HGG Advances

ARTICLE

AFA: Ancestry-specific allele frequency estimation in admixed populations: The Hispanic Community Health Study/Study of Latinos

Einat Granot-Hershkovitz,[1,2,*] Quan Sun,[3] Maria Argos,[4] Hufeng Zhou,[5] Xihong Lin,[5] Sharon R. Browning,[6] and Tamar Sofer[1,2,5,*]

Summary

Allele frequency estimates in admixed populations, such as Hispanics and Latinos, rely on the sample's specific admixture composition and thus may differ between two seemingly similar populations. However, ancestry-specific allele frequencies, i.e., pertaining to the ancestral populations of an admixed group, may be particularly useful for prioritizing genetic variants for genetic discovery and personalized genomic health. We developed a method, ancestry-specific allele frequency estimation in admixed populations (AFA), to estimate the frequencies of biallelic variants in admixed populations with an unlimited number of ancestries. AFA uses maximum-likelihood estimation by modeling the conditional probability of having an allele given proportions of genetic ancestries. It can be applied using either local ancestry interval proportions encompassing the variant (local-ancestry-specific allele frequency estimations in admixed populations [LAFAs]) or global proportions of genetic ancestries (global-ancestry-specific allele frequency estimations in admixed populations [GAFAs]), which are easier to compute and are more widely available. Simulations and comparisons to existing software demonstrated the high accuracy of the method. We implemented AFA on high-quality imputed data of ~9,000 Hispanics and Latinos from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), an understudied, admixed population with three predominant continental ancestries: Amerindian, European, and African. Comparison of the European and African estimated frequencies to the respective gnomAD frequencies demonstrated high correlations (Pearson $R^2 = 0.97–0.99$). We provide a genome-wide dataset of the estimated ancestry-specific allele frequencies for available variants with allele frequency between 5% and 95% in at least one of the three ancestral populations. Association analysis of Amerindian-enriched variants with cardiometabolic traits identified five loci associated with lipid traits in Hispanics and Latinos, demonstrating the utility of ancestry-specific allele frequencies in admixed populations.

Introduction

Admixed populations have multiple ancestral origins, with different admixture patterns within and between populations, resulting from historical worldwide migration of populations.[1] Estimation of ancestry-specific allele frequencies in admixed populations can identify ancestry-specific enriched variants, with higher minor allele frequencies (MAFs) in one ancestry compared with other ancestries. Fine mapping of association regions detected in admixture mapping, where one tests the association between local ancestry genomic interval (LAI) counts and a trait, can prioritize ancestry-specific enriched variants located in the identified regions for conditional association testing.[2,3] Similarly, genome-wide association studies (GWASs) of admixed populations can be followed by replication testing in unadmixed populations from a specific ancestry chosen based on the associated variant's ancestry-specific frequencies. More generally, allele frequencies are important for interpreting sequence variants, distinguishing between pathogenic and benign variants,[4] inferring demographic histories of populations, and determining susceptibility to disease.[5] Thus, ancestry-specific allele frequencies can contribute to both research and personalized medicine of admixed populations. This is especially relevant for modern-day populations that are becoming increasingly genetically admixed.[6]

Several population genetic software packages were previously developed for admixture and population structure analyses, producing a byproduct of ancestry-specific allele frequencies estimation in admixed populations.[7,8] Gravel et al.[9] developed an algorithm based on the expectation maximization (EM) framework relying on LAIs, but their method is not publicly available. A similar publicly available algorithm, Ancestry-Specific Allele Frequency Estimation (ASAFE), was developed. However, this method is available only for a three-way admixed diploid population and for genotyped markers located in LAIs, and it has not been implemented for large genome-wide analyses.[10] ASAFE was later extended to multi-way admixed populations in an algorithm that maximizes a multinomial likelihood.[11] Unfortunately, the software was not made public.

Here, we developed a computationally efficient method, ancestry-specific allele frequency estimation in admixed

[1]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA 02115, USA; [2]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA; [3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [4]School of Public Health, The University of Illinois, Chicago, Chicago, IL 60612, USA; [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [6]Department of Biostatistics, University of Washington, Seattle, WA 98105, USA
*Correspondence: egranot-hershkovitz@bwh.harvard.edu (E.G.-H.), tsofer@bwh.harvard.edu (T.S.)
https://doi.org/10.1016/j.xhgg.2022.100096.

populations (AFA), for the estimation of ancestry-specific allele frequencies for biallelic variants with no need for phased data. AFA can be applied to admixed populations with an unlimited predetermined number of ancestries. Our model is similar to that proposed by Gravel et al., using maximum-likelihood estimation by modeling the conditional probability of having a variant allele given local proportion ancestries (local-ancestry-specific allele frequency estimations in admixed populations [LAFA]). We further extended the model by leveraging global ancestry proportions (global-ancestry-specific allele frequency estimations in admixed populations [GAFA]), which are easier to compute and are more widely available, and we provide publicly available code. We further improved upon previous methods by adding accuracy estimates and developing common workflow language (CWL) workflows. We studied the performance of GAFA and LAFA in simulations and compared them with existing software. We then implemented the method on high-quality imputed genome-wide genetic data from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), an admixed population previously characterized with three predominant continental ancestries: Amerindian, European, and African, with varying proportions between individuals.[12] We provide estimated Hispanic or Latino ancestry-specific allele frequencies estimated based on the HCHS/SOL for all variants with allele frequency between 5% and 95% in at least one of the three ancestral populations. Finally, we performed association analyses of Amerindian-enriched variants with cardiometabolic traits in the HCHS/SOL population.

We computed ancestry-specific frequencies through AFA using the previous global proportion ancestries calculated by ADMIXTURE[12] and LAIs calculated by RFMix.[13,14] We hypothesized that frequency estimates of variants using local ancestries (LAFA) would be more precise than estimates using global proportion ancestries (GAFA). We compared our estimated ancestral-specific variant frequencies for European and African ancestries with their respective frequencies published in gnomAD, expecting them to be similar though not identical. We provide estimated Hispanic or Latino ancestry-specific allele frequencies estimated based on the HCHS/SOL for all variants with allele frequency between 5% and 95% in at least one of the three ancestral populations. Finally, we performed association analyses of Amerindian-enriched variants with cardiometabolic traits in the HCHS/SOL population.

Subjects and methods

Study population

The HCHS/SOL is a population-based longitudinal cohort study of US Hispanics and Latinos with participants recruited from four field centers (Bronx, NY; Chicago, IL; Miami, FL; and San Diego, CA) by a sampling design previously described.[15,16] A total of 16,415 self-identified Hispanic or Latino adults 18 to 74 years old were recruited during the first visit between 2008 and 2011, and various biospecimen and health information about risk and protective factors was collected. Most of the individuals self-identify with six Hispanic or Latino backgrounds: Cuban, Central American, Dominican, Mexican, Puerto Rican, and South American backgrounds (Table S1).[12] All participants in this analysis signed informed consent in their preferred language (Spanish or English) to use their genetic data. The study was reviewed and approved by the Institutional Review Boards at all collaborating institutions.

Genetic data

Genotyping and quality control were previously described.[12,17] In brief, genotyping was performed using Illumina MEGA array, and a total of 11,928 samples and 985,405 genotyped variants passed quality control. Genome-wide imputation was conducted using the multi-ethnic NHLBI Trans-Omics for Precision Medicine (TOPMed) freeze 8 reference panel (GRCh38 assembly).[18] This panel includes ~8,000 individuals from the HCHS/SOL, as well as additional Hispanic or Latino individuals. The improvement of imputation quality by a previous freeze of the TOPMed panel, which included multi-ethnic populations, was previously demonstrated.[18] Due to the overlap of samples in our target data and the TOPMed freeze 8 reference panel (n = 6,201), we recalculated the estimated imputation quality (R^2) using only non-overlapped samples to avoid over-estimates of the imputation quality. After filtering variants with $R^2 < 0.6$ and minor allele count ≤ 5, a total of 42,038,818 imputed variants remained for analysis. Coordinates of genotyped and imputed variants were converted from GRCh38 to GRCh37 using the liftOver tool from University of California, Santa Cruz (UCSC)[19] for LAFA analysis since the LAIs were based on GRCh37 (as described below).

Global proportion ancestries

Global continental ancestry proportions were previously estimated for 9,864 unrelated HCHS/SOL individuals using ADMIXTURE software under the assumption of three ancestral populations (Amerindian, African, and European), based on reference panels representing these ancestral populations.[12] For this analysis, we used genetic data from 8,933 individuals who consented to genetic data sharing with the broad scientific community. Overall, the average ancestral global proportion of the three ancestries in the total dataset are 55% European, 30.5% Amerindian, and 14.5% African. The distribution of the average global proportion of the three ancestries by background groups is presented in Table S1.

LAIs

Three-way LAIs (Amerindian, African, and European) were previously inferred in 12,793 HCHS/SOL individuals using the RFMix software with a reference panel derived from the combination of the Human Genome Diversity Project (HGDP) and the 1000 Genome Project (using the GRCh37 assembly) representing the relevant ancestral populations.[20] Overall, 15,500 are LAIs dispersed throughout the genome (14,815 LAI in autosomal chromosomes), each spanning ten to hundreds of thousands of base pairs. After excluding individuals to generate a dataset in which none of the individuals are third-degree relatives or closer and individuals who withdrew consent for genetic studies, 9,512 individuals remained.

All participants in this analysis signed informed consent in their preferred language (Spanish or English) to use their genetic data. The study was reviewed and approved by the Institutional Review Boards at all collaborating institutions.

