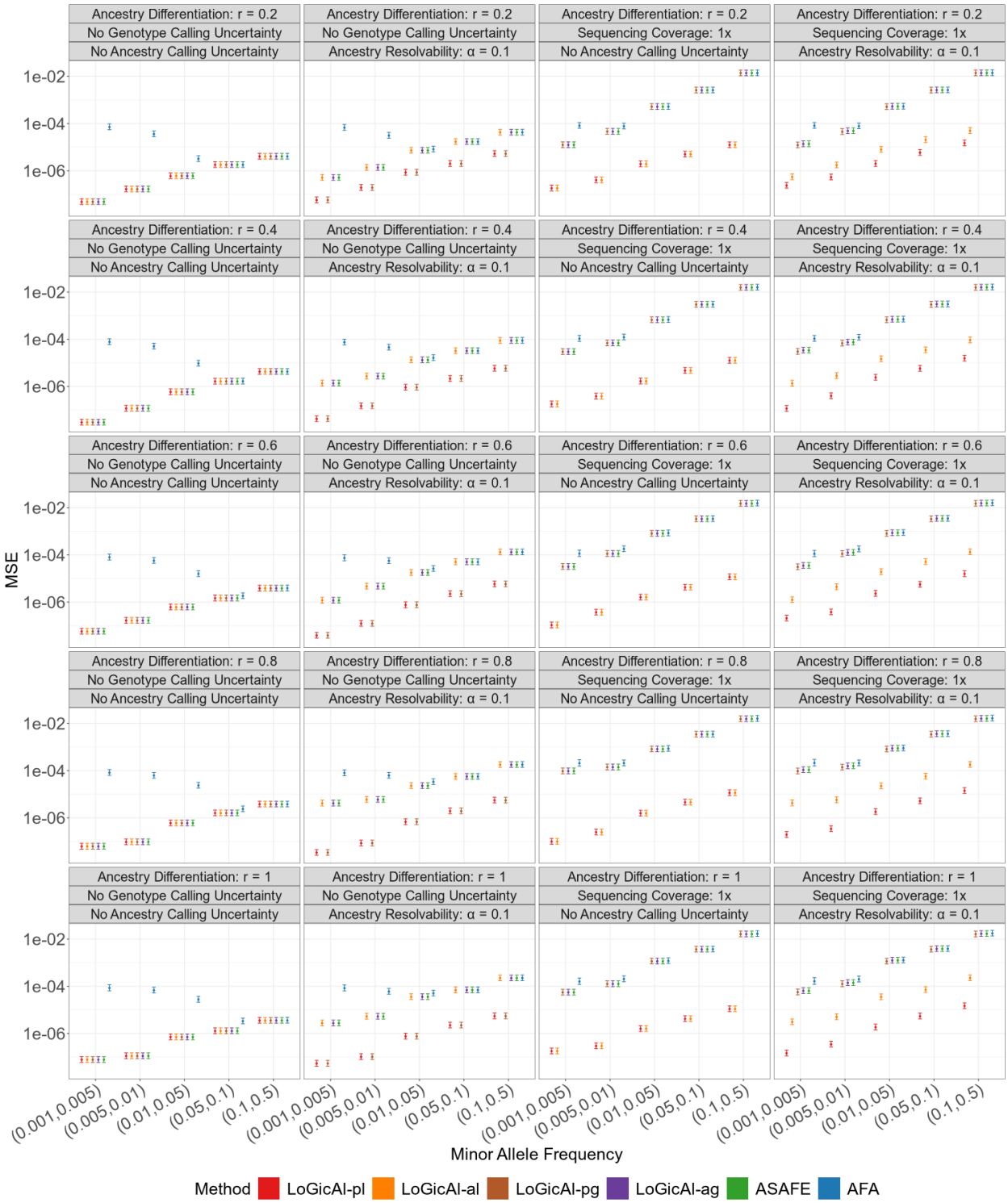

Figure S2: Mean squared error of ancestry-specific allele frequency estimation methods across uncertainty scenarios, allele frequency bins and levels of ancestral population differentiation. We simulated 100,000 unphased individuals from a three-way admixture model with intra-continental differentiation ($r = 0.2$) to inter-continental differentiation ($r = 1$). Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag), ASAFE, and AFA under four settings: no calling uncertainty, ancestry calling uncertainty only, genotype calling uncertainty only, and both ancestry and genotype calling uncertainty. Points denote replicate-averaged MSEs, and intervals indicate 95% confidence intervals for MSEs based on chi-squared approximation.

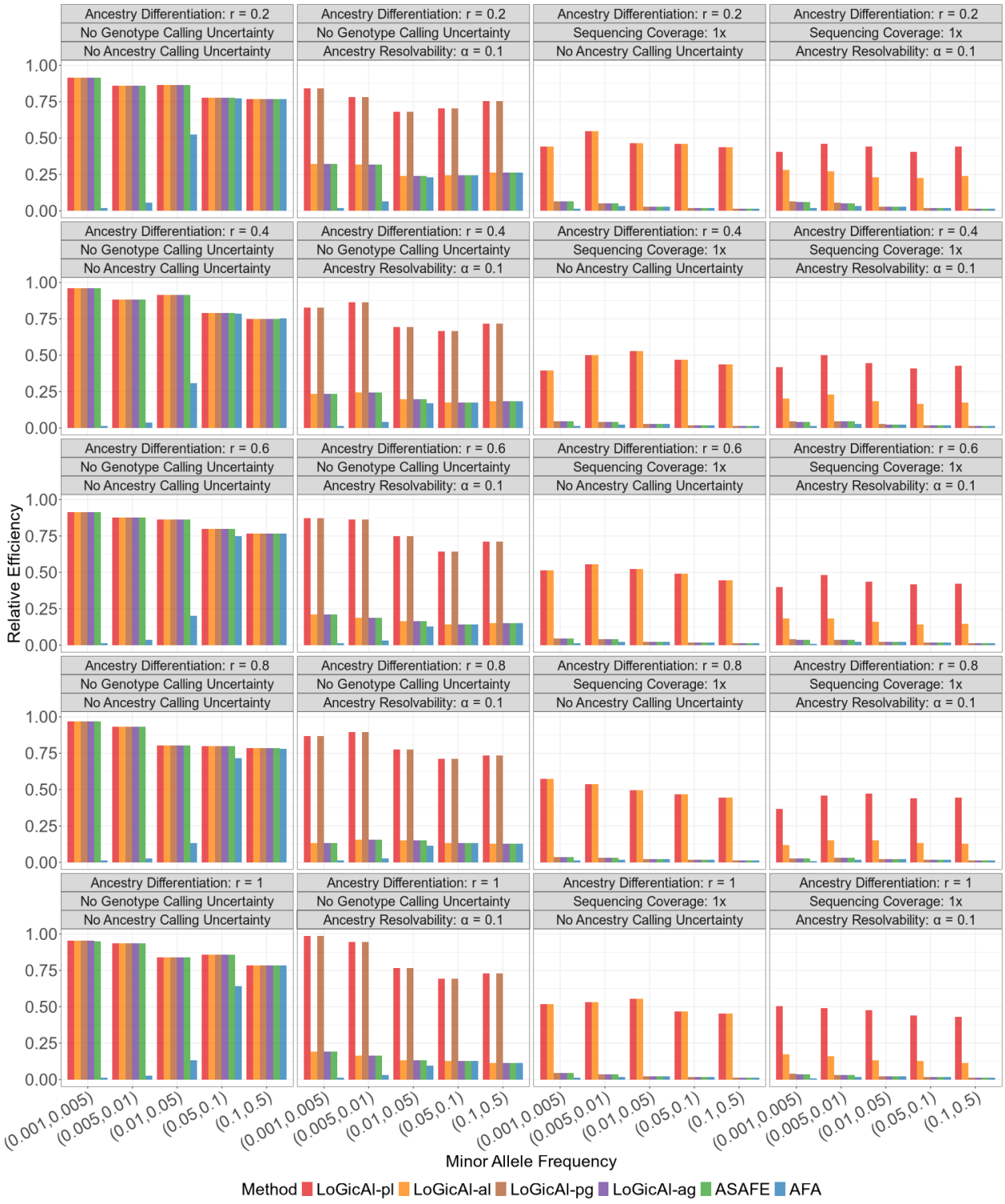

Figure S3: Relative efficiency of ancestry-specific allele frequency estimation methods across uncertainty scenarios, allele frequency bins and levels of ancestral population differentiation. We simulated 100,000 unphased individuals from a three-way admixture model with intra-continental differentiation ($r = 0.2$) to inter-continental differentiation ($r = 1$). Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag), ASAFE, and AFA under four settings: no calling uncertainty, ancestry calling uncertainty only, genotype calling uncertainty only, and both ancestry and genotype calling uncertainty. Bars denote relative efficiencies relative to the optimal estimator.

