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*]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA.

[2]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA.

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

Abstract

In genetic epidemiology and population genetic research of admixed populations, the allele frequencies across ancestral contributing populations, i.e., ancestry-specific allele frequencies, are a fundamental parameter, informing ancestry-enriched genetic drivers of disease etiologies, improving GWAS replication cohort design, enhancing polygenic risk prediction and portability, and providing insights into their demographic history. Current methods for estimating ancestry-specific allele frequencies typically rely on "best-guess" ancestry and genotype calls, thereby ignoring uncertainty from these upstream ancestry calling and genotyping procedures. Here, we introduce LoGicAl, a novel method for estimating ancestry-specific allele frequencies via accelerated expectation–maximization algorithm while simultaneously accounting for uncertainty from ancestry calling, genotyping, and statistical phasing to fine-tune its estimates. Simulation and real data applications demonstrate that ignoring these uncertainties inflates bias in allele frequency estimates and show that LoGicAl has superior accuracy and scalability in leveraging sequencing and array genotyping with different levels of local ancestry inference quality. Thus, LoGicAl can facilitate genomic analyses of admixed populations from biobank-scale data by providing precise ancestry-specific allele frequency estimates, which promote understandings of the 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 design[4], as well as portability of polygenic risk scores (PRS)[5, 6] in both single-ancestry populations and admixed populations.

However, the interpretation of allele frequencies is more complicated in admixed populations. Admixed populations arise when migration and gene flow occur between two or more previously isolated ancestral populations[7]. After generations of recombination, the genome of an admixed individual forms a mosaic structure of these ancestral populations. Disregarding this mosaic structure when estimating allele frequencies results in naive estimates that describe the patterns of genetic variation in the current admixed population. This naive estimate is a mixture of the allele frequencies of the ancestral populations weighted by the corresponding ancestral population contributions. Therefore, it creates challenges for interpreting admixed populations: as admixed population groups can differ regarding ancestral contributions, one simple allele frequency estimate is not transferable from one group to another. Moreover, it is often of interest to study pre-admixture ancestral populations from admixed populations. To overcome these challenges, ancestry-specific allele frequency has been suggested and broadly used in genetic analyses for admixed populations[8]. For example, to investigate ancestry-enriched genetic signals for complex traits[9], ancestry-specific allele frequencies have been used to identify ancestry-differentiated risk alleles for asthma[10], blood pressure[11], and mild cognitive impairment[12] in admixed populations. In GWAS design, integrating information from ancestry-specific allele frequency and matching it across corresponding ancestry cohorts can guide replication design for validating the discovered ancestry-enriched risk variants and improve polygenic prediction accuracy for admixed populations[13]. Moreover, when inferring population structure from admixed populations, incorporating reliable reference ancestry-specific allele frequencies can benefit statistical and computational performance of the inference tools compared with jointly modeling admixture proportions and allele frequencies[14]. The performance of all these approaches depends on the precision of the estimated ancestry-specific allele frequencies.

Several methods have been proposed to deconstruct allele frequencies of the underlying ancestral populations for admixed populations. Some methods primarily aim to infer population structure, usually 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 frequency[21, 22]. These methods integrate information from locus-specific ancestry background (i.e., local ancestry): ASAFE[21] estimates allele frequencies through expectation-maximization (EM) framework by taking inferred local ancestries and called genotypes from upstream analyses as inputs; AFA[22] provides ancestry-specific allele frequency estimates by modeling the weighted average of ancestry-specific allele frequencies given inferred local (or global) ancestry proportions. For computational efficiency, these local-ancestry-aware methods estimate

ancestry-specific allele frequency based on called (i.e., "best-guess") local ancestries and genotypes, thereby simply ignoring the uncertainty in the upstream local ancestry and genotype calling steps. These methods can provide relatively accurate ancestry-specific allele frequency estimates when both ancestry calling and genotype calling generate high-quality output. However, either step can suffer from high uncertainty (or high error rates). For example, the local ancestry inference accuracy can be relatively low at Amerindigenous tracks in admixed American cohorts[23], and low-coverage sequencing used to discover novel variants from underrepresented populations[24] generates genotype calls with high variance. Directly using these low-quality upstream outputs in existing ancestry-specific allele frequency estimation methods can violate their model assumptions and may bias their estimates.