Statistical analysis

The statistical model for estimation of ancestry-specific allele frequencies in admixed populations (AFA)

Suppose that we have a population of n individuals with predetermined K genetic ancestries. Consider a specific biallelic genetic variant in an autosomal chromosome. Each person has two copies of a variant potentially inherited from different ancestries. The genetic ancestry of each copy of the variant was inherited from the local ancestry encompassing the variant. For any given variant allele g, denote its ancestry-specific frequencies by $f_1, \ldots, f_K$ in ancestries $1, \ldots, K$, respectively. Denote further the probability that person i has local ancestry at the variant by $p_{i,k}$, $k = 1, \ldots, K$. We have that $p_{i,1} \ldots, p_{i,K}$ satisfy $0 \le p_{i,k} \le 1$ and $p_{i,1} + \ldots + p_{i,K} = 1$ for $i = 1, \ldots, n$, $k = 1, \ldots, K$. The allele count at the variant on a given chromosomal copy is sampled from a mixture of Bernoulli distributions, with

$$\begin{aligned}\Pr(g_i = 1) = \Pr(g_i = 1 \mid ancesry\ 1) \times p_{i,1} + \ldots \\ + \Pr(g_i = 1 \mid ancestry\ K) \times \mathrm{p_{i,K}} = f_1 p_{i,1} + \cdots \\ + f_K p_{i,K} = p_{i,mix}.\end{aligned}$$

For unphased data, or when using genetic ancestry probabilities that are not specific to the variant (e.g., global ancestries), the probabilities $p_{i,1}, \cdots, p_{i,K}$ are the same for the two copies of the allele. Under Hardy-Weinberg equilibrium (HWE) at each ancestry, we can extend the model above to a binomial distribution with two alleles. If g_i is now a biallelic variant, then

$$\Pr(\mathrm{g}_i = \mathrm{l}) = \binom{2}{l} p_{i,mix}^{l} \left(1 - p_{i,mix}\right)^{2-l},\ \mathrm{l} \in \{0, 1, 2\}. \quad \text{(Equation 1)}$$

Assuming ancestral probabilities $p_{i,1}, \ldots, p_{i,K}$ are known, the unknown frequencies $f_1, \ldots, f_K$ can now be estimated by maximizing the log likelihood across the sample of independent individuals. The standard errors of the estimated frequencies can be used to compute confidence intervals. We use the base R optim function with the "L-BFGS-B" optimization method for $K > 1$ ancestries and the "Brent" method when estimating allele frequency in one ancestry (for example, if $K - 1$ for $K > 1$ frequencies are known or assumed).

Choosing probabilities of genetic ancestry at the variant

To maximize the likelihood above, we assume that the ancestral probabilities $p_{i,1}, \ldots, p_{i,K}$ of the study individuals are known. In practice, they are estimated. We consider two estimators. First is the global proportion of ancestry (GAFA). These could be computed using software packages such as ADMIXTURE or RFMix with a subset of independent, genotyped genetic variants, with or without a reference panel.[6–8,14] The second estimator is based on LAIs (LAFA). Local ancestry analysis results in a segmentation of the genome in which each segment, LAI, is assigned a genetic ancestry. Thus, a given variant g is overlapping with a certain LAI, say LAI_g, which is annotated with two genetic ancestries. When these LAIs are unphased with respect to the allele counts, we first generate a vector of counts of local ancestries $(c_{i,1}, \ldots, c_{i,K})$ and divide all entries by two, the highest attained count. In mathematical notation,

$$\begin{aligned}p_{i,k} = c_{i,k}/2 &= (\#\ \text{genetic ancestries of type}\ k\ \text{in}\ LAI_g)/2, \quad k \\ &= 1, \ldots, K.\end{aligned}$$

The probabilities here take values 0, 0.5, 1. If using phased local ancestry data, in that local ancestry of each of the two estimated chromosomal alleles, the algorithm can use the Bernoulli distribution with the estimated ancestral distribution as the $p_{i,k}$, e.g., the posterior probabilities of ancestry that are provided by RFMix.

A potential limitation of the AFA algorithm is that it uses binomial distribution with the person-specific parameter being a mixture of the ancestry-specific frequencies. This implies that an HWE is assumed on the wrong, mixture-level parameter. With LAFA, we can instead use the Poisson binomial distribution: a sum of independent Bernoulli variables with potentially different parameters. More information about LAFA-Poisson binomial is provided in Note S1. Our simulations (below) address these issues by comparing implementation options and assessing estimation and inference accuracy.

Computing ancestry-specific allele frequencies on the X chromosome

The methodology for the X chromosome is similar, with a slight difference for males, where we use a Bernoulli distribution (or a Binomial distribution with parameters $(p_{mix},\ 1)$) to account for the fact that there is a single observed allele.

Handling of boundary conditions

The log likelihood of the binomial distribution cannot be maximized at the boundaries, i.e., when the data are consistent with an ancestry-specific frequency at the boundary of the parameter space, e.g., $f_k \in \{0, 1\}$ for some $k = 1, \ldots, K$. To prevent non-convergence of the estimation algorithm, we implemented a procedure that generates synthetic observations and adds them to the data. These are $2K$ synthetic observations, two for each ancestry, mimicking a reference and alternate allele from each of the genetic ancestries. For example, one synthetic observation will have a single reference allele for a (simulated) person, and the ancestral probabilities for this person are $p_{i,k} = 1$ for ancestry k and $p_{i,l} = 0$ for all other ancestries $l \neq k, l \in \{1, \ldots, K\}$. Another synthetic observation will have a single alternate allele for this variant and the same values of ancestral probabilities. In addition, the algorithm allows for setting box constraints on the boundaries.[21]

An approximation for computing ancestry-specific allele frequency using imputed data

When imputed data are confidently estimated, the extension of the algorithm to imputed genotypes is straightforward. For imputed genotypes with fractions, we cannot compute the log likelihood based on the probability in Equation 1. Instead, we notice that we can decompose the function into two parts: "2 choose l" and $p_{i,mix}^{l}(1 - p_{i,mix})^{2-l}$. The second part can be computed for any l, while the first part cannot. Instead, we apply linear interpolation to compute a value approximating 2 choose l based on the values of this function evaluated at the nearest integers higher and lower than l.

Simulation studies

We studied our method, AFA, in simulations to determine how the ancestry-specific allele frequency estimation accuracy is influenced by the effective sample size, effn, defined as $\mathrm{effn}_k = \sum_{i=1}^{n} p_{i,k}$ for ancestry $k = 1, \ldots, K$, by the expected allele frequencies (rare versus common variants) and by using the local versus global proportion ancestries (LAFA versus GAFA).

We simulated a three-way admixed population, using fixed $\mathrm{effn}_1 = \mathrm{effn}_2 = 1{,}000$ and varied effn_3 in the range 100–4,000, and focused on the estimation of f_3. We fixed $f_1 = 0.5$, $f_2 = 0.3$ throughout and varied the allele frequency $f_3 \in \{0.01, 0.05, 0.1, 0.2\}$. First, we simulated local ancestries based on global effn (where $\mathrm{n} = \mathrm{effn}_1 + \mathrm{effn}_2 + \mathrm{effn}_3$). We assumed that each person has two copies of 10 LAIs of equal lengths. Thus, the overall

number of LAIs of ancestry $k \in \{1, 2, 3\}$ was $n \times \text{effn}_k * 20$. Then, we randomly assigned 20 LAIs (two copies of 10 LAIs) to individuals and computed the global proportion of ancestries for each individual as the proportion of LAIs of each ancestry. The genetic variant was assumed to be in the first LAI. Next, we simulated the allele counts based on the allele frequencies f_1, f_2, and f_3. For each person and each copy of the first LAI, we sampled the allele from the Bernoulli distribution with a probability according to the ancestry at the interval copy. To mimic the real data, which are unphased, we then summed the allele count across the two copies for each person. Finally, we estimated ancestry-specific allele frequencies using the computed global proportion ancestries and using the ancestries of the first LAI. We performed $n_{sim} = 1{,}000$ simulation replicates for each setting. We also performed a similar simulation based on a homogeneous population derived from a single ancestry to compare the expected bias in frequency estimation in admixed populations with that in a non-admixed population when using the same algorithm.

Let $\overline{\widehat{f_3}} = \frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} \widehat{f_{3,j}}$ denote the mean estimated f_3 across simulations. We assessed the frequency estimation accuracy of f_3 using the following measures:

1. Bias: $(f_3 - \overline{\widehat{f_3}})$.
2. Inflation: $\overline{\widehat{f_3}} / f_3$.
3. RMSE (root-mean-squared error): $\sqrt{\frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} (f_3 - \widehat{f_{3,j}})^2}$.

Similarly, we also simulated two-way, four-way, and five-way admixed populations and assessed the frequency estimation accuracy for each scenario. We also simulated a three-way admixed population and compared the frequency estimation accuracy for GAFA, LAFA, and LAFA phased data (when the ancestry of each allele copy is known). We also simulated a two-way admixed population with various scenarios of ancestry-specific frequencies and ancestry-heterozygosity proportions and compared the frequency estimation accuracy for LAFA, LAFA phased, and LAFA Poisson binomial (more information on Note S2).

We further compared LAFA with ASAFE to compare both accuracy and computation time. Both methods operate on local ancestries and can be applied to a simulated three-way admixed simulated dataset. We fixed effn1 = effn2 = effn3 = 1,000 and $f_1 = 0.5, f_2 = 0.3$ and varied the allele frequency $f_3 \in \{0.1, 0.2, 0.4\}$. We simulated local ancestries similar to the description above. We estimated ancestry-specific allele frequencies using LAFA (Poisson binomial and binomial) and ASAFE, with a $n_{sim} = 1{,}000$ simulation replicates for each setting. We computed the average computing time, average bias, and RMSE per variant for LAFA and ASAFE.