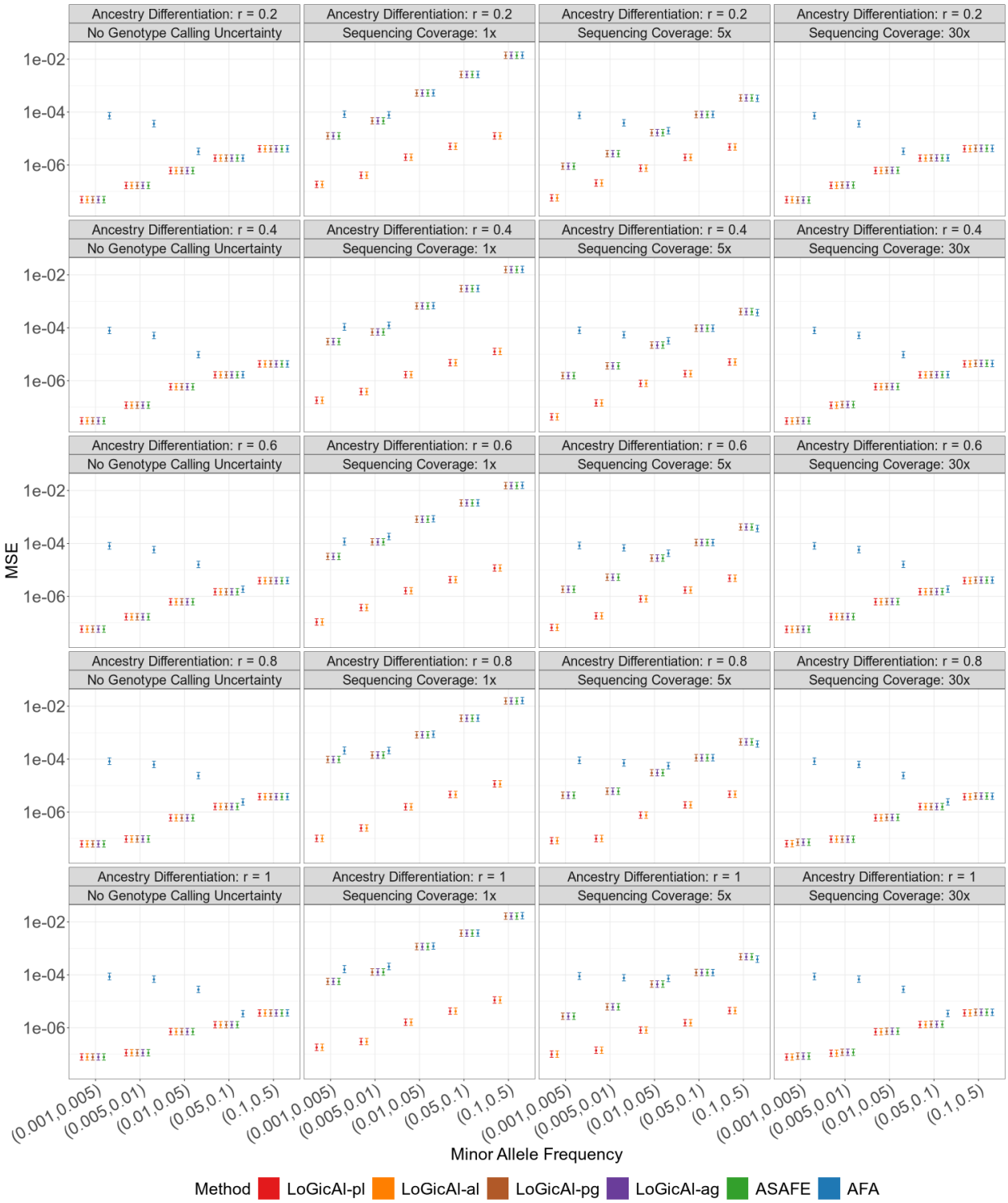

Figure S4: Mean squared error of ancestry-specific allele frequency estimation methods across sequencing depths, allele frequency bins and levels of ancestral population differentiation. We simulated 100,000 unphased individuals from a three-way admixture model with intra-continental differentiation ($r = 0.2$) to inter-continental differentiation ($r = 1$) while assuming no ancestry inference uncertainty. Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag), ASAFE, and AFA. Points denote replicate-averaged MSEs, and intervals indicate 95% confidence intervals for MSEs based on chi-squared approximation.

Figure S5: Relative efficiency of ancestry-specific allele frequency estimation methods across sequencing depths, allele frequency bins and ancestral population differentiation. We simulated $100,000$ unphased individuals from a three-way admixture model with intra-continental differentiation ($r = 0.2$) to inter-continental differentiation ($r = 1$) while assuming no ancestry inference uncertainty. Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag), ASAFE, and AFA. Bars denote relative efficiencies relative to the optimal estimator.

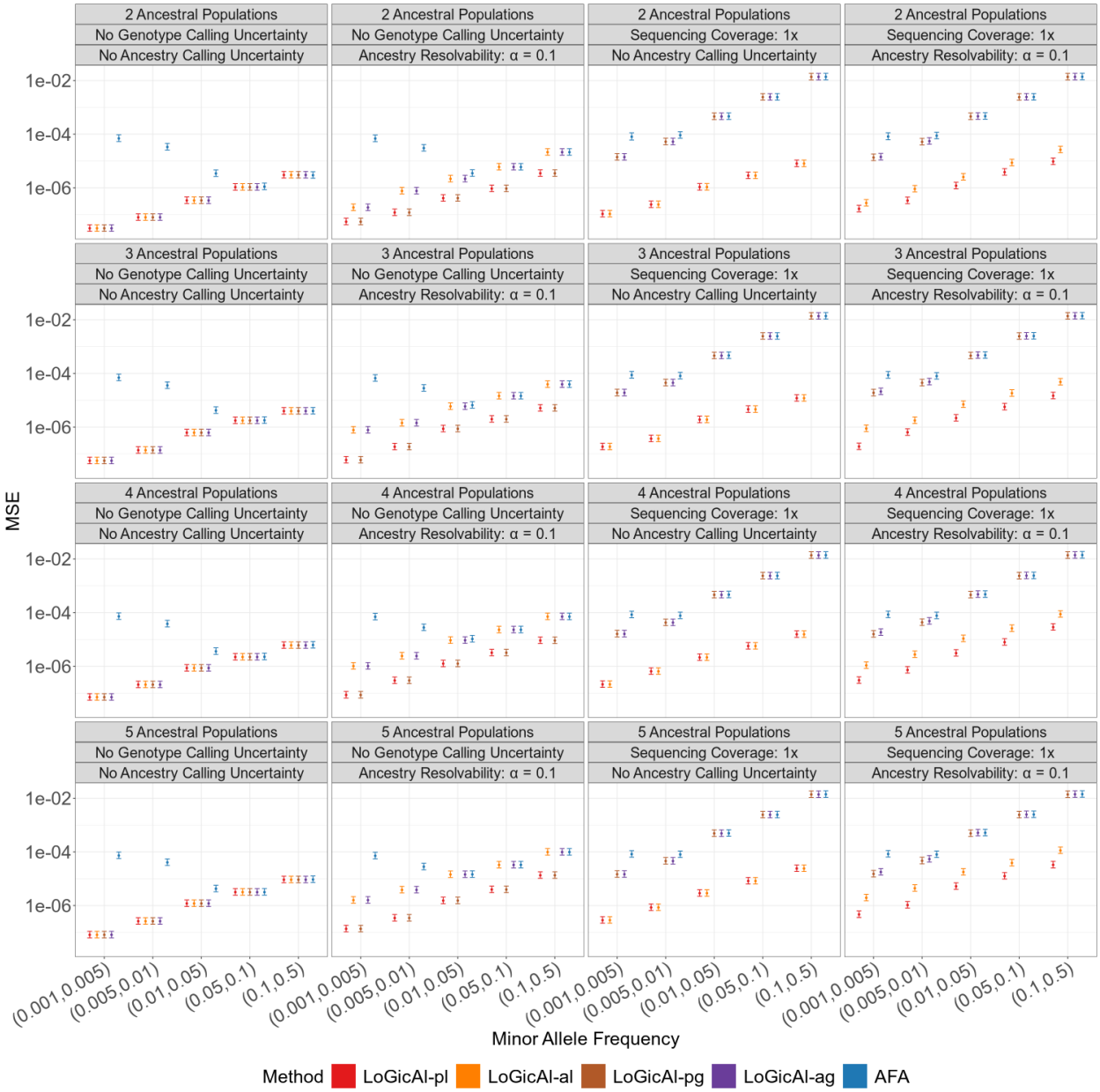

Figure S6: Mean squared error of ancestry-specific allele frequency estimation methods across uncertainty scenarios, allele frequency bins and numbers of ancestral populations. We simulated 100,000 unphased individuals under two- to five-way admixture with intra-continental differentiation. Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag) and AFA under four settings: no calling uncertainty, ancestry calling uncertainty only, genotype calling uncertainty only, and both ancestry and genotype calling uncertainty. ASAFE is not included as it is developed for three-way (and two-way) admixture scenario only. Points denote replicate-averaged MSEs, and intervals indicate 95% confidence intervals for MSEs based on chi-squared approximation.