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 primarily an accelerated EM algorithm robust to any uncertainty from upstream local ancestry calling and genotype calling simultaneously. LoGicAl informs ancestry-specific allele frequency estimates by calculating and updating sample-wide allele dosages defined by integrating probabilistic representation of ancestry calls, genotype likelihoods over each possible allele frequency distributions. We further enhance its computational performance for phased inputs. Various variations of LoGicAl are derived to accommodate more flexible input formats. We demonstrate the accuracy and efficiency of LoGicAl and benchmark its performance compared with other state-of-arts 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 inference and genotyping to fine-tune the estimates (Figure 1). LoGicAl is a fully parametric likelihood-based model, using a hierarchical mixture structure to incorporate the uncertainty from upstream ancestry calling and genotype calling procedures. The complete-data likelihood motivated by this hierarchical mixture model is decomposed into conditional independent parts corresponding to different sources of uncertainty as well as statistical phasing. Maximum likelihood estimates are calculated and evaluated through accelerated EM algorithm. LoGicAl can accommodate either probabilistic presentations or best-guess calls for local ancestries and genotypes. An alternative version of the algorithm leverages phasing information to generate rapid allele frequency estimates.

2.2 Model Setup

Suppose we have N diploid admixed individuals genotyped at J bi-allelic sites. Each individual's genome originates from K ancestral population groups. We aim to estimate the ancestry-specific allele frequencies $\mathbf{p} = (p_1, ..., p_K)$ for these K ancestral

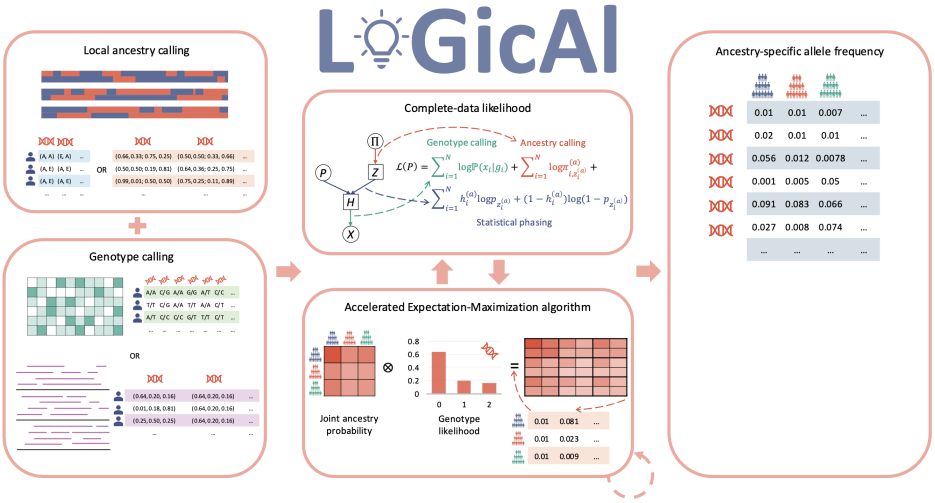

Fig. 1: Schematic overview of LoGicAl.

populations at each locus of interest (subscript j is omitted for simplicity). At each position, local ancestries are estimated using exiting inference tools from upstream analysis, such as RFMix[25] and FLARE[26]. These local ancestry inference programs usually output the most likely ancestry assignments and the corresponding probabilities of assigning each ancestral group. These assignment probabilities can be used to calculate the marginal probabilities of ancestral group 1 to K for individual i at site j, indicated by $\boldsymbol{\pi}_i = (\pi_{i,1}^{(1)}, ..., \pi_{i,K}^{(1)}; \pi_{i,1}^{(2)}, ..., \pi_{i,K}^{(2)})$. Assuming each individual is unrelated, the observed-data log-likelihood of the 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 the observed genotype call and G_i is the unobserved true genotype for individual i at the site.

2.3 Estimation Algorithm

LoGicAl computes maximum likelihood estimates of ancestry-specific allele frequencies, using an expectation-maximization (EM) algorithm[27] by introducing latent variables $(Z_i^{(1)}, Z_i^{(2)})$ and $(H_i^{(1)}, H_i^{(2)})$, representing true local ancestry and true genotype per chromosome, respectively. Assuming Hardy-Weinberg equilibrium within each ancestral group, LoGicAl estimates ancestry-specific allele frequencies via the following EM algorithm (see Supplemental information for details):

E-Step

$$Q(\mathbf{p}|\mathbf{p}^{(t)}) = \sum_i \sum_{(\mathbf{z}_i,\mathbf{h}_i)} \sum_a \phi_i^{(t)}(\mathbf{z}_i,\mathbf{h}_i)(h_i^{(a)} \log(p_{z_i^{(a)}}) + (1-h_i^{(a)}) \log(1-p_{z_i^{(a)}})) \quad (2)$$