Comparing ancestry-specific allele frequency estimates with previously published estimates

We compared the estimated ancestry-specific frequencies of nine variants using GAFA and LAFA with previously published estimated ancestry-specific allele frequencies based on the ASAFE method in the HCHS/SOL dataset.[3,22,23] We also compared the estimated Amerindian frequency of four variants with the previously published frequencies in Pima-Indians.[3]

Using ADMIXTURE to estimate ancestry-specific allele frequencies

We used ADMIXTURE to estimate ancestry-specific allele frequencies in HCHS/SOL and to compare them with AFA estimates. We first performed linkage disequilibrium (LD) pruning in the HCHS/SOL unrelated dataset, using the –indep-pairwise command in Plink (window size 50 kb, a shift of 10 variants at each step, and LD between variants $r^2 > 0.1$), as suggested by the ADMIXTURE tutorial. We then ran the ADMIXTURE algorithm on the output file that included n = 1,274,187 genome-wide variants. The resulting P file included ancestry-specific estimates for these variants. We compared these estimates with the corresponding GAFA estimates.

Comparing estimated ancestry-specific allele frequencies with gnomAD allele frequencies

We compared the estimated European and African frequencies in the admixed HCHS/SOL population using GAFA and LAFA with the gnomAD v.2 liftover (GRCh38) non-Finnish European and African frequencies, respectively, by plotting and calculating the Pearson squared correlation coefficient. We assessed only gnomAD variants passing quality control filters (FILTER = = "PASS"), with an ancestral minor allele count of ≥ 100 respective to the assessed ancestry. We performed the same comparison between gnomAD frequencies and ADMIXTURE-estimated frequencies. Finally, we calculated the percentage of estimated confidence intervals (CIs) using GAFA and LAFA, which include the corresponding gnomAD MAFs, binned by MAF categories.

Identification of Amerindian-enriched variants in the HCHS/SOL population

In populations that have undergone bottlenecks and genetic drifts, such as the Native Americans and American Indians, it is expected that some risk variants of large effects have risen in frequency compared with the population of origin.[9,24] We thus sought variants with a substantially higher Amerindian-specific frequency in the HCHS/SOL compared with the European- and African-specific allele frequencies. We defined Amerindian-enriched variants as those with both European and African MAF < 0.01 and Amerindian frequency between 0.05 and 0.95.

Associations of Amerindian-enriched variants with cardiometabolic traits

We performed association tests for the Amerindian enriched variants using the "GENESIS" R package, with 12 cardiometabolic-related traits in ~11,700 HCHS/SOL individuals who had both genetic data and cardiometabolic traits. We adjusted for age, sex, center, log of the sample weights, first five principal components (PCs), and genetic analysis groups. Genetic analysis groups were constructed based on a combination of self-identified Hispanic or Latino background and genetic similarity and are classified as Central American, Cuban, Dominican, Mexican, Puerto Rican, and South American.[12] We further adjusted for both linear and squared age for systolic and diastolic blood pressure and hypertension and adjusted for BMI for all outcomes except BMI and obesity. Table S2 lists the 12 cardiometabolic outputs we analyzed and their corresponding covariates and medication adjustments. We used the Bonferroni correction to determine the p value threshold. To identify independently associated SNPs, we performed conditional analysis using the index (most significant) SNP as a covariate.

Results

Simulation studies

Table 1 and Figure 1 summarize the results from simulation studies of frequency estimation in a three-way admixed population, based on GAFA or LAFA. For comparison, simulation results based on an unadmixed population under the same framework, essentially reducing to standard maximum likelihood estimation, are presented in Table S3

Table 1. Results from simulation studies of frequency estimation of a biallelic variant in a three-way admixed population, based on AFAs, by different effective sample sizes and different expected minor allele frequencies

		Three-way admixed population using GAFA							Three-way admixed population using LAFA						
Expected MAF	Effective N	MAF mean	MAF difference	MAF ratio	RMSE	CI % coverage	Interval 2.5%	Interval 97.5%	MAF mean	MAF difference	MAF ratio	RMSE	CI % coverage	Interval 2.5%	Interval 97.5%
0.005	100	0.153	0.148	30.556	0.164	0.90	0.0569	0.3284	0.040	0.035	8.001	0.039	0.97	0.0172	0.0797
	200	0.109	0.104	21.736	0.115	0.90	0.0407	0.2264	0.023	0.018	4.643	0.021	0.98	0.0102	0.0499
	500	0.063	0.058	12.683	0.065	0.90	0.0251	0.1327	0.009	0.004	1.856	0.006	0.99	0.0038	0.0194
	1,000	0.040	0.035	7.981	0.040	0.91	0.0145	0.0866	0.007	0.002	1.380	0.003	0.99	0.0029	0.0123
	2,000	0.023	0.018	4.542	0.020	0.93	0.0088	0.0471	0.006	0.001	1.105	0.002	0.94	0.0027	0.0091
	4,000	0.012	0.007	2.468	0.009	0.93	0.0047	0.0253	0.005	0.000	1.060	0.001	0.94	0.0035	0.0072
0.01	100	0.156	0.146	15.631	0.164	0.90	0.0609	0.3369	0.043	0.033	4.306	0.038	0.97	0.0174	0.0865
	200	0.110	0.100	11.013	0.112	0.91	0.0424	0.2340	0.027	0.017	2.651	0.020	0.98	0.0103	0.0563
	500	0.066	0.056	6.645	0.064	0.92	0.0247	0.1395	0.014	0.004	1.401	0.007	0.99	0.0043	0.0272
	1,000	0.044	0.034	4.383	0.039	0.91	0.0171	0.0928	0.012	0.002	1.156	0.004	0.96	0.0050	0.0195
	2,000	0.026	0.016	2.622	0.020	0.93	0.0093	0.0562	0.011	0.001	1.053	0.002	0.95	0.0067	0.0151
	4,000	0.016	0.006	1.574	0.009	0.95	0.0061	0.0303	0.010	0.000	1.022	0.001	0.94	0.0074	0.0129
0.05	100	0.176	0.126	3.524	0.149	0.94	0.0685	0.3589	0.071	0.021	1.410	0.034	0.97	0.0271	0.1327
	200	0.134	0.084	2.677	0.102	0.94	0.0500	0.2681	0.060	0.010	1.208	0.024	0.95	0.0235	0.1071
	500	0.091	0.041	1.810	0.056	0.94	0.0344	0.1757	0.053	0.003	1.063	0.013	0.93	0.0290	0.0797
	1,000	0.069	0.019	1.388	0.033	0.95	0.0259	0.1258	0.051	0.001	1.020	0.008	0.96	0.0373	0.0660
	2,000	0.056	0.006	1.117	0.018	0.97	0.0265	0.0887	0.050	0.000	1.002	0.004	0.96	0.0413	0.0588
	4,000	0.052	0.002	1.035	0.010	0.95	0.0336	0.0716	0.050	0.000	1.004	0.003	0.95	0.0446	0.0561
0.1	100	0.200	0.100	2.003	0.134	0.95	0.0744	0.4059	0.115	0.015	1.152	0.038	0.96	0.0512	0.1863
	200	0.164	0.064	1.641	0.093	0.94	0.0611	0.3204	0.107	0.007	1.068	0.026	0.96	0.0627	0.1587
	500	0.126	0.026	1.259	0.052	0.95	0.0504	0.2293	0.102	0.002	1.023	0.015	0.96	0.0742	0.1320
	1,000	0.110	0.010	1.104	0.033	0.95	0.0512	0.1744	0.101	0.001	1.009	0.010	0.95	0.0823	0.1214
	2,000	0.103	0.003	1.031	0.021	0.95	0.0640	0.1433	0.100	0.000	1.001	0.006	0.95	0.0882	0.1131
	4,000	0.101	0.001	1.009	0.011	0.97	0.0796	0.1217	0.100	0.000	1.003	0.004	0.95	0.0929	0.1084
0.2	100	0.264	0.064	1.322	0.123	0.96	0.0984	0.4848	0.209	0.009	1.044	0.043	0.96	0.1275	0.2906
	200	0.234	0.034	1.171	0.086	0.95	0.0955	0.4126	0.203	0.003	1.015	0.029	0.96	0.1466	0.2588
	500	0.214	0.014	1.069	0.059	0.95	0.1062	0.3281	0.201	0.001	1.006	0.018	0.94	0.1655	0.2379
	1,000	0.205	0.005	1.023	0.036	0.96	0.1351	0.2749	0.200	0.000	1.001	0.012	0.96	0.1764	0.2217
	2,000	0.201	0.001	1.005	0.023	0.95	0.1571	0.2468	0.200	0.000	1.002	0.008	0.95	0.1844	0.2152
	4,000	0.200	0.000	1.000	0.013	0.96	0.1755	0.2240	0.200	0.000	1.000	0.005	0.95	0.1895	0.2103

For each of the settings, we tested 1,000 simulation replicates and calculated the mean frequency estimate; the difference; ratio of the mean observed frequency and the expected frequency; the RMSE of the estimate frequencies; the percentage of CI, including the expected frequency (coverage); and the 95% interval of the estimated frequencies. The results refer to one of the ancestries. The characteristics of the other two ancestries were the same in all simulations, with effective sample size of effn = 1,000, one ancestry with MAF = 0.5, and the other with MAF = 0.3.
Abbreviations: AFA, ancestry-specific allele frequency estimation in admixed populations; CI, confidence interval; GAFA, global-ancestry-specific allele frequency estimation in admixed populations; LAFA, local-ancestry-specific allele frequency estimation in admixed populations; MAF, minor allele frequency; RMSE, root-mean-squared error.

and Figure S1. As expected, estimated frequencies become more accurate with increasing effective sample size and increasing MAF. Likely due to the boundaries of the parameter space, the estimated frequencies tend to be biased toward more common MAFs until large-enough effective sample sizes or allele frequencies (or, in other words, enough counts of the minor allele) are available. In addition, accuracy increased when using LAFA compared with GAFA. For example, for MAF = 0.01 and effn = 4,000, we had bias = 0.00574 for GAFA and bias = 0.00022 for LAFA; for MAF = 0.2 and effn = 1,000, we had bias = 0.00453 for GAFA and bias = 0.00028 for LAFA (Table 1). Similar trends of improved accuracy of frequency estimation with larger effective sample sizes and higher MAFs