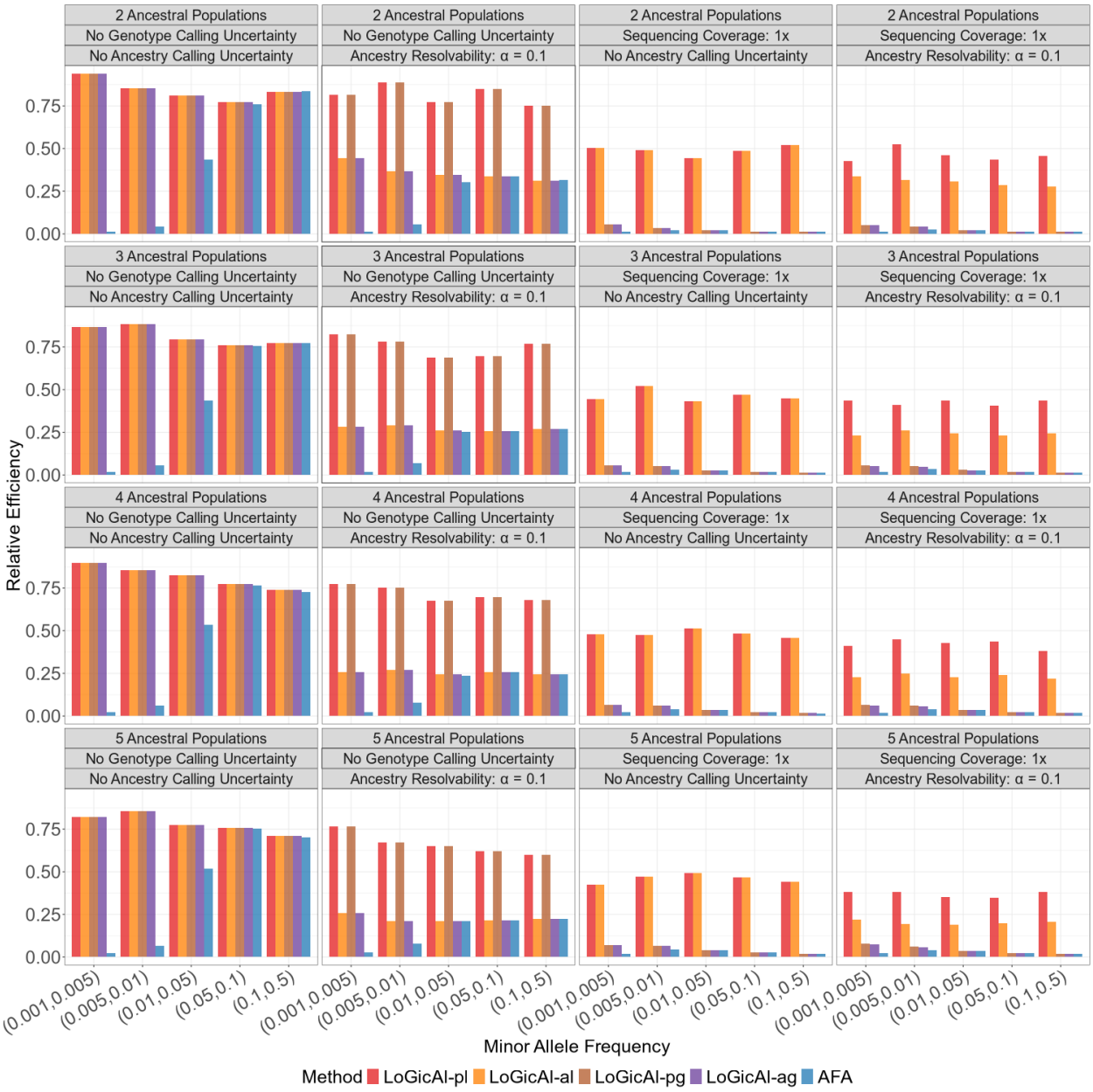

Figure S7: Relative efficiency of ancestry-specific allele frequency estimation methods across uncertainty scenarios, allele frequency bins and numbers of ancestral populations. We simulated 100,000 unphased individuals under two- to five-way admixture with intra-continental differentiation. Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag) and AFA under four settings: no calling uncertainty, ancestry calling uncertainty only, genotype calling uncertainty only, and both ancestry and genotype calling uncertainty. ASAFE is not included as it is developed for three-way (and two-way) admixture scenario only. Bars denote relative efficiencies relative to the optimal estimator.

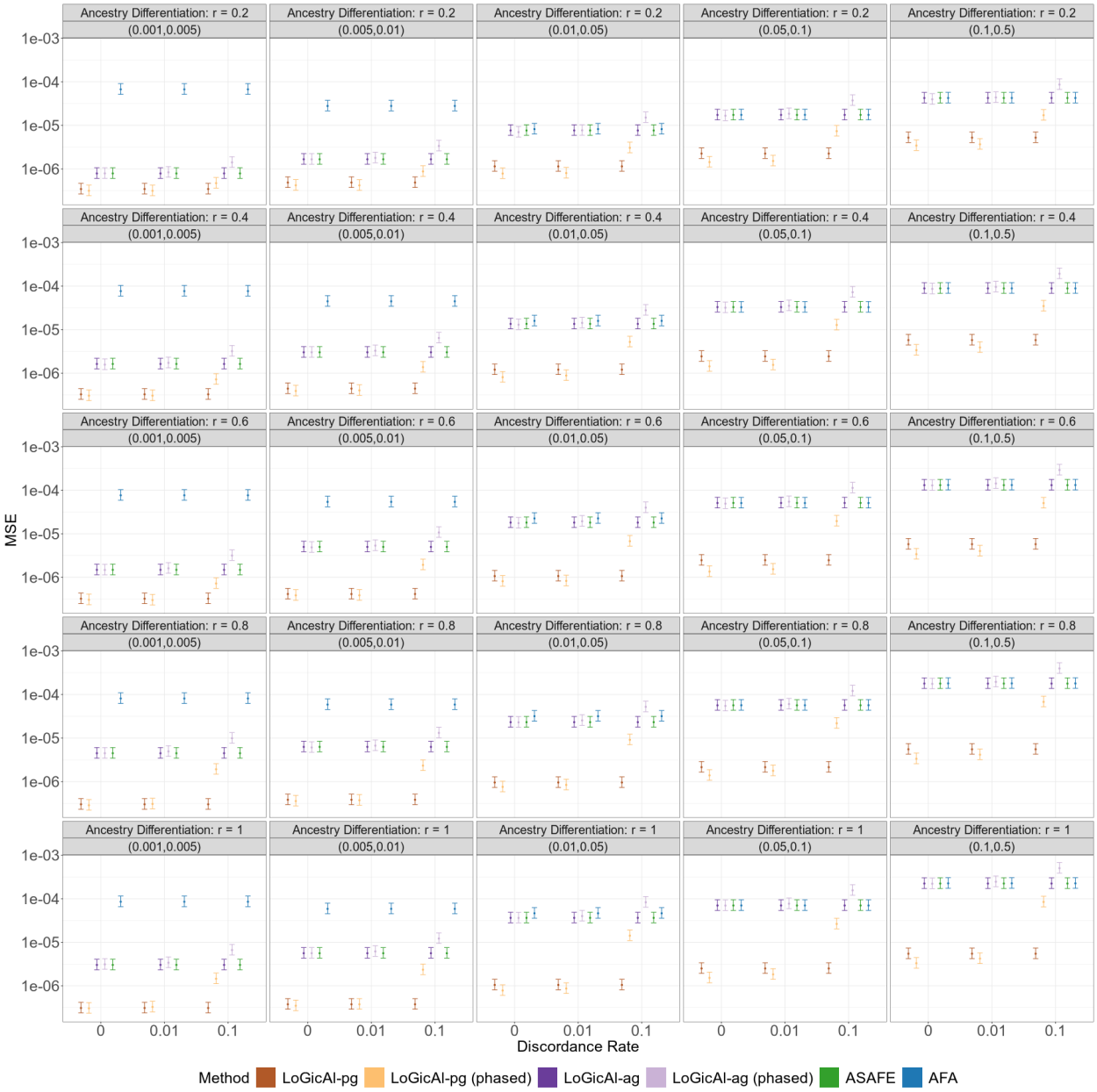

Figure S8: Mean squared error of ancestry-specific allele frequency estimation methods across discordance rates, allele frequency bins and levels of ancestral population differentiation. We simulated 100,000 phased individuals from a three-way admixture model under a scenario with probabilistic ancestry calling ($\alpha = 0.1$) and a realistic genotyping error rate ($\theta = 5 \times 10^{-4}$). Methods compared include LoGicAl (LoGicAl-pg, LoGicAl-pg (phased), LoGicAl-ag, LoGicAl-ag (phased)), ASAFE and AFA. Points denote replicate-averaged MSEs, and intervals indicate 95% confidence intervals for MSEs based on chi-squared approximation.