M-step

$$p_k^{(t+1)} = \frac{\sum_{i=1}^N \{\sum_{(z_i^{(1)},h_i^{(1)})} \phi_i^{(t)}(z_i^{(1)},k,h_i^{(1)},1) + \sum_{(z_i^{(2)},h_i^{(2)})} \phi_i^{(t)}(k,z_i^{(2)},1,h_i^{(2)})\}}{\sum_{i=1}^N \{\sum_{(z_i^{(1)},h_i^{(1)},h_i^{(2)})} \phi_i^{(t)}(z_i^{(1)},k,h_i^{(1)},h_i^{(2)}) + \sum_{(z_i^{(2)},h_i^{(1)},h_i^{(2)})} \phi_i^{(t)}(k,z_i^{(2)},h_i^{(1)},h_i^{(2)})\}}, \quad (3)$$

where

$$\phi_i(z^{(1)},z^{(2)},h^{(1)},h^{(2)}) = \frac{Pr(X_i=x_i|G_i=h^{(1)}+h^{(2)})\pi^{(1)}_{i,z^{(1)}}\pi^{(2)}_{i,z^{(2)}}(p_{z^{(1)}})^{h^{(1)}}(1-p_{z^{(1)}})^{1-h^{(1)}}(p_{z^{(2)}})^{h^{(2)}}(1-p_{z^{(2)}})^{1-h^{(2)}}}{\sum_{(\mathbf{z},\mathbf{h})} Pr(X_i=x_i|G_i=h^{(1)}+h^{(2)})\pi^{(1)}_{i,z^{(1)}}\pi^{(2)}_{i,z^{(2)}}(p_{z^{(1)}})^{h^{(1)}}(1-p_{z^{(1)}})^{1-h^{(1)}}(p_{z^{(2)}})^{h^{(2)}}(1-p_{z^{(2)}})^{1-h^{(2)}}} \quad (4)$$

2.4 Algorithm Implementation

The classical EM algorithm can suffer from slow convergence, especially when sample size is large or the uncertainty structure is complicated, limiting its applications to biobank-scale genomic data of admixed samples. Several accelerating techniques were proposed recently to tackle the computational burden of the fixed-point algorithms[28], including squared iterative methods (SQUAREM)[29, 30], Anderson acceleration and DAAREM[31], and quasi-Newton acceleration[32]. We investigated the empirical performance of these state-of-art methods (see Supplemental information), and incorporated SQUAREM acceleration in LoGicAl to enhance its computing performance.

2.5 Simulation Study

We designed simulation study to investigate estimation accuracy of LoGicAl and its variations across wide-range settings, and to benchmark with other state-of-arts tools (Figure 2). For this purpose, we first sample allele frequencies in the ancestral populations based on a population genetic model, then generate the distribution of underlying local ancestries in each individual. Based on these ancestries and simulated ancestry-specific allele frequencies, we sample the true genotypes in each individual. After specifying genotyping error models, we generate observed genotypes from the underlying true genotypes.

To generate the ancestry-specific allele frequencies, we follow the Balding-Nichols model[33]. The model assumes each ancestral population descends independently from a common reference population. Then the genetic differentiation between the reference

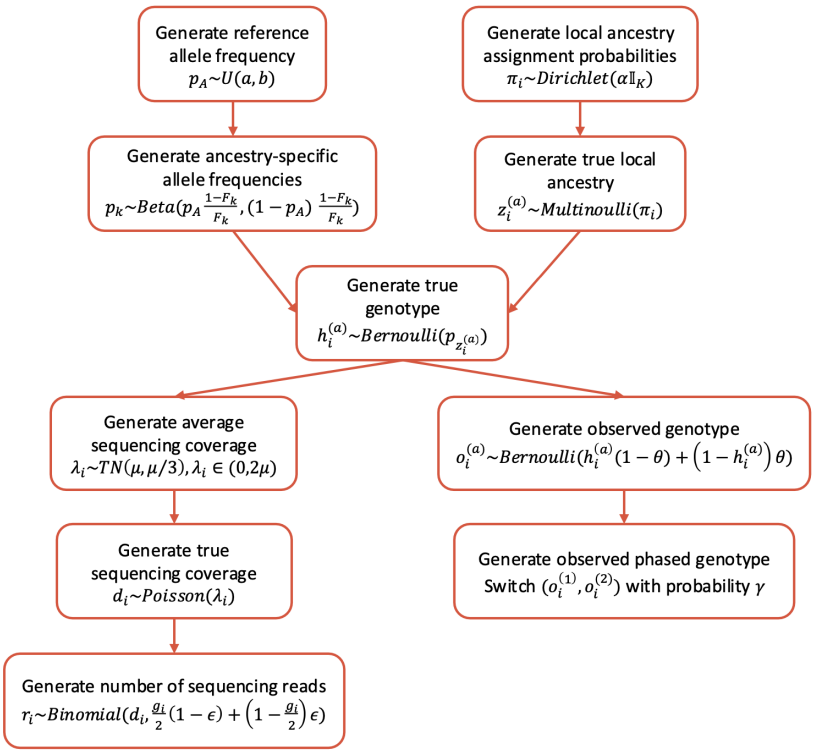

Fig. 2: Flow chart of simulation study design.