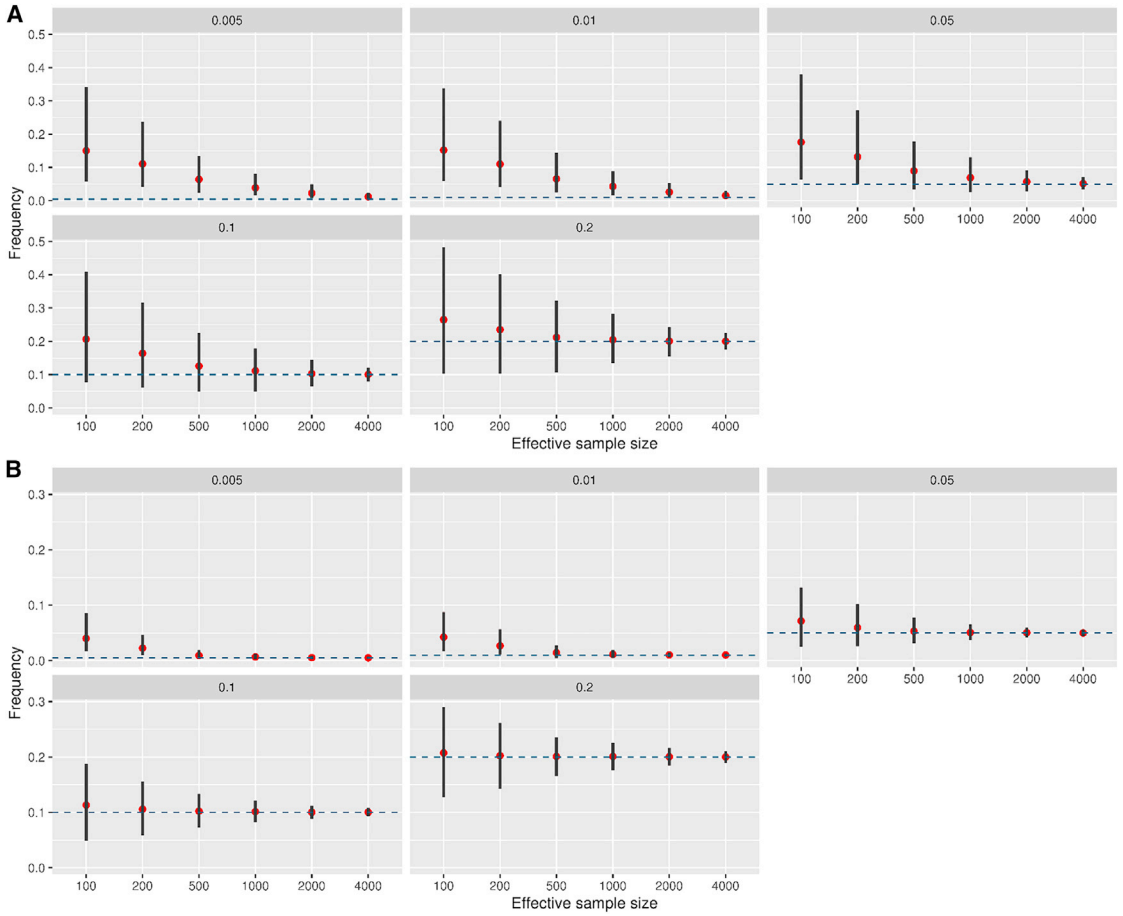

Figure 1. AFA simulation studies
Results from simulation studies of frequency estimation of a biallelic variant in a three-way admixed population, based on (A) global-ancestry-specific allele frequency estimation in admixed populations (GAFA) and (B) local-ancestry-specific allele frequency estimation in admixed populations (LAFA). Various settings include a different effective sample size of effn (x axis) and different expected minor allele frequencies (indicated in the upper title of each graph). We performed 1,000 simulation replicates of each scenario. Each dot represents the mean frequency of 1,000 simulation replicates, and each line represents the 95% interval estimated frequencies across the simulation replicates.

are observed in the unadmixed population analysis (Table S3; Figure S1) with, unsurprisingly, higher accuracy compared with the admixed population. Figure S2 summarizes the results from simulation studies of GAFA or LAFA ancestry-specific frequency estimation in two- to five-way admixed populations, showing similar accuracy estimations for each case (while holding the effective sample size from the ancestral population of interest fixed), emphasizing the compatibility of our method to admixed populations with multiple ancestries. Figure S3 summarizes the results from simulation studies in a three-way admixed population of GAFA, LAFA, and LAFA using phased data (LAFA-phased), showing slightly higher accuracy of LAFA-phased compared with LAFA, as expected. Figure S4 summarizes the results from LAFA simulation studies in a two-way admixed population with different ancestry-heterozygosity proportions. LAFA-phased method is the most accurate; its performance did not depend on ancestry proportion heterozygosity. However, a high proportion of ancestry heterozygosity affects the accuracy of both LAFA and LAFA Poisson binomial, with LAFA being the most sensitive to high heterozygosity proportion.

Table S4 summarizes the simulation results comparison between LAFA and ASAFE. Using three ancestries, the average ancestry-specific allele frequency estimates are similar for both methods, and so are the average bias and RMSE, both low and similar. The average computing time per single variant was higher for LAFA compared with ASAFE, and it was substantially higher when LAFA used the Poisson binomial distribution. However, we developed CWL workflows for both GAFA and LAFA, which are portable to multiple computational environments and demonstrate the time efficiency of our method when applied on a genome-wide scale (see more details in the following section).

Hispanic Community Health Study/Study of Latinos

We applied AFA to the HCHS/SOL imputed dataset, excluding variants with minor allele count less than or equal to five, setting frequency boundary conditions

Table 2. Number of estimated variant frequencies per chromosome in HCHS/SOL that are common, with frequency between 5% and 95%, in at least one of the three ancestral populations, stratified by boundary condition, calculated via GAFA or LAFA

	GAFA			LAFA		
Chromosome	Total	Boundary 1 × 10^{-5} (%)	Boundary 1 × 10^{-2}	Total	Boundary 1 × 10^{-5} (%)	Boundary 1 × 10^{-2}
1	731,372	583,809 (79.82)	147,563	733,709	413,042 (56.3)	320,667
2	789,792	629,106 (79.65)	160,686	812,360	448,927 (55.26)	363,433
3	683,573	547,793 (80.14)	135,780	692,240	389,055 (56.2)	303,185
4	699,617	560,836 (80.16)	138,781	703,178	407,366 (57.93)	295,812
5	608,090	488,543 (80.34)	119,547	622,061	343,828 (55.27)	278,233
6	623,354	516,241 (82.82)	107,113	625,767	355,100 (56.75)	270,667
7	557,900	455,514 (81.65)	102,386	557,945	321,508 (57.62)	236,437
8	530,427	419,391 (79.07)	111,036	545,404	294,384 (53.98)	251,020
9	411,577	332,874 (80.88)	78,703	417,816	232,124 (55.56)	185,692
10	483,953	394,154 (81.44)	89,799	481,662	276,438 (57.39)	205,224
11	470,444	378,998 (80.56)	91,446	479,053	269,368 (56.23)	209,685
12	458,340	367,421 (80.16)	90,919	459,471	257,015 (55.94)	202,456
13	348,128	288,100 (82.76)	60,028	353,459	208,717 (59.05)	144,742
14	307,247	250,059 (81.39)	57,188	307,967	171,930 (55.83)	136,037
15	271,244	219,250 (80.83)	51,994	275,865	155,716 (56.45)	120,149
16	285,290	226,155 (79.27)	59,135	279,954	158,444 (56.6)	121,510
17	256,758	207,176 (80.69)	49,582	250,340	140,177 (55.99)	110,163
18	270,605	221,435 (81.83)	49,170	273,281	158,435 (57.98)	114,846
19	215,685	176,195 (81.69)	39,490	202,739	115,992 (57.21)	86,747
20	212,499	169,496 (79.76)	43,003	212,818	118,307 (55.59)	94,511
21	129,185	105,026 (81.3)	24,159	131,628	75,669 (57.49)	55,959
22	127,717	104,683 (81.96)	23,034	123,688	69,062 (55.84)	54,626
X	335,292	235,553 (70.25)	99,739	301,688	139,741 (46.32)	161,947
Total	9,808,089	7,877,808 (80.32)	1,930,281	9,844,093	5,520,345 (56.08)	4,323,748

(low = 0.00001; high = 0.99999) as arguments to the optimization function. If AFA did not converge for a given variant, we applied it again with a stricter boundary condition (low = 0.01; high = 0.99). We developed workflows for GAFA and LAFA on BioData Catalyst Powered by Seven Bridges.[25] We processed data in a parallel manner by batching the workflows by chromosomes and scattering jobs by blocks of 3,000 variants, using the c5.18xlarge spot instance provisioned on Amazon Web Services. The workflows are described (represented) in the CWL open standard[26] and are therefore portable to multiple computational environments. The computation time for the shortest chromosome (chromosome 21; n = 552,556 variants) was 57 min using GAFA and 110 min using LAFA, with ~50 jobs running in parallel. The number of estimated variant frequencies per chromosome is summarized in Table S5 stratified by boundary condition for both GAFA and LAFA. The number of variants for which we provide estimated variant frequencies, under the condition that they have a frequency between 5% and 95% in at least one of the three ancestral populations, is summarized in Table 2 stratified by boundary condition for both GAFA and LAFA. In general, rare variants required strict boundary conditions (0.01 rather than 0.00001) on the estimated frequencies for algorithm convergence.