Figure S9: Relative efficiency of ancestry-specific allele frequency estimation methods across discordance rates, allele frequency bins and levels of ancestral population differentiation. We simulated 100,000 phased individuals from a three-way admixture model under a scenario with probabilistic ancestry calling ($\alpha = 0.1$) and a realistic genotyping error rate ($\theta = 5 \times 10^{-4}$). Methods compared include LoGicAl (LoGicAl-pg, LoGicAl-pg (phased), LoGicAl-ag, LoGicAl-ag (phased)), ASAFE and AFA. Bars denote relative efficiencies relative to the optimal estimator.

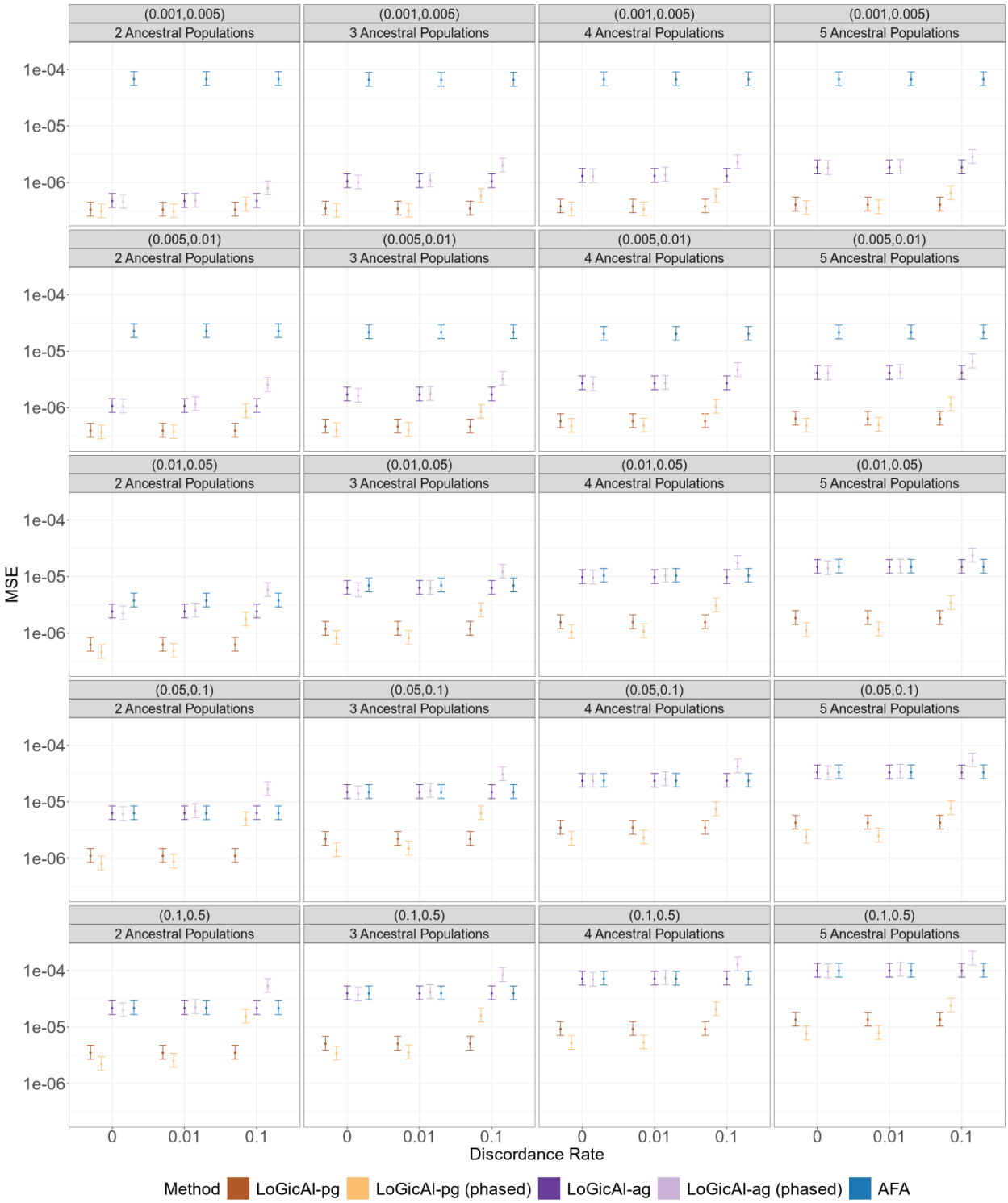

Figure S10: Mean squared error of ancestry-specific allele frequency estimation methods across discordance rates, allele frequency bins and numbers of ancestral populations. We simulated $100,000$ phased individuals under two- to five-way admixture with intra-continental differentiation. Methods compared include LoGicAl (LoGicAl-pg, LoGicAl-pg (phased), LoGicAl-ag, LoGicAl-ag (phased)) and AFA under a scenario with probabilistic ancestry calling ($\alpha = 0.1$) and a realistic genotyping error rate ($\theta = 5 \times 10^{-4}$). ASAFE is not included as it is developed for three-way (and two-way) admixture scenario only. Points denote replicate-averaged MSEs, and intervals indicate 95% confidence intervals for MSEs based on chi-squared approximation.

Figure S11: Relative efficiency of ancestry-specific allele frequency estimation methods across discordance rates, allele frequency bins and numbers of ancestral populations. We simulated 100,000 phased individuals under two- to five-way admixture with intra-continental differentiation. Methods compared include LoGicAl (LoGicAl-pg, LoGicAl-pg (phased), LoGicAl-ag, LoGicAl-ag (phased)) and AFA under a scenario with probabilistic ancestry calling ($\alpha = 0.1$) and a realistic genotyping error rate ($\theta = 5 \times 10^{-4}$). ASAFE is not included as it is developed for three-way (and two-way) admixture scenario only. Bars denote relative efficiencies relative to the optimal estimator.