population and each ancestral population k is characterized by Wright's F-statistics F_k[34]: the allele frequency p_k of contributing population k at the site with reference allele frequency p_A is

$$p_k \sim Beta(p_A \frac{1-F_k}{F_k}, (1-p_A)\frac{1-F_k}{F_k}), k = 1, ..., K \tag{5}$$

Reference allele frequency p_A follows uniform distributions $p_A \sim U(a,b)$ with varying choices of (a,b) from $(5 \times 10^{-4}, 0.001)$ to $(0.4, 0.5)$ to simulate reference allele frequency spectrum from rare to common variants. Based on p_A, we sample true ancestry-specific allele frequencies p_k from equation (5). Specifically, to simulate the genetic differentiation across $K = 3$ different populations in humans, we set $(F_1, F_2, F_3) = (0.01r, 0.05r, 0.1r)$, where $r \in \{0.1, 0.2, ..., 1\}$ is a scaling parameter, as implemented[19] and evaluated[35][36] in past literature. Larger r describes more rapid genetic drift rate from the reference population, making the descending contributing populations more distinct. Setting $r = 1$ results in allele frequency differentiation observed between population on different continents (e.g., between Yoruba, Han, and

Maya populations[35]), whereas setting $r = 0.1$ corresponds to differentiation between populations on the same continent.

To generate the true underlying local ancestries and their corresponding assignment probabilities π, we sample the assignment probabilities of each ancestral group for each individual independently from a Dirichlet distribution $\boldsymbol{\pi_i} \sim Dir(\alpha \mathbb{1}_{\boldsymbol{K}})$, where $\alpha \in \{0.1, ..., 0.5\}$ is a scaling parameter describing the resolvability of local ancestry assignment: when α is close to 0, the local ancestry background is more distinguishable. Based on the assignment probabilities, we generate the true ancestry of each allele a from a multinoulli distribution $z_i^{(a)} \sim Multinoulli(\boldsymbol{\pi_i}), a = 1, 2$.

Combining the simulated ancestry-specific allele frequencies $\mathbf{p}$ and local ancestries $\mathbf{z^{(a)}}, a = 1, 2$, we sample true genotype per chromosome for each individual independently from Bernoulli distribution $h_i^{(a)} \sim Bernoulli(p_{z_i^{(a)}}), a = 1, 2$. To model genotyping errors and statistical phasing, we generate two types of observed genotype inputs: (1) genotype likelihoods, and (2) genotype calls (see Supplemental information for details).

2.6 Evaluation Metrics

We evaluate the estimation accuracy of LoGicAl using average root mean square error (RMSE) and statistical efficiency, and benchmark its performance against two existing methods, ASAFE[21] and AFA[22]. The RMSE compares the true ancestry-specific allele frequencies $\mathbf{p} = (p_1, ..., p_K)$ and their estimates $\hat{\mathbf{p}}$ as

$$RMSE(\mathbf{p}, \hat{\mathbf{p}}) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (p_k - \hat{p}_k)^2} \tag{6}$$

The statistical efficiency measures the precision of an estimator T by comparing the RMSE of the minimal-RMSE estimator against the RMSE of the estimator T

$$Eff(T) = \frac{RMSE(\mathbf{p}, \hat{\mathbf{p}}_{minimal})}{RMSE(\mathbf{p}, \hat{\mathbf{p}}_T)} \tag{7}$$

where the minimal-RMSE estimates are empirically calculated as the point estimates with known true ancestry, genotype, and ancestry-genotype phasing information; thus, they reflect the sampling variation only.