Comparing ancestry-specific allele frequency estimates with previously published estimates

Table 3 summarizes nine previously published HCHS/SOL ancestry-specific allele frequencies estimated by ASAFE for comparison with our GAFA and LAFA frequency estimations. Frequency estimations for all nine variants are highly comparable, with absolute mean frequency differences for African = 0.0008, European = 0.0153, and Amerindian = 0.0101 for GAFA and African = 0.0023, European = 0.019, and Amerindian = 0.0094 for LAFA. Table 4 summarizes four previously published allele frequencies of Pima Indians to the Amerindian-specific allele frequency estimated in HCHS/SOL based on GAFA and LAFA. Here, too, the

Table 3. HCHS/SOL ancestry-specific allele frequencies previously published, estimated by ASAFE, compared with GAFA and LAFA frequency estimations

					Method								
					ASAFE			GAFA			LAFA		
SNP	CHR	POS (hg38)	Ref.	Alt.	African	European	Amerindian	African	European	Amerindian	African	European	Amerindian
rs4133185[a]	7	15,461,794	A	T	0.126	0.180	0.823	0.123	0.172	0.817	0.124	0.175	0.829
rs4628172[a]	7	15,455,525	T	G	0.101	0.175	0.820	0.097	0.168	0.814	0.100	0.171	0.827
rs4721442[a]	7	15,466,382	T	G	0.877	0.821	0.177	0.884	0.828	0.179	0.879	0.826	0.164
rs1458038[b]	4	80,243,569	T	C	0.030	0.250	0.310	0.065	0.257	0.289	0.035	0.248	0.322
rs9366626[b]	6	25,684,725	G	A	0.750	0.620	0.250	0.748	0.627	0.324	0.744	0.615	0.268
rs73156692[b]	12	101,214,917	A	G	0.130	0.230	0.010	0.165	0.242	0.012	0.140	0.244	0.011
rs113719683[c]	4	40,431,429	T	C	1.000	0.926	0.997	0.974	0.872	0.952	0.990	0.866	0.959
rs112178366[c]	4	40,431,425	A	G	1.000	0.927	0.997	0.974	0.873	0.952	0.990	0.867	0.959
rs112927755[c]	4	40,431,443	G	A	1.000	0.927	0.997	0.976	0.879	0.952	0.990	0.873	0.959

Frequencies refer to the ref. allele. ASAFE, ancestry-specific allele frequency estimation.
[a]Burkart,[22] 2017
[b]Sofer et al.,[3,17] 2017
[c]Jian et al.,[23] 2020

absolute mean frequency differences are low with GAFA = 0.03 and LAFA = 0.01.

Comparing estimated ancestry-specific allele frequencies with gnomAD

Figure 2 compares the estimated European- and African-specific allele frequencies in HCHS/SOL for variants on chromosome 2 using GAFA and LAFA with the gnomAD non-Finnish European and African frequencies, respectively, and stratifies to genotyped variants and imputed variants with different imputation quality thresholds. Correlations with gnomAD frequencies were similar for all sets of variants. All other chromosomes' comparisons with gnomAD are presented in Figures S6 (GAFA) and S7 (LAFA). All estimated frequencies were highly correlated, with Pearson R^2 = 0.97–0.99. We further calculated the percentage of gnomAD allele frequencies that are included in the corresponding CI estimated in HCHS/SOL by GAFA or LAFA, binned by gnomAD frequency categories (Table 5). The mean range of CIs was also calculated for each category and was consistently smaller for LAFA compared with GAFA since the ancestral determination for each variant is more accurate when using LAIs. Thus, LAFA resulted in a lower percentage of included gnomAD allele frequencies relative to GAFA; however, this does not indicate a superiority of GAFA over LAFA because of potentially true differences in ancestry-specific allele frequencies in HCHS/SOL compared with gnomAD. The mean ranges of CIs are lower in low-frequency variant bins compared with the common frequency bins, both for GAFA and LAFA.

We also estimated ancestry-specific allele frequencies on a pruned dataset subset of 1,274,187 variants from HCHS/SOL using ADMIXTURE. The total running time was ~45 h on a Linux cluster. Figure S8A compares the estimated ancestry-specific allele frequencies using GAFA and ADMIXTURE for all three ancestries. Correlations were high (R^2 > 0.99), as expected. Figures S8B and S8C compare the estimated ancestry-specific allele frequencies for the pruned subset using ADMIXTURE or GAFA with the corresponding gnomAD frequencies, presenting similar correlations.

Correlation of estimated ancestry-specific allele frequencies between the GAFA and LAFA for each of the three ancestries

Figure 3 presents strong correlations of the chromosome 2 estimated ancestry-specific allele frequencies in the HCHS/SOL population between GAFA and LAFA for each of the three ancestral populations. The European's correlation is stronger than the Africans and Amerindians. This is probably due to their larger effective sample size in the HCHS/SOL, enabling a more precise estimation of the alleles' frequencies (effn based on global proportion ancestries: African = 1,296, European = 4,912, and Amerindian = 2,725). All other chromosomes' correlations are presented in Figure S9.

Correlation of the estimated ancestry-specific allele frequencies between different ancestries

Figure 4 presents weak correlations of the estimated ancestry-specific allele frequencies for chromosome 2 variants in the HCHS/SOL population between the three ancestral populations, for both GAFA and LAFA. The squared Pearson correlation coefficient is strongest when comparing Amerindian-specific to European-specific frequencies (GAFA: R^2 = 0.78; LAFA: R^2 = 0.76), followed by the comparison of African to European (GAFA: R^2 = 0.71; LAFA: R^2 = 0.71), and weakest in the comparison of African to Amerindian (GAFA: R^2 = 0.61; LAFA: R^2 = 0.6). Similar correlations of all other chromosomes are presented in Figures S10 (GAFA) and S11 (LAFA).

Table 4. Previously published Pima Indians allele frequencies, compared with our GAFA and LAFA Amerindian frequency estimations in the HCHS/SOL

						GAFA			LAFA		
SNP	CHR	POS (hg38)	Ref.	Alt.	Pima Indians freq.	Freq. est.	CI low	CI high	Freq. est.	CI low	CI high
rs75432840[a]	6	34,143,031	C	G	0.28	0.398	0.379	0.417	0.287	0.272	0.303
rs138977532[a]	6	36,382,025	C	T	0.59	0.635	0.620	0.650	0.587	0.572	0.602
rs139139046[a]	11	71,452,308	G	C	0.87	0.831	0.819	0.844	0.823	0.813	0.833
rs72849841[a]	17	80,298,494	C	T	1.00	0.987	0.978	0.996	0.996	0.994	0.998

Frequencies refer to the ref. allele.
[a]Sofer et al.,[3,17] 2017.

Evaluating the algorithm convergence rate of GAFA and LAFA by frequency boundary conditions

Summary statistics of HCHS/SOL alleles calculated using AFA versus alleles that failed calculation are presented in Table S6. For variants on chromosome 2, 92.3% were calculated using GAFA (n = 3,299,310,366) and 88.8% were calculated using LAFA (n = 3,175,914). Low MAF is likely the main reason for failed AF calculations. LAFA's successful calculation percentage is lower compared with GAFA, since the LAIs do not encompass the whole genome, and the liftover from GRCh38 to GRCh37 (in order to match each variant to its LAI) also failed for some variants. The number of overlapped calculated variants in both methods on chromosome 2 is n = 3,102,863, while n = 192,993 variants were successfully calculated only in GAFA and n = 70,641 variants were successfully calculated only in LAFA. This emphasizes the importance of developing both methods and their potential to complement each other.

Association of Amerindian-enriched variants with cardiometabolic traits

Results for association analyses of the Amerindian-enriched variants (n = 112,824) with 12 cardiometabolic traits are presented in Figure S12 (qq-plots) and Figure S13 (Manhattan plots). At the Bonferroni significance level of $0.05/112{,}824 = 4.4 \times 10^{-7}$, 13 variants were significantly associated with one or two outcomes (Table 6), comprising five independent loci. Table S7 summarizes the annotation for these variants using Functional Annotation of Variants – Online Resource (FAVOR). Three of these associations (rs17119918 with triglycerides, rs78950101, and rs4939873 with high-density lipoprotein [HDL]) would not pass the traditional genome-wide significance threshold of 5×10^{-8}, demonstrating the advantage of testing only the ancestry-enriched variants. Our main finding is a region spanning seven variants (~350 Kbp, with Amerindian allele frequencies between 0.13 and 0.32), located in chr11q23.3 and associated with triglycerides. One of the variants, rs191206329, was also associated with HDL. The most significant association was rs139961185 with triglycerides ($p = 1.4 \times 10^{-15}$), an imputed variant in an intronic region in the *SIK3* gene. Conditional analyses adjusting for this variant suggest this region is composed of two parts, one associated with higher triglycerides and lower HDL and the other region associated with lower triglycerides (Table 6). This variant was previously associated with triglycerides in a Mexican cohort[27] and a multi-ethnic non-European GWAS[28] showing similar direction effects. Another interesting region is chr1p13.3 (111 Kbp) showed a significant negative association for both total cholesterol and low-density lipoprotein (LDL), with Amerindian allele frequencies between 0.1 and 0.2. All four variants fall on intronic regions spanning three genes: *KIAA1324*, *SARS*, and *CELSR2*.

Global Lipids Genetics Consortium (GLGC) has recently published aggregated GWAS results, including results for n = ~48,000 Hispanic individuals, including the HCHS/SOL.[29] Table S8 presents these results for the lipid-associated variants that we identified in HCHS/SOL. All associations presented the same direction of effect as in HCHS/SOL and passed the traditional GWAS threshold of 5×10^{-8}, including the three associations identified only by the Bonferroni significance level, suggesting that these are true associations.

Discussion

We developed a method for estimating AFA based on either the rather widely available global proportion ancestry (GAFA) or LAIs (LAFA). Simulations have shown high accuracy of the estimated frequencies for both options, with increasing accuracy dependent on effective ancestry-specific sample size and MAF and with a slight advantage for LAFA over GAFA. We applied our method to the high-quality imputed genomic data of admixed Hispanics or Latinos from HCHS/SOL with three predominant continental ancestries, European, African, and Amerindian, and demonstrated speed, simplicity of calculation, and a highly successful ancestry-specific frequency estimation rate. This enabled us to select Amerindian-enriched variants and identify previously known associations with cardiometabolic-related traits in Hispanics or Latinos.