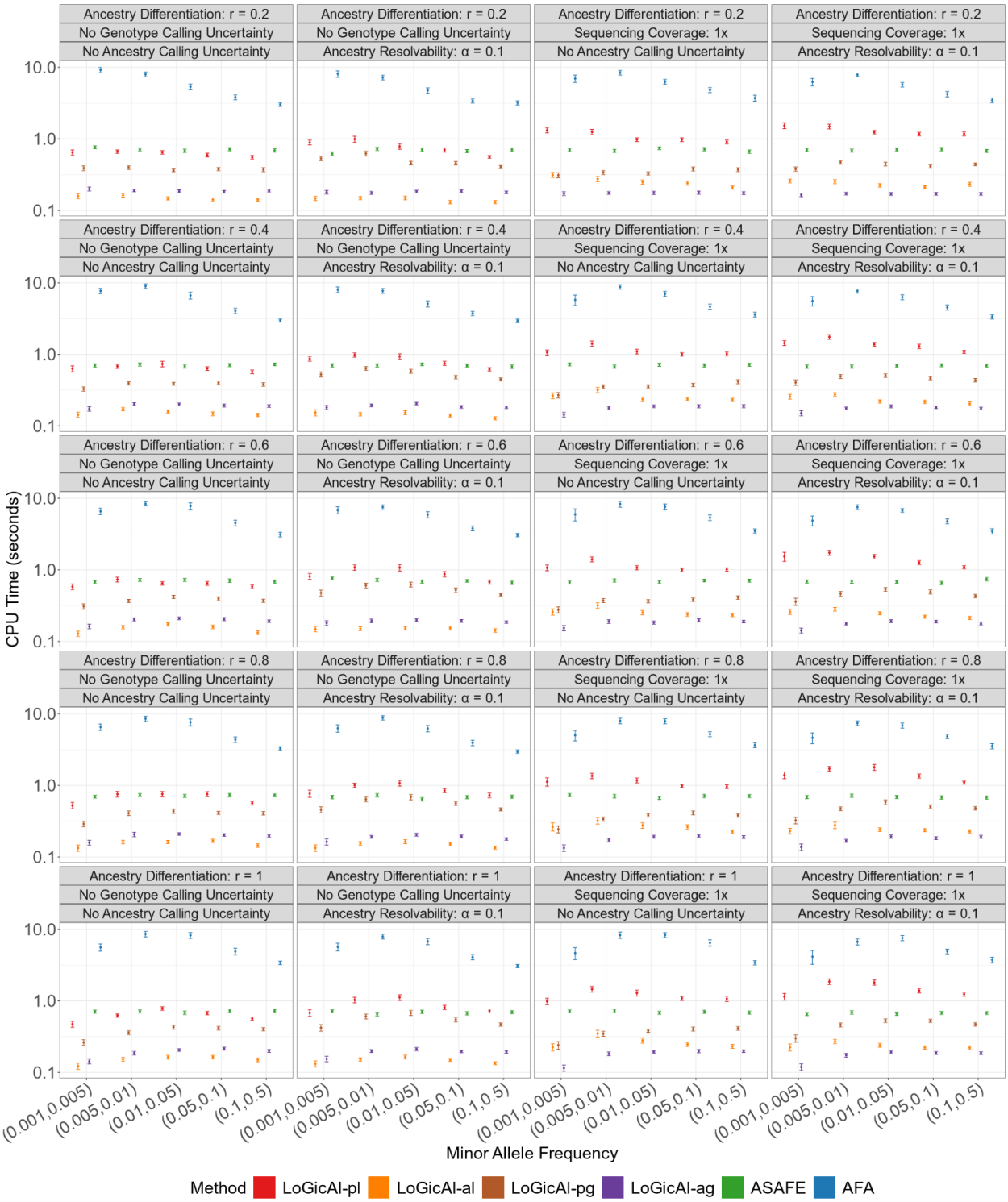

Figure S12: CPU runtime (seconds) of ancestry-specific allele frequency estimation methods across uncertainty scenarios, allele frequency bins and levels of ancestral population differentiation. We simulated 100,000 unphased individuals from a three-way admixture model with intra-continental differentiation ($r = 0.2$) to inter-continental differentiation ($r = 1$). Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag), ASAFE, and AFA under four settings: no calling uncertainty, ancestry calling uncertainty only, genotype calling uncertainty only, and both ancestry and genotype calling uncertainty. Points denote replicate-averaged runtime, and intervals indicate 95% confidence intervals for runtime based on normal approximation.

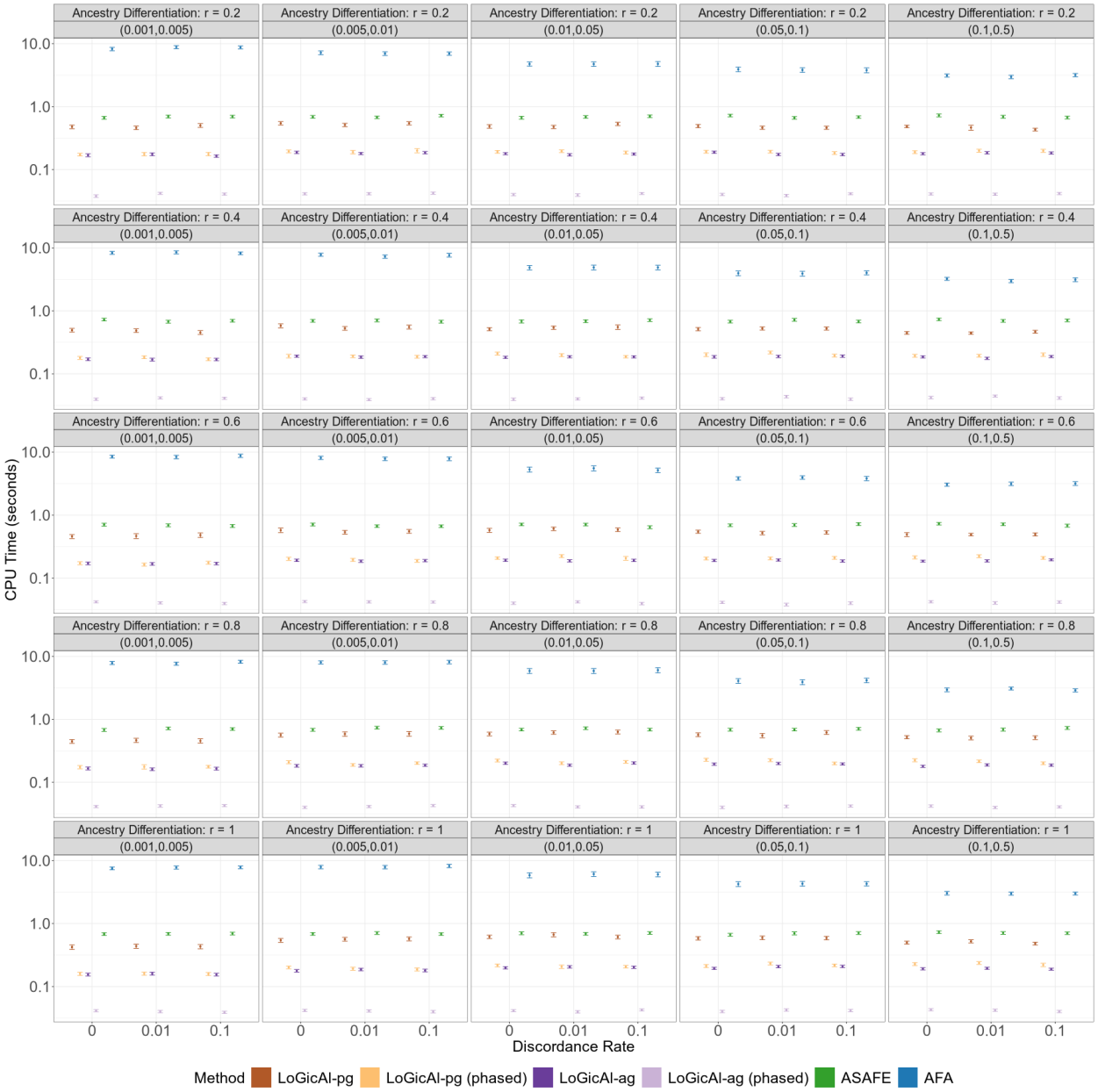

Figure S13: CPU runtime (seconds) of ancestry-specific allele frequency estimation methods across discordance rates, allele frequency bins and levels of ancestral population differentiation. We simulated 100,000 phased individuals from a three-way admixture model under a scenario with probabilistic ancestry calling ($\alpha = 0.1$) and a realistic genotyping error rate ($\theta = 5 \times 10^{-4}$). Methods compared include LoGicAl (LoGicAl-pg, LoGicAl-pg (phased), LoGicAl-ag, LoGicAl-ag (phased)), ASAFE and AFA. Points denote replicate-averaged runtime, and intervals indicate 95% confidence intervals for runtime based on normal approximation.