2.7 Application to 1000 Genomes Project Data

We estimate ancestry-specific allele frequencies from the 1000 Genomes project cohort[37] using several variations of LoGicAl. In this analysis, a total number of 2,504 unrelated individuals from 26 populations were generated at the New York Genome Center. With these pre-phased genotypes, we inferred the local ancestries for each individual using RFMix[25] (version 2) with the following option: $--node-size = 5$. We used the Human Genome Diversity Panel (HGDP) dataset[38] as reference panel for local ancestry calling and pre-processed the reference cohort in alignment with

the quality control steps described in past literature[39], ending up with 938 individuals and 639,958 autosomal SNPs. We condensed the 53 populations in the reference panel into seven super-populations: (1) Sub-Saharan Africa, (2) Central/South Asia, (3) East Asia, (4) Europe, (5) Native America, (6) Oceania and (7) Middle East. All seven super-populations were used for local ancestry inference over each individual in the 1000 Genomes Project cohort. After running RFMix, we applied LoGicAl-ag and LoGicAl-phase to calculate ancestry-specific allele frequencies by integrating the inferred local ancestry information and previously called genotypes. Ancestry-specific allele frequencies were estimated from (1) combined cohort of all 26 populations, (2) separate cohorts defined by super-populations, and (3) each population in the 1000 Genomes Project.

3 Results

We evaluated estimation accuracy and computational costs of LoGicAl and its variations (LoGicAl-pl, LoGicAl-al, LoGicAl-pg, LoGicAl-ag, and LoGicAl-phase; p: local ancestry assignment probability, a: best-guess local ancestry, l: genotype likelihood, g: best-guess genotype call) when estimating ancestry-specific allele frequencies through simulation and data applications to the 1000 Genomes Project cohort. We further benchmarked their performance against other state-of-arts estimation methods, i.e., ASAFE[21] and AFA[22].

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

To assess the impact of ancestry calling uncertainty and genotype calling uncertainty on different estimation methods, we vary the resolution of ancestry calling from (1) deterministic, i.e., no uncertainty in local ancestry inference, (2) distinguishable ($\alpha = 0.1$), and (3) ambiguous ($\alpha = 0.5$), as well as consider genotypes called from (1) hypothetical scenario with no genotyping uncertainty, (2) high-coverage sequencing ($30\times$), and (3) low-coverage sequencing ($1\times$). In this comparison, we generate types of ancestry and genotype information for $100,000$ individuals to approximate the sample size of real-world biobank-scale cohorts and apply aforementioned methods to estimate ancestry-specific allele frequencies from the simulated cohorts.

We first consider an ideal scenario when there is no uncertainty in either ancestry calling or genotype calling (Figure 3A). In spite of no ancestry and genotype calling uncertainty, unknown genotype phases can induce loss in efficiency compared to the optimal estimation which knows full information including genotype phasing. This explains why the efficiency is always smaller than 1 even in absence of uncertainty. Under this hypothetical scenario, all LoGicAl variations as well as ASAFE show precise estimation results across allele frequency bins with the highest efficiency of 0.9816 at the bin $(5 \times 10^{-4}, 0.001)$ and the lowest efficiency of 0.7515 at $(0.4, 0.5)$. That's because increasing allele frequency contributes to more heterozygotes, allowing for more possibility of containing unknown genotype phases and thereby reducing efficiency. While AFA achieves comparable estimation accuracy to LoGicAl and ASAFE

when estimating ancestry-specific allele frequencies for common variants, its accuracy decreases for rare variants.

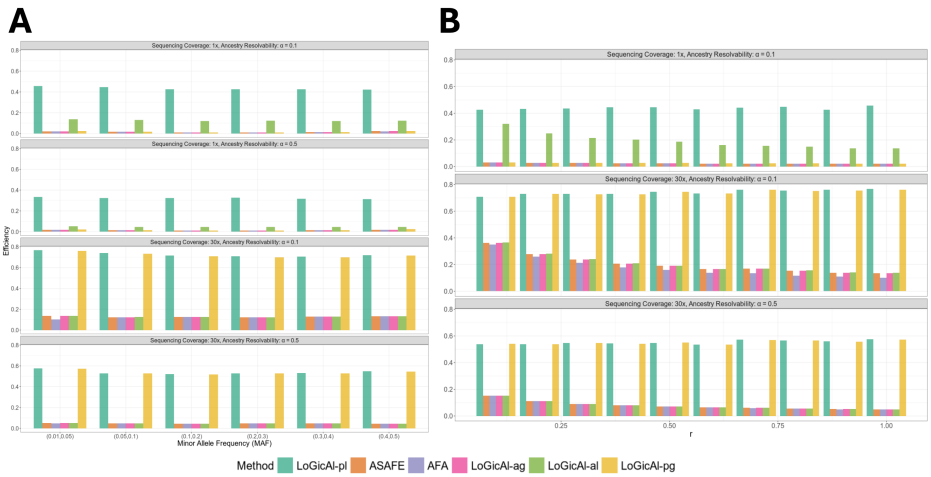

Fig. 3: Accuracy of LoGicAl, ASAFE and AFA in simulated dataset.