Comparison of the GAFA and LAFA European and African estimated ancestry-specific allele frequencies in HCHS/SOL to the respective gnomAD frequencies demonstrated strong positive correlations. We did not expect

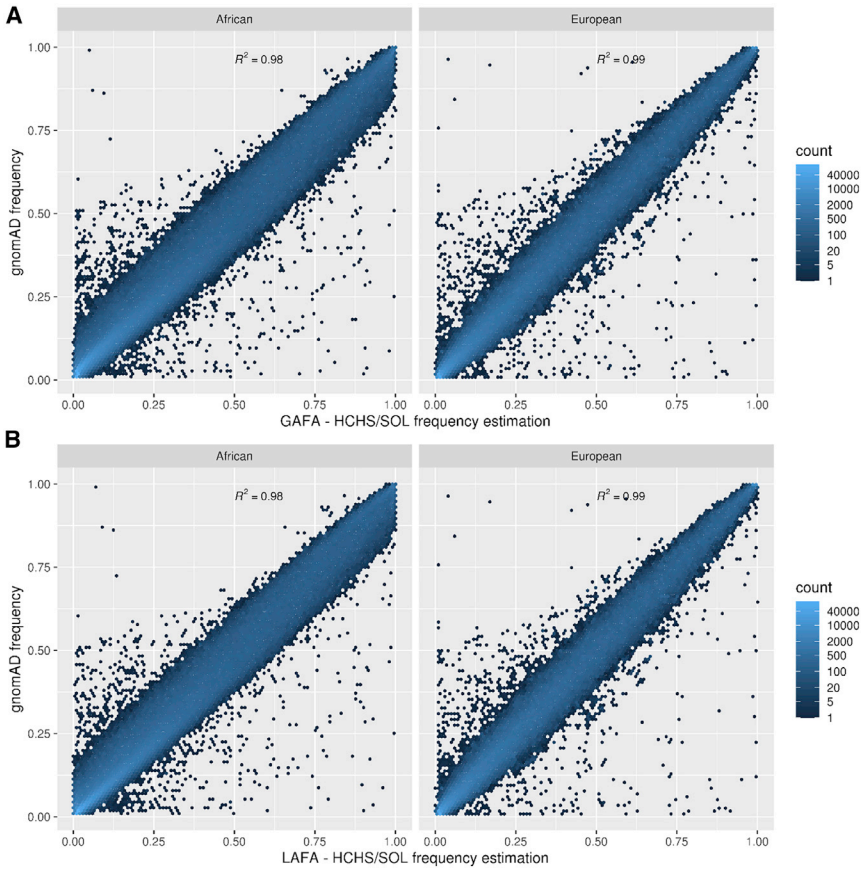

Figure 2. Ancestry-specific variant frequencies' comparison of HCHS/SOL (based on AFA) and gnomAD

Scatterplots of estimated ancestry-specific allele frequencies in HCHS/SOL chromosome 2 to corresponding gnomAD non-Finnish European and African frequencies, respectively, (A) using GAFA (no. variants: African = 1,239,958; European = 819,710) and (B) using LAFA (no. variants: African = 1,168,271; European = 775,749), stratified to genotyped variants and imputed variants with different imputation quality thresholds.

perfect correlation with the respective gnomAD frequencies, since evolutionary forces, such as genetic drift, mutagenesis, and natural selection, are expected to accumulate and result in frequency differences between the gnomAD allele frequencies and the Hispanics or Latinos ancestry-specific allele frequencies. Similarly, there are likely differences in ancestry-specific allele frequencies between the various Hispanic or Latino background groups, due to their different population histories. In this work, we did not evaluate such differences due to low power; however, we expect them to be minor, based on the results from the gnomAD comparison. The correlation between estimated European-specific frequencies and gnomAD non-Finnish Europeans frequencies is somewhat stronger compared with the respective African comparison. This is likely due to two reasons: first, individuals of African ancestries are characterized by a greater level of genetic diversity compared with Europeans,[30] so allele frequency comparisons between two populations of African ancestral origin will demonstrate a larger difference compared with frequency comparisons between two populations of European ancestral origin. Second, the effective sample size of European ancestry in the HCHS/SOL was substantially larger than the African effective sample size, enabling a more precise estimation of allele frequencies. We provide a genome-wide dataset of US Hispanic or Latino ancestry-specific allele frequencies estimated based on the HCHS/SOL for all variants with a frequency between 5% and 95% in at least one of the three ancestral populations, using GAFA (n = 9,808,089) and LAFA (n = 9,844,093). A previous publication based on sequence data of 66 Mexicans, 60 Colombians, and 55 Puerto Rican individuals from the 1000 Genomes project has published a list of estimated allele frequencies of the Native American ancestry proportion.[9] Here, we add a substantial increase in the sample size (n = ~9,000), in the diversity of the Hispanic or Latino background groups, and in the relative proportion of the Amerindian ancestry (30%), by using high-quality imputed genomic data based on a reference panel that includes whole-genome sequences of >8,000 Hispanic or Latino individuals. Inter-HCHS/SOL ancestry-specific allele frequencies present the strongest correlations between Amerindians and Europeans, followed by the Africans and Europeans, followed by Africans and Amerindians. These findings agree with the dominant paleoanthropology hypothesis of the African origin of modern humans, followed by migration to Europe, followed by other migrations to Asia and America.[31] Stronger bottlenecks (founder effect) in Amerindians led to more drifts and hence more differences in Amerindian compared with African frequencies. Thus, our dataset can serve as a unique resource for genetic epidemiology studies supporting research of personalized health in admixed populations.

We demonstrated the utilization of the ancestry-specific allele frequency in HCHS/SOL by conducting association analysis of Amerindian-enriched variants with 12 cardiometabolic traits. We detected 13 variants located in five association regions significantly associated with lipid traits. GWAS results from the GLGC for Hispanics (n = 48,000) support all our findings, demonstrating proof of concept. This highlights the advantage of focusing on ancestry-enriched variants when studying small understudied admixed populations, enabling more lenient thresholds and ancestry-unique findings. Further analysis of the associated variants and genes encompassing them is needed to determine biological insights.

Table 5. Percentage of non-Finnish European and African gnomAD frequencies included in the corresponding CI estimated in HCHS/SOL by GAFA and LAFA, binned by gnomAD frequency categories

		GAFA				LAFA			
	AF categories	In (%)	Out (%)	N[a]	CI length[b]	In (%)	Out (%)	N[a]	CI length[b]
African	<0.05	0.47	0.53	6,154,256	0.019	0.71	0.29	5,722,679	0.015
	0.05–0.1	0.47	0.53	2,484,933	0.032	0.53	0.47	2,306,660	0.023
	0.1–0.2	0.58	0.42	2,179,468	0.049	0.38	0.62	2,024,874	0.031
	0.2–0.3	0.53	0.47	1,139,908	0.063	0.31	0.69	1,059,791	0.039
	0.3–0.4	0.49	0.51	749,960	0.072	0.31	0.69	696,775	0.044
	0.4–0.5	0.48	0.52	557,597	0.076	0.31	0.69	516,965	0.046
	0.5–0.6	0.48	0.52	454,150	0.077	0.31	0.69	421,517	0.047
	0.6–0.7	0.48	0.52	404,407	0.074	0.30	0.70	376,436	0.045
	0.7–0.8	0.48	0.52	365,295	0.068	0.30	0.70	339,991	0.040
	0.8–0.9	0.49	0.51	341,980	0.056	0.29	0.71	319,042	0.032
	0.9–1	0.56	0.44	298,032	0.033	0.39	0.61	279,494	0.018
European	<0.01	0.81	0.19	764,738	0.009	0.64	0.36	713,699	0.006
	0.01–0.05	0.40	0.60	2,630,848	0.010	0.30	0.70	2,444,070	0.007
	0.05–0.1	0.31	0.69	1,180,341	0.016	0.22	0.78	1,097,724	0.011
	0.1–0.2	0.31	0.69	1,379,163	0.023	0.21	0.79	1,287,160	0.015
	0.2–0.3	0.31	0.69	913,809	0.028	0.21	0.79	853,249	0.018
	0.3–0.4	0.30	0.70	698,053	0.031	0.20	0.80	652,017	0.020
	0.4–0.5	0.31	0.69	555,041	0.033	0.21	0.79	518,228	0.021
	0.5–0.6	0.31	0.69	460,852	0.033	0.21	0.79	430,399	0.021
	0.6–0.7	0.31	0.69	396,445	0.032	0.21	0.79	369,703	0.021
	0.7–0.8	0.32	0.68	329,217	0.029	0.21	0.79	306,753	0.019
	0.8–0.9	0.32	0.68	272,693	0.024	0.21	0.79	254,014	0.016
	0.9–1	0.32	0.68	391,595	0.013	0.14	0.86	366,541	0.008

We assessed only gnomAD variants passing quality control filters (FILTER = "PASS"), with an ancestry-specific minor allele count of ≥100 respective to the assessed ancestry. The Europeans have an extra category for rare variants (<0.01), since their calculation is based on a larger dataset compared with Africans.
AF, allele frequency.
[a]Number of variants.
[b]Mean confidence interval lengths.