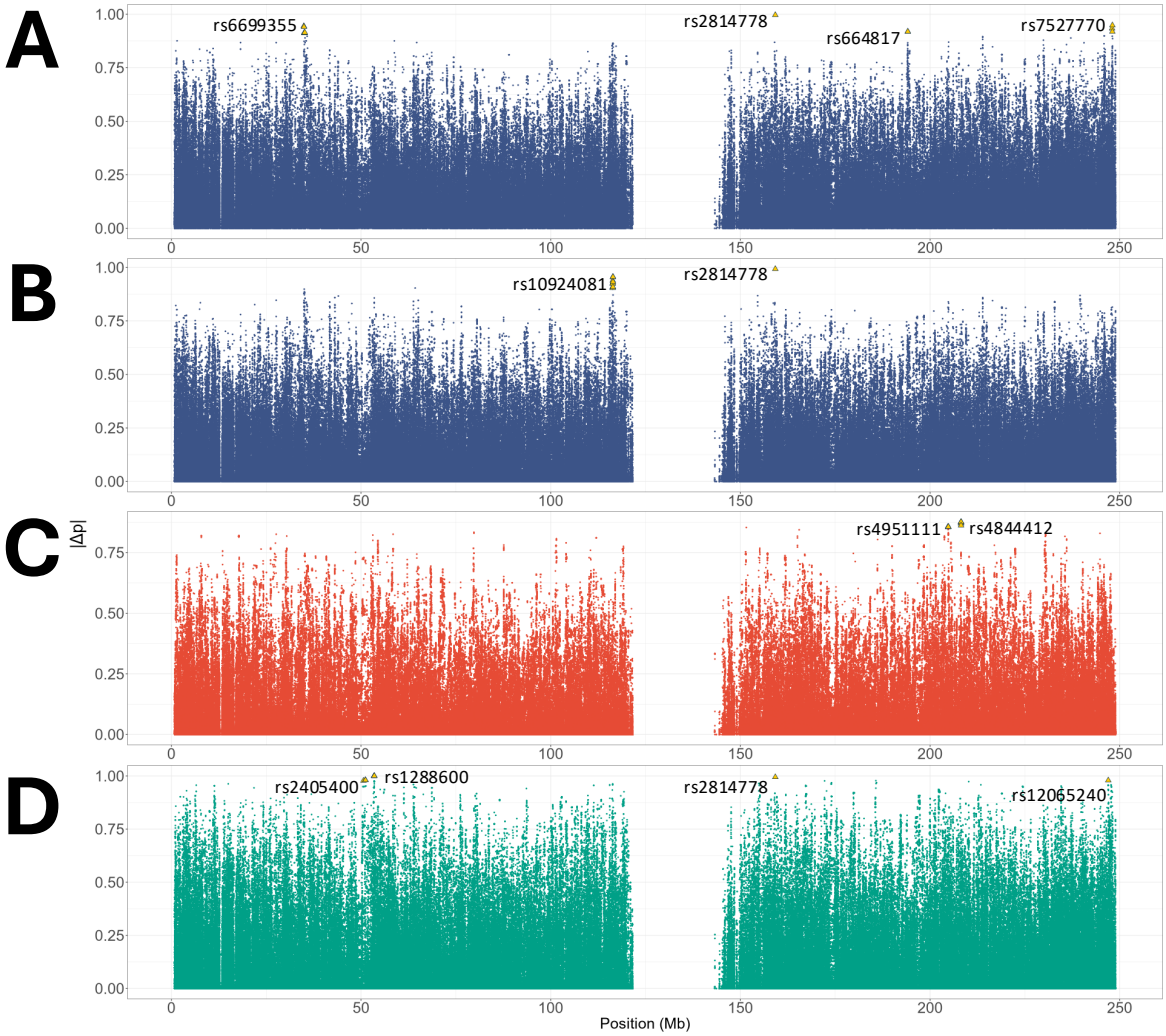

Figure S14: Absolute differences of ancestry-specific allele frequency estimates across chromosome 1. (A) Absolute differences between African and European ancestral estimates in 1kGP admixed African cohorts. (B) Absolute differences between African and European ancestral estimates in 1kGP admixed American cohorts. (C) Absolute differences between Native American and European ancestral estimates in 1kGP admixed American cohorts. (D) Absolute differences between Native American and African ancestral estimates in 1kGP admixed American cohorts. Outliers (yellow) denote the top 10 most differentiated variants per comparison; among these, the most differentiated variant within 1 Mb window is labeled by its rsID.

Supplemental tables

Table S1: Dirichlet regression estimates of resolvability of local ancestry assignment across 1000 Genomes Project individuals. We fitted an intercept-only Dirichlet regression model to the posterior probabilities of assigning each ancestral superpopulation generated by RFMix, using the `DirichReg()` function from the `DirichletReg` R package. We calculated the estimated α parameters as the exponential of the coefficients corresponding to each superpopulation from the fitted model.

Superpopulation	AFR	SAS	EAS	EUR	AMR	OCN	MES
$\hat{\alpha}$	0.118	0.107	0.110	0.118	0.094	0.092	0.092

AFR: Sub-Saharan Africa, SAS: Central & South Asia, EAS: East Asia, EUR: Europe, AMR: Native America, OCN: Oceania, MES: Middle East.

Table S2: CPU runtime (seconds) of ancestry-specific allele frequency estimation methods across allele frequency bins and uncertainty scenarios. We simulated $100,000$ unphased individuals from a three-way admixture model with inter-continental differentiation ($r = 1$). Methods compared include LoGicAl (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag), ASAFE, and AFA under four settings: (O) no calling uncertainty, (A) ancestry calling uncertainty only, (G) genotype calling uncertainty only, and (A+G) both ancestry and genotype calling uncertainty. Runtimes are reported for a single variant and are averaged over 100 simulation replicates.

Allele Frequency	Scenario	LoGicAl-pl	LoGicAl-al	LoGicAl-pg	LoGicAl-ag	ASAFE	AFA
(0.001,0.005)	O	0.47	0.12	0.26	0.14	0.71	5.59
	A	0.68	0.13	0.42	0.15	0.72	5.68
	G	0.98	0.23	0.24	0.12	0.72	4.65
	A+G	1.14	0.22	0.30	0.12	0.65	4.15
(0.005,0.01)	O	0.62	0.15	0.36	0.19	0.71	8.59
	A	1.03	0.15	0.61	0.20	0.65	7.92
	G	1.46	0.35	0.35	0.18	0.72	8.30
	A+G	1.85	0.27	0.46	0.17	0.69	6.70
(0.01,0.05)	O	0.78	0.16	0.43	0.21	0.68	8.23
	A	1.11	0.16	0.68	0.21	0.70	6.78
	G	1.29	0.28	0.38	0.19	0.68	8.33
	A+G	1.80	0.24	0.53	0.19	0.66	7.55
(0.05,0.1)	O	0.68	0.16	0.41	0.22	0.73	4.91
	A	0.81	0.15	0.55	0.20	0.67	4.08
	G	1.08	0.25	0.40	0.20	0.70	6.47
	A+G	1.39	0.22	0.53	0.19	0.68	4.91
(0.1,0.5)	O	0.56	0.15	0.40	0.20	0.72	3.40
	A	0.73	0.13	0.47	0.19	0.69	3.06
	G	1.07	0.23	0.41	0.20	0.68	3.41
	A+G	1.24	0.22	0.47	0.19	0.68	3.74

Table S3: CPU runtime (seconds) of ancestry-specific allele frequency estimation methods across allele frequency bins and discordance rates. We simulated $100,000$ phased individuals from a three-way admixture model with inter-continental differentiation ($r = 1$). Methods compared include LoGicAl (LoGicAl-pg, LoGicAl-pg (phased), LoGicAl-ag, LoGicAl-ag (phased)), ASAFE, and AFA under a scenario with probabilistic ancestry calling ($\alpha = 0.1$) and a realistic genotyping error rate ($\theta = 5 \times 10^{-4}$). Runtimes are reported for a single variant and are averaged over 100 simulation replicates.

Allele Frequency	Discordance	LoGicAl-pg	LoGicAl-pg (phased)	LoGicAl-ag	LoGicAl-ag (phased)	ASAFE	AFA
(0.001,0.005)	0	0.42	0.16	0.16	0.04	0.68	7.53
	0.01	0.44	0.16	0.16	0.04	0.69	7.76
	0.1	0.43	0.16	0.15	0.04	0.69	7.81
(0.005,0.01)	0	0.54	0.20	0.18	0.04	0.69	7.90
	0.01	0.56	0.19	0.19	0.04	0.71	7.88
	0.1	0.57	0.19	0.18	0.04	0.68	8.21
(0.01,0.05)	0	0.62	0.22	0.20	0.04	0.70	5.86
	0.01	0.67	0.21	0.21	0.04	0.69	6.11
	0.1	0.61	0.21	0.20	0.04	0.71	6.04
(0.05,0.1)	0	0.58	0.21	0.19	0.04	0.66	4.24
	0.01	0.59	0.23	0.21	0.04	0.70	4.30
	0.1	0.59	0.22	0.21	0.04	0.71	4.29
(0.1,0.5)	0	0.50	0.23	0.19	0.04	0.73	3.04
	0.01	0.52	0.24	0.19	0.04	0.71	2.99
	0.1	0.48	0.22	0.19	0.04	0.70	3.00

Table S4: Allele frequency estimates for LoGicAl-identified strongest ancestry-differentiated variants on chromosome 1 in 1kGP admixed American cohorts.

		LoGicAl			gnomAD			
rsID	Gene Region	AFR	EUR	NA	AFR	EUR	AMR	Ancestry Pair
$rs2814778$	$ACKR1$	> 0.999	0.007	0.005	0.822	0.004	0.090	AFR-EUR
$rs4844412$	$PLXNA2$	0.221	0.108	0.985	0.145	0.151	0.378	NA-EUR
$rs1288600$	$SLC25A3P1$	> 0.999	0.314	< 0.001	0.858	0.273	0.275	NA-AFR

AFR: Sub-Saharan Africa, EUR: Europe, NA: Native America, AMR: Admixed America.

Supplemental notes

Note S1: Derivation of LoGicAl algorithm

Following the notations in section (2.2), the complete-data log-likelihood of ancestry-specific allele frequencies is

$$
\begin{aligned}
&l(\boldsymbol{p}_{K\times 1}|\boldsymbol{x}_{N\times 1},\boldsymbol{\pi}_{N\times 2K};\boldsymbol{z}_{N\times 2},\boldsymbol{h}_{N\times 2}) = \log\Big\{f(\boldsymbol{x}|\boldsymbol{h})\times f(\boldsymbol{h}|\boldsymbol{z};\boldsymbol{p})\times f(\boldsymbol{z}|\boldsymbol{\pi})\Big\}\\
=&\sum_{i=1}^{N}\log\mathbb{P}\mathrm{r}(X_i=x_i|H_i^{(1)}=h_i^{(1)},H_i^{(2)}=h_i^{(2)})+\sum_{i=1}^{N}\sum_{a=1}^{2}\Big(h_i^{(a)}\log p_{z_i^{(a)}}+(1-h_i^{(a)})\log(1-p_{z_i^{(a)}})\Big)+\\
&\sum_{i=1}^{N}\sum_{a=1}^{2}\log\pi^{(a)}_{i,z_i^{(a)}}
\end{aligned}
\tag{4}
$$