To evaluate the impact of ancestry calling uncertainty, we consider results with ancestry resolvability $\alpha = 0.1$ while assuming no genotype calling uncertainty (Figure 3A). This choice of ancestry resolvability parameter reflects the local ancestry posterior distribution in the 1000 Genomes Project (see Supplementary materials for detailed derivation). In this scenario, the efficiency of the methods accounting for ancestry calling uncertainty (LoGicAl-pl, LoGicAl-pg) is similar and ranges from 0.7081 to 0.9615 across allele frequency bins. Compared with the earlier scenario of no ancestry calling uncertainty, the median loss in efficiency of these two methods is 0.0643 across allele frequency bins. Methods which do not model ancestry inference uncertainty have uniformly lower efficiency: LoGicAl-al, LoGicAl-ag and ASAFE range from 0.1266 to 0.2362, while AFA ranges from 0.0038 to 0.1349. Compared with the earlier scenario of no ancestry calling uncertainty, the median loss in efficiency of LoGicAl-al, LoGicAl-ag and ASAFE is 0.6932 while the median loss of AFA is 0.4748. Consistent with the previous hypothetical scenario, AFA shows similar efficiency as LoGicAl-al, LoGicAl-ag and ASAFE for common variants but lower efficiency for rare variants. Compared with these ancestry-calling-uncertainty-unaware methods, LoGicAl-pl and LoGicAl-pg increase the statistical efficiency by four folds on average.

To evaluate the impact of genotype calling uncertainty, we compare the performance of each method in the scenarios when the genotypes are called through low-coverage ($1\times$) or high-coverage ($30\times$) sequencing in absence of local ancestry inference uncertainty (Figure 3A). When the average sequencing depth is $1\times$, the efficient

of the LoGicAl variations adjusting for genotyping uncertainty (LoGicAl-pl, LoGicAl-al) is similar and ranges from 0.4220 to 0.5155 across allele frequency bins. Compared with the no calling uncertainty scenario, the median loss in efficiency of these two methods is 0.3440. Methods which do not account for genotyping uncertainty have uniformly lower statistical efficiency: LoGicAl-pg, LoGicAl-ag as well as ASAFE range from 0.0105 to 0.0536 while AFA ranges from 0.0034 to 0.0210. Compared with the no calling uncertainty scenario, the median loss in efficiency of LoGicAl-pg, LoGicAl-ag and ASAFE is 0.8116 while the median loss of AFA is 0.5930. Again, AFA shows comparable efficiency as LoGicAl-pg, LoGicAl-ag and ASAFE for common variants but reduced efficiency for rare variants. Compared with these genotyping-uncertainty-unaware methods, LoGicAl-pl and LoGicAl-al improve the statistical efficiency by 22-fold to 29-fold on average. When the genotypes are called via high-coverage sequencing (30×), the estimation accuracy of each method is similar to the ideal scenario when there is no uncertainty in either genotype calling or ancestry calling.

With both uncertain local ancestry calls ($\alpha = 0.1$) and low-coverage sequencing (1×), LoGicAl-pl outperforms all other methods and its statistical efficiency ranges from 0.4129 to 0.4749 across allele frequency bins, which is about half of the efficiency for the same method in the hypothetical no uncertainty scenario. Methods which do not account for either ancestry calling uncertainty or genotyping uncertainty show uniformly lower efficiency: LoGicAl-al ranges from 0.1178 to 0.1758, LoGicAl-pg, LoGicAl-ag and ASAFE have similar efficiency and range from 0.0106 to 0.0399 while AFA ranges from 0.0031 to 0.0202. Compared with these uncertainty-unaware methods, the statistical efficiency of LoGicAl-pl is about twofold to fourfold of the efficiency of LoGicAl-al, about tenfold to 42-fold of the efficiency of LoGicAl-pg, LoGicAl-ag, ASAFE and about 21-fold to 135-fold of the efficiency of AFA.

3.2 Impact of ancestral population genetic differentiation

The precision of estimated ancestry-specific allele frequencies depends on the interplay of the uncertainty from the upstream analyses and the genetic differentiation between the ancestral contributing populations. In the ideal scenario of no ancestry calling uncertainty and no genotyping uncertainty (Figure 3B), the statistical efficiency of LoGicAl and ASAFE is similar and ranges from 0.7754 to 0.8284 across different degrees of ancestral population differentiation. AFA shows comparable estimation accuracy with LoGicAl and ASAFE at within-continental level differentiation while its efficiency decreases when estimating ancestry-specific allele frequencies from the admixed sample with cross-continental level ancestral population differentiation.