The advantages of our method are the ability to estimate ancestry-specific allele frequencies and CIs of genotyped or imputed variants in admixed populations with an unlimited number of ancestries, with no need for phased data, on a genome-wide scale. The estimated ancestry-specific frequencies are similar to their corresponding frequencies in ADMIXTURE; however, ADMIXTURE is tuned to apply to a limited number of pruned genome-wide variants only and does not produce accuracy estimates. Our method can be applied to imputed data; however, imputation quality depends on the representation of the ancestral populations in the reference panel and will likely affect the quality of the ancestry-specific allele frequency estimation. Therefore, we generally recommend applying this method on variants with high imputation quality and interpreting results with caution when the imputation quality is low. The algorithm is applicable for phased data as well. Thus, our method is simple, computationally efficient, versatile, and enables a wider usage compared with previous methods. It can be applied using either global proportions of genetic ancestries (GAFA) or LAI proportions encompassing the variant (LAFA). GAFA is a computationally simpler process compared with LAFA, and it encompasses all regions of the genome. However, it assumes a uniform distribution of ancestries throughout the genome, which is slightly less precise. Comparison of both GAFA and LAFA shows strong correlations for variants calculated by both methods and shows some variants could be calculated by using only one of the methods, complementing each other and emphasizing the advantage of using both options. Specifically, LAFA is more precise, but the algorithm may not converge when using LAFA so that frequency estimates were not obtained, while GAFA may converge for these variants. We think that this is likely due to local ancestry inference errors: when using LAFA, the ancestral probabilities assigned by the algorithm at

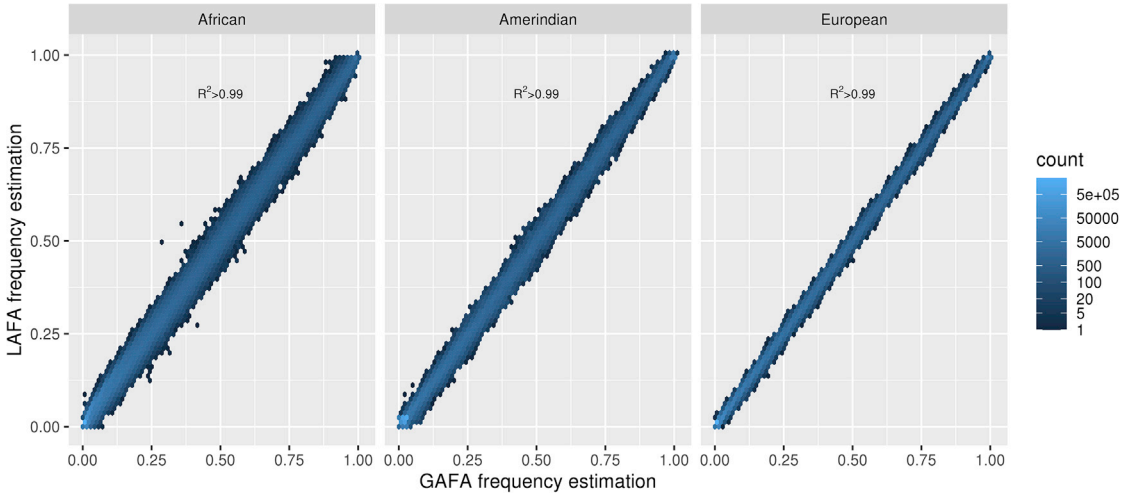

Figure 3. Scatterplots of the estimated ancestry-specific allele frequencies in chromosome 2 in the HCHS/SOL population between GAFA and LAFA for each of the three ancestral populations.

the segment take values $p_{i1},\ldots,p_{iK}\in\{0,0.5,1\}$. Thus, if in all LAIs from a specific ancestry, the observed MAC is 0, it may lead to non-convergence. Non-convergence may also arise from a lack of HWE in LAIs from a certain ancestry. Depending on effective population sample sizes, AFA may perform less well for low MAF variants. First, estimation depends on the effective sample sizes of the ancestral origins and the ancestry-specific frequencies (e.g., having enough counts). Second, AFA methods apply maximum-likelihood estimation of binomial likelihoods, which cannot be evaluated by the optimization algorithm at the boundaries of the parameter space (i.e., at frequencies of 0 or 1, though the likelihood is computable at the boundary). Therefore, very few minor allele counts in one of the genetic ancestries may lead to non-convergence of the algorithm, unless box constraints are placed (e.g., limiting the frequencies to

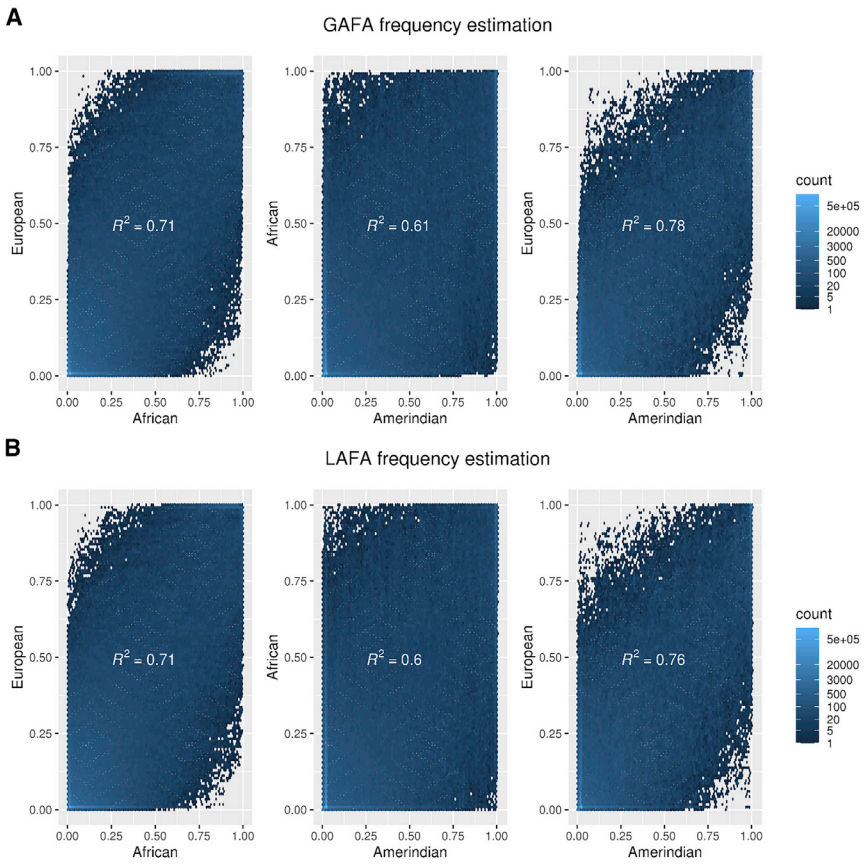

Figure 4. Scatterplots of the estimated ancestry-specific allele frequencies in chromosome 2 in the HCHS/SOL population between the three ancestral populations
(A) GAFA and (B) LAFA.

Table 6. Significant associations for Amerindian-enriched variants in the HCHS/SOL with cardiometabolic traits

									Ancestry-specific MAF based on GAFA HCHS/SOL			Association			Conditional analysis I			Conditional analysis II		
Trait	rsID	Chr	Position (hg38)	Effect allele	Non-effect allele	R^2[a]	MAC	MAF	African	European	Amerindian	Effect estimate	SE	p value	Effect estimate	SE	p value	Effect estimate	SE	p value
TC	rs142336293	1	109,152,113	G	A	0.99	1,387	0.059	0.005	0.006	0.193	−6.244	1.235	4.29×10^{-7}	−3.340	1.695	4.88×10^{-2}			
TC	rs146236384	1	109,220,029	T	TG	0.97	1,268	0.054	0.003	0.009	0.169	−7.203	1.309	3.78×10^{-8}	−28.735	11.908	0.0158[b]			
Trig	rs17119918	11	116,714,897	A	G	0.99	998	0.043	0.003	0.002	0.144	−15.844	2.997	1.24×10^{-7}	−14.444	2.993	1.40×10^{-6}	16.363	10.932	1.34×10^{-1}
Trig	rs146714678	11	116,726,114	G	T	0.99	981	0.042	0.003	0.003	0.141	−17.789	3.015	3.65×10^{-9}	−16.111	3.015	9.16×10^{-8}	17.732	74.427	0.812[b]
Trig	rs141882698	11	116,730,622	C	A	0.99	985	0.042	0.003	0.003	0.142	−17.729	3.011	3.92×10^{-9}	−16.026	3.012	1.03×10^{-7}	−1.720	32.901	9.58×10^{-1}
Trig	rs191206329	11	116,755,683	G	C	1.00	1,649	0.070	0.002	0.006	0.228	18.516	2.384	8.09×10^{-15}	12.003	2.987	5.85×10^{-5}	17.246	2.397	6.22×10^{-13}
Trig	rs145796806	11	116,779,468	T	C	0.99	959	0.041	0.003	0.003	0.138	−17.885	3.056	4.83×10^{-9}	−16.229	3.055	1.09×10^{-7}	−4.247	13.520	7.53×10^{-1}
Trig	rs139961185	11	116,936,627	A	G	0.95	2,368	0.101	0.003	0.004	0.329	16.558	2.073	1.40×10^{-15}	35.099	16.535	0.0338[b]	15.832	2.077	2.47×10^{-14}
Trig	rs144818596	11	117,067,138	G	T	0.91	2,280	0.097	0.003	0.003	0.318	16.886	2.131	2.31×10^{-15}	11.886	6.562	7.01×10^{-2}	16.226	2.134	2.85×10^{-14}
HDL	rs191206329	11	116,755,683	G	C	1.00	1,650	0.070	0.002	0.006	0.228	−1.770	0.320	3.09×10^{-8}						
HDL	rs78950101	16	56,898,811	C	T	0.98	2,203	0.094	0.009	0.005	0.310	−1.475	0.289	3.35×10^{-7}						
HDL	rs4939873	18	49,535,684	T	G	0.95	2,047	0.087	0.007	0.005	0.283	1.589	0.297	9.29×10^{-8}						
LDL	rs142336293	1	109,152,113	G	A	0.99	1,387	0.059	0.005	0.006	0.193	−5.885	1.068	3.60×10^{-8}	−1.928	1.467	1.89×10^{-1}			
LDL	rs1815307	1	109,182,202	T	C	0.97	1,258	0.054	0.004	0.006	0.173	−6.693	1.133	3.45×10^{-9}	−1.534	2.015	4.46×10^{-1}			
LDL	rs146236384	1	109,220,029	T	TG	0.97	1,268	0.054	0.003	0.009	0.169	−7.773	1.133	6.80×10^{-12}	−27.076	10.308	0.0086[b]			
LDL	rs189575997	1	109,263,516	T	C	0.90	686	0.029	0.005	0.006	0.092	−8.572	1.563	4.16×10^{-8}	−3.730	1.943	5.49×10^{-2}			

HDL, high-density lipoprotein; LDL, low-density lipoprotein; MAC, minor allele count; MAF, minor allele frequency; SE, standard error; TC, total cholesterol; Trig, triglycerides.
[a]Imputation quality.
[b]Index variant for conditional analysis.

be estimated within the interval [0.01, 0.99]), so that frequencies outside the interval cannot be estimated.