The joint conditional distribution of latent variables $(Z_i^{(1)},Z_i^{(2)},H_i^{(1)},H_i^{(2)})$ given the observed data and ancestry-specific allele frequencies is

$$
\begin{aligned}
&w_i(z^{(1)},z^{(2)},h^{(1)},h^{(2)}) = \mathbb{P}\mathrm{r}(Z_i^{(1)}=z^{(1)},Z_i^{(2)}=z^{(2)},H_i^{(1)}=h^{(1)},H_i^{(2)}=h^{(2)}|x_i,\boldsymbol{\pi}_i;\boldsymbol{p})\\
=&\frac{\mathbb{P}\mathrm{r}(X_i=x_i|H_i^{(1)}=h^{(1)},H_i^{(2)}=h^{(2)})p_{z^{(1)}}^{h^{(1)}}(1-p_{z^{(1)}})^{1-h^{(1)}}p_{z^{(2)}}^{h^{(2)}}(1-p_{z^{(2)}})^{1-h^{(2)}}\pi^{(1)}_{i,z^{(1)}}\pi^{(2)}_{i,z^{(2)}}}{\sum_{u_1=0}^{1}\sum_{u_2=0}^{1}\sum_{v_1=1}^{K}\sum_{v_2=1}^{K}\mathbb{P}\mathrm{r}(X_i=x_i|H_i^{(1)}=u_1,H_i^{(2)}=u_2)p_{v_1}^{u_1}(1-p_{v_1})^{1-u_1}p_{v_2}^{u_2}(1-p_{v_2})^{1-u_2}\pi^{(1)}_{i,v_1}\pi^{(2)}_{i,v_2}}
\end{aligned}
\tag{5}
$$

Therefore, the expected log-likelihood of ancestry-specific allele frequencies is

$$
\begin{aligned}
&Q(\boldsymbol{p}|\hat{\boldsymbol{p}}^{(t)}) = \mathbb{E}[l(\boldsymbol{p}|\boldsymbol{x},\boldsymbol{\pi};\boldsymbol{z},\boldsymbol{h})|\boldsymbol{x},\boldsymbol{\pi};\hat{\boldsymbol{p}}^{(t)}]\\
=&\sum_{i=1}^{N}\sum_{z^{(1)}=1}^{K}\sum_{z^{(2)}=1}^{K}\sum_{h^{(1)}=0}^{1}\sum_{h^{(2)}=0}^{1}w_i^{(t)}(z^{(1)},z^{(2)},h^{(1)},h^{(2)})\Big[\log\mathbb{P}\mathrm{r}(X_i=x_i|H_i^{(1)}=h^{(1)},H_i^{(2)}=h^{(2)})+\\
&\sum_{a=1}^{2}\big(h^{(a)}\log p_{z^{(a)}}+(1-h^{(a)})\log(1-p_{z^{(a)}})\big)+\sum_{a=1}^{2}\log\pi^{(a)}_{i,z^{(a)}}\Big]
\end{aligned}
\tag{6}
$$

Taking derivative for each element p_k in $\boldsymbol{p}$

$$
\begin{aligned}
&\frac{\partial Q(\boldsymbol{p}|\hat{\boldsymbol{p}}^{(t)})}{\partial p_k}\\
=&\sum_{i=1}^{N}\sum_{h^{(1)}=0}^{1}\sum_{h^{(2)}=0}^{1}\frac{\sum_{z^{(2)}=1}^{K}w_i^{(t)}(k,z^{(2)},h^{(1)},h^{(2)})h^{(1)}+\sum_{z^{(1)}=1}^{K}w_i^{(t)}(z^{(1)},k,h^{(1)},h^{(2)})h^{(2)}}{p_k}-\\
&\sum_{i=1}^{N}\sum_{h^{(1)}=0}^{1}\sum_{h^{(2)}=0}^{1}\frac{\sum_{z^{(2)}=1}^{K}w_i^{(t)}(k,z^{(2)},h^{(1)},h^{(2)})(1-h^{(1)})+\sum_{z^{(1)}=1}^{K}w_i^{(t)}(z^{(1)},k,h^{(1)},h^{(2)})(1-h^{(2)})}{1-p_k}
\end{aligned}
\tag{7}
$$

By setting $\frac{\partial Q(\boldsymbol{p}|\hat{\boldsymbol{p}}^{(t)})}{\partial \boldsymbol{p}}=\mathbf{0}$, the expected log-likelihood $Q(\boldsymbol{p}|\hat{\boldsymbol{p}}^{(t)})$ is maximized at

$$
\hat{p}_k^{(t+1)} = \frac{\sum_{i=1}^{N}\big[\sum_{h^{(1)}=0}^{1}\sum_{z^{(1)}=1}^{K}w_i^{(t)}(z^{(1)},k,h^{(1)},1)+\sum_{h^{(2)}=0}^{1}\sum_{z^{(2)}=1}^{K}w_i^{(t)}(k,z^{(2)},1,h^{(2)})\big]}{\sum_{i=1}^{N}\sum_{h^{(1)}=0}^{1}\sum_{h^{(2)}=0}^{1}\big[\sum_{z^{(1)}=1}^{K}w_i^{(t)}(z^{(1)},k,h^{(1)},h^{(2)})+\sum_{z^{(2)}=1}^{K}w_i^{(t)}(k,z^{(2)},h^{(1)},h^{(2)})\big]}
\tag{8}
$$

Note S2: Simulation study details

We simulated ancestry-specific allele frequencies following the Balding-Nichols model[32] assuming each ancestral population descends independently from a common reference population. Under this model, the allele frequency p_k of the ancestral population k with reference allele frequency p_A is

$$p_k \sim Beta(p_A \frac{1-F_k}{F_k}, (1-p_A)\frac{1-F_k}{F_k}), k = 1, \ldots, K \tag{9}$$

where F_k is Wright's F-statistic[48, 49] measuring the genetic differentiation between the reference population and ancestral population k. Specifically, we simulated reference allele frequency p_A from a uniform distribution $p_A \sim U(a, b)$ with varying (a, b) from $(0.001, 0.005)$ to $(0.1, 0.5)$ to generate allele frequency spectrum from rare to common variants. Based on the simulated p_A, we sampled true ancestry-specific allele frequencies p_k from the distribution (9). Here, to generate ancestral population differentiation, we adopted the design of [19]: we first fixed the total number of ancestral populations at $K = 3$ and set $(F_1, F_2, F_3) = (0.01r, 0.05r, 0.1r)$, where $r \in \{0.2, 0.4, 0.6, 0.8, 1\}$ is a scaling parameter controlling the degrees of population differentiation. This approach has been previously evaluated by [33, 50]. Larger r describes more rapid genetic drift rate from the reference population, making the descending populations more distinct from each other: setting $r = 1$ results in allele frequency differentiation observed between populations across different continents (e.g., between San, Han, and Maya populations), whereas setting $r = 0.2$ corresponds to differentiation between populations on the same continent[33]. Additionally, we varied the number of ancestral populations from $K = 2$ to $K = 5$ and set $F_k = F, k = 1, \ldots, K$, with $F = 0.01$ to simulate weak population differentiation and $F = 0.05$ to simulate strong population differentiation.

To simulate true underlying local ancestries and their assignment probabilities, we sampled the assignment probabilities of ancestral groups for each individual independently from a Dirichlet distribution $\boldsymbol{\pi}_i \sim Dir(\alpha \mathbb{1}_K)$, where α quantifies the resolvability of local ancestry inference and we fixed $\alpha = 0.1$ as estimated from the 1000 Genomes Project cohort (Table S1). Given the simulated assignment probabilities, we generated true local ancestry of each allele a from a multinoulli distribution $z_i^{(a)} \sim Multinoulli(\boldsymbol{\pi}_i), a = 1, 2$, while generating the best-guess local ancestry call as the ancestral group where the assignment probability is maximized.

After obtaining the simulated ancestry-specific allele frequencies $\mathbf{p}$ and true local ancestries $\mathbf{z^{(a)}}, a = 1, 2$, we sampled true genotype per chromosome for each individual independently from a Bernoulli distribution $h_i^{(a)} \sim Bernoulli(p_{z_i^{(a)}}), a = 1, 2$. To imitate the effects of genotyping errors and statistical phasing uncertainty, we generated two types of observed genotype inputs: (1) genotype likelihoods representing probabilistic genotype information, and (2) best-guess genotype calls representing discreet genotype assignments.