To assess the impact of ancestral population differentiation in presence of the uncertainty from the upstream analyses, we compare the estimation results under single source of uncertainty separately (Figure 3B). When the uncertainty only arises from upstream ancestry calling, the statistical efficiency of the methods adjusting for ancestry calling uncertainty (LoGicAl-pl, LoGicAl-pg) ranges from 0.6879 to 0.7477 across levels of ancestral population differentiation. Methods which do not account for ancestry calling uncertainty (LoGicAl-al, LoGicAl-ag, ASAFE, AFA) have uniformly lower efficiency compared with LoGicAl-pl and LoGicAl-pg. The efficiency of these ancestry-calling-uncertainty-unaware methods decreases from 0.3747

at intra-continental level differentiation ($r = 1$) to 0.1352 at inter-continental level differentiation ($r = 1$). Compared with these methods, LoGicAl-pl and LoGicAl-pg improve the efficiency by about threefold on average. When the uncertainty only arises from upstream genotype calling, the efficiency of the genotyping-uncertainty-aware methods (LoGicAl-pl, LoGicAl-al) ranges from 0.4644 to 0.5094 while the genotyping-uncertainty-unaware methods (LoGicAl-pg, LoGicAl-ag, ASAFE, AFA) decrease from 0.0201 to 0.0169 as ancestral populations are more divergent. Compared with these genotyping-uncertainty-unaware methods, LoGicAl-pl and LoGicAl-al improve the efficiency by about 22-fold to 29-fold.

3.3 Pre-phased genotypes

To evaluate the estimation accuracy of different methods when the input genotypes are pre-phased (Figure 3C), we consider the scenario with distinguishable ancestry calling ($\alpha = 0.1$) and realistic genotyping error rate ($\theta = 5 \times 10^{-4}$). We generate ancestry and genotype information for $100,000$ individuals and vary the discordance rate from 0 to 20%. We benchmark the performance of LoGicAl-phase and the LoGicAl variation which also takes best-guess ancestry and genotypes as inputs (LoGicAl-ag) against the state-of-arts methods (ASAFE and AFA).

When the discordance rate is 0, LoGicAl-ag, ASAFE and AFA show similar statistical efficiency of 0.1275 while LoGicAl-phase has slightly higher efficiency of 0.1286. When the discordance rate increases, the efficiency of these three phasing-adjusted methods (LoGicAl-ag, ASAFE, AFA) is 0.1289 for LoGicAl-ag and ASAFE and 0.1286 for AFA. The statistical efficiency is constant across different discordant rates. The statistical efficiency of LoGicAl-phase decreases with the highest of 0.1234 at the discordance rate of 0.01 and lowest of 0.0650 at 0.2. These patterns are consistent in the admixed samples with cross-continental level genetic differentiation among contributing populations as well as across varying levels of data quality with respect to ancestry calling and genotyping.

3.4 Applications to 1000 Genomes Project

We applied LoGicAl to estimate ancestry-specific allele frequencies from the admixed samples in the 1000 Genomes Project dataset. We estimated (1) the African and European ancestral allele frequencies in the admixed African cohorts (ASW and ACB cohorts), and (2) the African, European and Native American ancestral allele frequencies in the admixed American cohorts (CLM, MXL, PEL and PUR cohorts). In the admixed African cohorts, the ancestry-specific allele frequencies always show the highest correlation with the allele frequencies in their corresponding population groups compared with others in the reference panel: the Pearson's correlations of the African ancestral allele frequencies with allele frequencies in the HGDP African and European cohorts are 0.9830 and 0.7577, respectively; the correlations of the European ancestral allele frequencies with the HGDP African and European cohorts are 0.7293 and 0.9699. This pattern is replicated in the admixed American cohorts. When comparing the estimated ancestry-specific allele frequencies with the allele frequencies in their non-admixed counterparts from the 1000 Genomes Project, the Pearson's correlation

is 0.9954 for the African ancestry and 0.9729 for the European ancestry. Further, the correlation between the African ancestral allele frequencies estimated from the admixed African cohorts and admixed American cohorts is 0.9704 and the correlation between the European estimates is 0.9678.