Data and code availability

We provide a publicly available GitHub repository, https://github.com/tamartsi/Ancestry_specific_freqs, which includes (1) code for GAFA and LAFA for computing ancestry-specific allele frequencies, (2) simulation code, and (3) a dataset of Hispanic or Latino ancestry-specific allele frequencies and their CIs estimated based on the HCHS/SOL using GAFA and LAFA for all variants (genotyped or imputed) with an estimated frequency between 5% and 95% in at least one of the three ancestral populations. This dataset will also be available through FAVOR v.2 data release in both the single variant query (Allele Frequency Block) and batch query, http://favor.genohub.org. CWL workflows for GAFA and LAFA are also available via dockstore and https://github.com/cwl-apps/ancestral-maf-admixed-population.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100096.

Acknowledgments

The authors thank the staff and participants of HCHS/SOL for their important contributions. The investigator's website is http://www.cscc.unc.edu/hchs/. The HCHS/SOL is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute to the University of North Carolina (HHSN268201300001I/N01-HC-65233), University of Miami (HHSN268201300004I/N01-HC-65234), Albert Einstein College of Medicine (HHSN 268201300002I/N01-HC-65235), University of Illinois at Chicago (HHSN268201300003I/N01- HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I/N01-HC-65237). The following institutes, centers, and offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, and NIH Institution-Office of Dietary Supplements. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03). The authors would also like to acknowledge the use of the TransOmics in Precision Medicine (TOPMed) program imputation panel (freeze-8 version), supported by the National Heart, Lung and Blood Institute (NHLBI); see http://www.nhlbiwgs.org. TOPMed study investigators contributed data to the reference panel, which was accessed through https://imputation.biodatacatalyst.nhlbi.nih.gov. The panel was constructed and implemented by the TOPMed Informatics Research Center at the University of Michigan (3R01HL-117626-02S1; contract HHSN268201800002I). The TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I) provided additional data management, sample identity checks, and overall program coordination and support. We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. Support for this work was provided by the National Institutes of Health, National Heart, Lung, and Blood Institute, through the BioData Catalyst program (awards 1OT3HL142479-01, 1OT3HL142478-01, 1OT3HL142481-01, 1OT3HL142480-01, and 1OT3HL147154-01). Any opinions expressed in this document are those of the author(s) and do not necessarily reflect the views of NHLBI, individual BioData Catalyst team members, or affiliated organizations and institutions.

Declaration of interests

The authors declare no competing interests.

Received: October 7, 2021
Accepted: February 18, 2022

Web resources

BioData Catalyst: https://biodatacatalyst.nhlbi.nih.gov/
FAVOR: http://favor.genohub.org
GitHub: https://github.com/tamartsi/Ancestry_specific_freqs
GitHub: https://github.com/cwl-apps/ancestral-maf-admixed-population

References

1. Montinaro, F., Busby, G.B.J., Pascali, V.L., Myers, S., Hellenthal, G., and Capelli, C. (2015). Unravelling the hidden ancestry of American admixed populations. Nat. Commun. *6*, 1–7.
2. Atkinson, E.G., Maihofer, A.X., Kanai, M., Martin, A.R., Karczewski, K.J., Santoro, M.L., Ulirsch, J.C., Kamatani, Y., Okada, Y., Finucane, H.K., et al. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. Nat. Genet. *53*, 195–204.
3. Sofer, T., Baier, L.J., Browning, S.R., Thornton, T.A., Talavera, G.A., Wassertheil-Smoller, S., Daviglus, M.L., Hanson, R., Kobes, S., Cooper, R.S., et al. (2017). Admixture mapping in the Hispanic community health study/study of Latinos reveals regions of genetic associations with blood pressure traits. PLoS One *12*, e0188400.
4. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical genetics and genomics and the association for molecular pathology. Genet. Med. *17*, 405–424.
5. Choudhury, A., Hazelhurst, S., Meintjes, A., Achinike-Oduaran, O., Aron, S., Gamieldien, J., Jalali Sefid Dashti, M., Mulder, N., Tiffin, N., and Ramsay, M. (2014). Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. BMC Genomics *15*, 437.
6. Uren, C., Hoal, E.G., and Möller, M. (2020). Putting RFMix and ADMIXTURE to the test in a complex admixed population. BMC Genet. *21*, 1–8.
7. Falush, D., Stephens, M., and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics *164*, 1567–1587.

8. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
9. Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J.K., Muzzio, M., Rodriguez-Flores, J.L., Kenny, E.E., Gignoux, C.R., Maples, B.K., Guiblet, W., et al. (2013). Reconstructing native American migrations from whole-genome and whole-exome data. PLoS Genet. *9*, 1004023.
10. Zhang, Q.S., Browning, B.L., Browning, S.R., and Stegle, O. (2016). Genetics and population analysis ASAFE: ancestry-specific allele frequency estimation. Bioinformatics *32*, 2227–2229.
11. Zhang, Q.S. (2018). Statistical Genetic Methods and Applications for Population Structure (Ph.D. Dissertation). https://books.google.com/books?id=QOL5wQEACAAJ.
12. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernández-Rhodes, L., Justice, A.E., Graff, M., et al. (2016). Genetic diversity and association studies in US hispanic/Latino populations: applications in the hispanic community health study/study of latinos. Am. J. Hum. Genet. *98*, 165–184.
13. Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T., Conomos, M.P., Schurmann, C., McHugh, C.P., et al. (2016). Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. Am. J. Hum. Genet. *98*, 229–242.
14. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
15. Lavange, L.M., Kalsbeek, W.D., Sorlie, P.D., Avilés-Santa, L.M., Kaplan, R.C., Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. (2010). Sample design and cohort selection in the hispanic community health study/study of Latinos. Ann. Epidemiol. *20*, 642–649.
16. Sorlie, P.D., Avilé S-Santa, L.M., Wassertheil-Smoller, S., Kaplan, R.C., Daviglus, M.L., Giachello, A.L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., et al. (2010). Design and implementation of the Hispanic community health study/study of Latinos. Ann. Epidemiol. *20*, 629–641.
17. Sofer, T., Wong, Q., Hartwig, F.P., Taylor, K., Warren, H.R., Evangelou, E., Cabrera, C.P., Levy, D., Kramer, H., Lange, L.A., et al. (2017). Genome-wide association study of blood pressure traits by Hispanic/Latino background: the Hispanic community health study/study of Latinos. Sci. Rep. *7*, 10348.
18. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M., Arnett, D.K., Avery, C., et al. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. *15*, e1008500.
19. Maintainer, B.P. (2021). liftOver: Changing Genomic Coordinate Systems with rtracklayer::liftOver. R Package Version 1.16.0. https://bioconductor.org/packages/release/workflows/html/liftOver.html.
20. Browning, S.R., Grinde, K., Plantinga, A., Gogarten, S.M., Stilp, A.M., Kaplan, R.C., Avilés-Santa, M.L., Browning, B.L., and Laurie, C.C. (2016). Local ancestry inference in a large US-based Hispanic/Latino study: Hispanic community health study/study of Latinos (HCHS/SOL). G3 (Bethesda) *6*, 1525–1534.
21. Deb, K., and Jain, H. (2014). An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, Part I: solving problems with box constraints. IEEE Trans. Evol. Comput. *18*, 577–601.
22. Burkart, K.M., Sofer, T., London, S.J., Manichaikul, A., Hartwig, F.P., Yan, Q., Artigas, M.S., Avila, L., Chen, W., Thomas, S.D., et al. (2018). A genome-wide association study in hispanics/latinos identifies novel signals for lung function the hispanic community health study/study of latinos. Am. J. Respir. Crit. Care Med. *198*, 208–219.
23. Jian, X., Sofer, T., Tarraf, W., Bressler, J., Faul, J.D., Zhao, W., Ratliff, S.M., Lamar, M., Launer, L.J., Laurie, C.C., et al. (2020). Genome-wide association study of cognitive function in diverse Hispanics/Latinos: results from the Hispanic community health study/study of Latinos. Transl. Psychiatry *10*, 1–13.
24. Hatzikotoulas, K., Gilly, A., and Zeggini, E. (2014). Using population isolates in genetic association studies. Brief. Funct. Genomics *13*, 371–377.
25. National Heart, Lung, and Blood Institute; National Institutes of Health; and U.S. Department of Health and Human Services (2020). The NHLBI BioData Catalyst (Zenodo). https://doi.org/10.5281/zenodo.3822858.
26. Amstutz, P., Crusoe, M.R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Kern, J., Leehr, D., Ménager, H., et al. (2016). Common Workflow Language, v1.0. Specification (Common Workflow Language Working Group).
27. Ko, A., Cantor, R.M., Weissglas-Volkov, D., Nikkola, E., Reddy, P.M.V.L., Sinsheimer, J.S., Pasaniuc, B., Brown, R., Alvarez, M., Rodriguez, A., et al. (2014). Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nat. Commun. *5*, 3983.
28. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514.
29. Graham, S.E., Clarke, S.L., Wu, K.-H.H., Kanoni, S., Zajac, G.J.M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. Nature *600*, 1–11.
30. Campbell, M.C., and Tishkoff, S.A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu. Rev. Genomics Hum. Genet. *9*, 403–433.
31. Henn, B.M., Cavalli-Sforza, L.L., and Feldman, M.W. (2012). The great human expansion. Proc. Natl. Acad. Sci. U S A *109*, 17758–17764.