We generated genotype likelihoods by adopting the framework in [51]. We simulated sequencing coverage for individual i from a Poisson distribution $d_i \sim Poisson(\lambda_i)$ after simulating average sequencing coverage from a truncated normal distribution $\lambda_i \sim TN(\mu, \mu/3), \lambda_i \in (0, 2\mu)$. Here, μ is average sequencing coverage across individuals and we chose $\mu \in \{1, 5, 30\}$ to model low-coverage to high-coverage sequencing data. We then sampled the number of sequencing reads presenting the alternate allele from a Binomial distribution $r_i \sim Bin(d_i, \frac{g_i}{2}(1-\epsilon)+(1-\frac{g_i}{2})\epsilon)$, where $g_i = h_i^{(1)} + h_i^{(2)}$ and ϵ is per-base sequencing error rate. We fixed $\epsilon = 0.001$ across all simulation settings. Therefore, the genotype likelihoods for reference homozygote, heterozygote, and alternate homozygote are $((1-\epsilon)^{d_i - r_i}\epsilon^{r_i}, 0.5^{d_i}, (1-\epsilon)^{r_i}\epsilon^{d_i - r_i})$, respectively. We used simulated overall allele frequency $\bar{p} = \frac{1}{K}\sum_{k=1}^{K} p_k$ as prior and called genotypes for each individual as maximum-a-posteriori estimates based on the posterior probabilities $((1-\bar{p})^2 \times (1-\epsilon)^{d_i - r_i}\epsilon^{r_i}, 2\bar{p}(1-\bar{p}) \times 0.5^{d_i}, \bar{p}^2 \times (1-\epsilon)^{r_i}\epsilon^{d_i - r_i})$.

In parallel, to simulate genotype calls as observed genotype inputs, we generated observed genotype per chromosome from a Bernoulli distribution $o_i^{(a)} \sim Bern(h_i^{(a)}(1-\theta) + (1-h_i^{(a)})\theta), a = 1, 2$, where θ is genotyping error rate and we fixed $\theta = 5 \times 10^{-4}$ as estimated from previous literature[38] to mimic error rates from either array-based (e.g., UK Biobank[52]) or sequence-based (e.g., TOPMed[45]) genotype calling. To simulate observed phased genotypes, we exchanged the values of $o_i^{(1)}$ and $o_i^{(2)}$ with probability γ, generating observed phase $(o_i^{(2)}, o_i^{(1)})$. γ is a compound discordance rate measuring deviations of observed ancestry-genotype phase from true ancestry-genotype phase. These deviations can arise from switch errors and double

switch errors (i.e., flip errors) in statistical phasing[38] and inferred ancestry segment flips[39]. Following reported values[44], we varied $\gamma \in \{0, 1\%, 10\%\}$.

References

[1] Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics* **5**, e1000695 (2009).

[2] Myles, S., Davison, D., Barrett, J., Stoneking, M. & Timpson, N. Worldwide population differentiation at disease-associated snps. *BMC medical genomics* **1**, 1–10 (2008).

[3] Itan, Y., Jones, B. L., Ingram, C. J., Swallow, D. M. & Thomas, M. G. A worldwide correlation of lactase persistence phenotype and genotypes. *BMC evolutionary biology* **10**, 1–11 (2010).

[4] Palmer, C. & Pe'er, I. Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLoS genetics* **13**, e1006916 (2017).

[5] Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature communications* **11**, 3865 (2020).

[6] Saitou, M., Dahl, A., Wang, Q. & Liu, X. Allele frequency impacts the cross-ancestry portability of gene expression prediction in lymphoblastoid cell lines. *The American Journal of Human Genetics* **111**, 2814–2825 (2024).

[7] Korunes, K. L. & Goldberg, A. Human genetic admixture. *PLoS genetics* **17**, e1009374 (2021).

[8] Nelson, S. C. *et al.* Social and scientific motivations to move beyond groups in allele frequencies: The topmed experience. *The American Journal of Human Genetics* **109**, 1582–1590 (2022).

[9] Montana, G. & Pritchard, J. K. Statistical tests for admixture mapping with case-control and cases-only data. *The American Journal of Human Genetics* **75**, 771–789 (2004).

[10] Mersha, T. B. Mapping asthma-associated variants in admixed populations. *Frontiers in genetics* **6**, 140765 (2015).

[11] Sofer, T. *et al.* Admixture mapping in the hispanic community health study/study of latinos reveals regions of genetic associations with blood pressure traits. *PLoS One* **12**, e0188400 (2017).

[12] Granot-Hershkovitz, E. *et al.* Interaction analysis of ancestry-enriched variants with apoe-4 on mci in the study of latinos-investigation of neurocognitive aging. *Scientific reports* **13**, 5114 (2023).

[13] Wang, Y. *et al.* Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. *Cell Genomics* **3** (2023).

[14] Chiu, A. M., Molloy, E. K., Tan, Z., Talwalkar, A. & Sankararaman, S. Inferring population structure in biobank-scale genomic data. *The American Journal of Human Genetics* **109**, 727–737 (2022).

[15] Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

[16] Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).

[17] Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **28**, 289–301 (2005).

[18] Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655–1664 (2009).

[19] Raj, A., Stephens, M. & Pritchard, J. K. faststructure: variational inference of population structure in large snp data sets. *Genetics* **197**, 573–589 (2014).

[20] Ko, S. *et al.* Unsupervised discovery of ancestry-informative markers and genetic admixture proportions in biobank-scale datasets. *The American Journal of Human Genetics* **110**, 314–325 (2023).

[21] Zhang, Q. S., Browning, B. L. & Browning, S. R. Asafe: ancestry-specific allele frequency estimation. *Bioinformatics* **32**, 2227–2229 (2016).

[22] Granot-Hershkovitz, E. *et al.* Afa: Ancestry-specific allele frequency estimation in admixed populations: The hispanic community health study/study of latinos. *Human Genetics and Genomics Advances* **3** (2022).

[23] Honorato-Mauer, J. *et al.* Characterizing features affecting local ancestry inference performance in admixed populations. *The American Journal of Human Genetics* (2025).

[24] Martin, A. R. *et al.* Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *The American Journal of Human Genetics* **108**, 656–668 (2021).

[25] Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics* **93**, 278–288 (2013).

[26] Browning, S. R., Waples, R. K. & Browning, B. L. Fast, accurate local ancestry inference with flare. *The American Journal of Human Genetics* **110**, 326–335 (2023).

[27] Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)* **39**, 1–22 (1977).

[28] Wu, C. J. On the convergence properties of the em algorithm. *The Annals of statistics* 95–103 (1983).

[29] Varadhan, R. & Roland, C. Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics* **35**, 335–353 (2008).

[30] Du, Y. & Varadhan, R. Squarem: An r package for off-the-shelf acceleration of em, mm and other em-like monotone algorithms. *Journal of Statistical Software* **92**, 1–41 (2020).

[31] Tang, B., Henderson, N. C. & Varadhan, R. Accelerating fixed-point algorithms in statistics and data science: A state-of-art review. *Journal of Data Science* **21** (2023).

[32] Balding, D. J. & Nichols, R. A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12 (1995).

[33] Kitada, S., Nakamichi, R. & Kishino, H. Understanding population structure in an evolutionary context: population-specific f st and pairwise f st. *G3* **11**, jkab316 (2021).

[34] Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).

[35] Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *science* **319**, 1100–1104 (2008).

[36] Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).

[37] Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nature genetics* **46**, 409–415 (2014).

[38] Browning, B. L. & Browning, S. R. Genotype error biases trio-based estimates of haplotype phase accuracy. *The American Journal of Human Genetics* **109**, 1016–1025 (2022).

[39] Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in gwas and to boost power. *Nature genetics* **53**, 195–204 (2021).

[40] Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

[41] Chat, V., Ferguson, R., Morales, L. & Kirchhoff, T. Ultra low-coverage whole-genome sequencing as an alternative to genotyping arrays in genome-wide association studies. *Frontiers in Genetics* **12**, 790445 (2022).

[42] Cavazos, T. B. & Witte, J. S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Human Genetics and Genomics Advances* **2** (2021).

[43] DeFelice, M. *et al.* Blended genome exome (bge) as a cost efficient alternative to deep whole genomes or arrays. *bioRxiv* 2024–04 (2024).

[44] Avadhanam, S. & Williams, A. L. Phase-free local ancestry inference mitigates the impact of switch errors on phase-based methods. *G3: Genes, Genomes, Genetics* jkaf122 (2025).

[45] Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature* **590**, 290–299 (2021).

[46] Hadley, T. J. & Peiper, S. C. From malaria to chemokine receptor: the emerging physiologic role of the duffy blood group antigen. *Blood, The Journal of the American Society of Hematology* **89**, 3077–3091 (1997).

[47] Shindo, H., Chigira, H., Tanaka, J., Kamatani, N. & Inoue, M. Grouping preprocess to accurately extend application of em algorithm to haplotype inference. *Journal of human genetics* **53**, 747–756 (2008).

[48] Wright, S. Coefficients of inbreeding and relationship. *The American Naturalist* **56**, 330–338 (1922).

[49] Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting f st. *Nature Reviews Genetics* **10**, 639–650 (2009).

[50] Goudet, J. & Weir, B. S. An allele-sharing, moment-based estimator of global, population-specific and population-pair f st under a general model of population structure. *PLoS genetics* **19**, e1010871 (2023).

[51] Meisner, J. & Albrechtsen, A. Inferring population structure and admixture proportions in low-depth ngs data. *Genetics* **210**, 719–731 (2018).

[52] Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).