4 Discussion

In this article, we have developed LoGicAl, a suite of novel accurate, flexible and scalable programs designed to estimate ancestry-specific allele frequencies from admixed samples when the upstream ancestry calling and genotyping are not perfect. These programs include likelihood-based estimation methods designed for unphased genotypes and these variations resolve the ancestry-and-genotype-informed hierarchical mixture model via accelerated EM algorithm. LoGicAl also includes a fast heuristic approach designed for pre-phased genotypes with reliable ancestry-genotype phases. By effectively adjusting for the uncertainty arising from ancestry calling and genotyping as well as statistical phasing errors, LoGicAl outperforms the state-of-arts estimation methods in both accuracy and scalability.

Several features facilitate the usage of LoGicAl in modern genetic and genomic analyses of admixed populations. Firstly, LoGicAl improves ancestry-specific allele frequency estimation accuracy by modeling uncertainty from upstream ancestry inference. This uncertainty measure is usually provided as posterior distribution of the ancestry calls in practice, e.g., RFMix[25], MOSAIC[40], and FLARE[26]. Considering ancestry information as probability distributions rather than merely point estimates is particularly important when inferring local ancestry is challenging, for example, because the contributing ancestral populations split recently and thereby are closely related[26]. Ignoring potential local ancestry misspecifications biases ancestry-specific allele frequency estimates by shrinking the estimates towards unstratified allele frequency. Additionally, when the contributing populations are more differentiated, the estimation bias accumulates through the misspecifications of either ancestry or genotype, resulting in exacerbated bias in estimation for cross-continental admixture.

Secondly, LoGicAl provides accurate ancestry-specific allele frequency estimates for genotypes sequenced from low-coverage whole genome sequencing. Recent research[24] has highlighted the capacity of using low-coverage sequencing to capture novel variants in underrepresented populations as a cost-effective alternative to high-coverage sequencing. Further, low-coverage sequencing has been successfully and broadly used in genome-wide association studies[41] and can improve the transferability of polygenic risk scores[42]. Our simulation results demonstrate that neglecting genotyping uncertainty from low-coverage sequencing induces bias in ancestry-specific allele frequency estimates. By modeling the genotyping uncertainty, LoGicAl substantially mitigates the estimation bias and provides more efficient estimates compared with existing methods. This improvement in estimation accuracy is consistent across allele frequency bins and persists in estimation for rare variants.

Thirdly, the implementation of squared iterative methods has drastically accelerated the classical EM algorithm without compromising on accuracy or stability. Moreover, incorporating phased genotype inputs via LoGicAl-phase further reduces

computational costs. These improvements in computational performance altogether enable the utilities of LoGicAl for analyzing biobank-scale genomic data of admixed samples.

Several potential limitations could be addressed to further enhance the statistical and computational performance of LoGicAl in future. Firstly, LoGicAl requires accurate uncertainty measures to efficiently account for the uncertainty from the upstream analyses. In particular, LoGicAl requires well-calibrated ancestry calling uncertainty measure to precisely adjust for the potential local ancestry misspecifications. Secondly, LoGicAl is a marginal model for estimating ancestry-specific allele frequencies and it does not explicitly model the linkage disequilibrium (LD) structure across variants. Computational benefits may be achieved via grouping input data according to their LD structure and local ancestry block and performing group-wide analyses[43]. Thirdly, we do not leverage relatedness or family information in the current LoGicAl framework and LoGicAl is primarily designed for unrelated admixed individuals. The violation of unrelatedness assumption and the inclusion of related individuals will not impact the estimation accuracy in theory; however, cautions are required if inference on ancestry-specific allele frequency is of research interest.

To summarize, we develop an accurate computational tool, LoGicAl, to estimate ancestry-specific allele frequency from admixed samples. LoGicAl provides comprehensive and precise descriptions of genetic variation patterns across whole genomes and informs ancestry-specific allele frequency spectrum for admixed populations. LoGicAl further contributes to constructing a precise spatial landscape and dynamics of genetic variations across geographical regions, and promote the understandings in population structure and demographic history at a finer scale.

Acknowledgments. We would like to thank Gonçalo Abecasis for a helpful discussion of phase-aware local ancestry inference. This work was supported by R01 HG011031 and R01 HG005855.

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] 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).

[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] Henderson, N. C. & Varadhan, R. Damped anderson acceleration with restarts and monotonicity control for accelerating em and em-like algorithms. *Journal of Computational and Graphical Statistics* **28**, 834–846 (2019).

[32] Zhou, H., Alexander, D. & Lange, K. A quasi-newton acceleration for high-dimensional optimization algorithms. *Statistics and computing* **21**, 261–273 (2011).

[33] 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).

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

[35] 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).

[36] 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).

[37] 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).

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

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

[40] Salter-Townshend, M. & Myers, S. Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics* **212**, 869–889 (2019).

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