*Bioinformatics*, 32(14), 2016, 2227–2229
doi: 10.1093/bioinformatics/btw220
Advance Access Publication Date: 3 May 2016
Applications Note

OXFORD

Genetics and population analysis

# ASAFE: ancestry-specific allele frequency estimation

**Qian S. Zhang[1,2,*], Brian L. Browning[2,3] and Sharon R. Browning[2]**

[1]Department of Medicine, [2]Department of Biostatistics and [3]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on August 31, 2015; revised on March 16, 2016; accepted on April 17, 2016

## Abstract

**Summary:** In a genome-wide association study (GWAS) of an admixed population, such as Hispanic Americans, ancestry-specific allele frequencies can inform the design of a replication GWAS. We derive an EM algorithm to estimate ancestry-specific allele frequencies for a bi-allelic marker given genotypes and local ancestries on a 3-way admixed population, when the phase of each admixed individual's genotype relative to the pair of local ancestries is unknown. We call our algorithm Ancestry Specific Allele Frequency Estimation (ASAFE). We demonstrate that ASAFE has low error on simulated data.

**Availability and implementation:** The R source code for ASAFE is available for download at https://github.com/BiostatQian/ASAFE

**Contact:** qszhang@uw.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Genome-wide association studies (GWASs) in humans have focused on populations of European origin. Hispanic Americans however, have descended from ancestral Africans, Europeans and Native Americans. Hispanics exemplify an admixed population, one that has descended from multiple ancestral populations.

After discovering that an allele is significantly associated with a trait in a GWAS of admixed individuals, it would be useful to know the allele's ancestry-specific frequencies. For a sample of Hispanic individuals, an ancestry-specific allele frequency would be the frequency of the allele amongst the chromosomes in the sample that have a particular ancestral origin (either African, European, or Native American) at the marker. Such frequencies provide information about which populations the discovered association might also be present in, thereby informing the choice of study population for replication.

To obtain estimates of ancestry-specific allele frequencies, one might find the allele frequency in each reference panel. However, this approach is impossible for a marker that is present in the admixed sample, but absent from a reference panel. This situation arises when the reference panels are genotyped on a SNP array and when the admixed individuals have sequence data. Even for a marker that is typed in the reference panels, this approach suffers when the reference panels do not exactly match the ancestral populations. For example, none of the populations in the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) are non-admixed Native American. The ADMIXTURE and STRUCTURE programs can use admixed genotypes to estimate ancestry-specific allele frequencies, but require genotypes from reference individuals of known ancestry in order to identify the frequencies with particular ancestries. (Alexander *et al.*, 2009; Pritchard *et al.*, 2000) Furthermore, ADMIXTURE assumes linkage equilibrium amongst markers, and STRUCTURE's linkage model assumes linkage equilibrium amongst markers descended from the same ancestral population. (Falush *et al.*, 2003) In contrast, we utilize estimates of local ancestry, such as those from the RFMix program (Maples *et al.*, 2013). Estimates of local ancestry from such programs utilize haplotype frequencies, and thus make use of linkage disequilibrium. At markers genotyped in all the reference panels, one can directly obtain ancestry specific allele frequencies from the phased genotypes and corresponding phased local ancestry estimates. However, at

**Table 1.** Mean and SD of errors (error = estimated ancestry-specific allele frequency – true ancestry-specific allele frequency) for 56 003 SNPs, grouped by true ancestry-specific allele frequency bins for African, European (Eur.), and Native American (Nat. Am.) ancestries

| | | True Ancestry-Specific Allele Frequency Bins | | | | |
|---|---|---|---|---|---|---|
| Ancestry | Statistic | (0–0.2] | (0.2–0.4] | (0.4–0.6] | (0.6–0.8] | (0.8–1] |
| African | Mean | −0.0011 | −0.0003 | −0.0004 | 0.0004 | −0.0004 |
| African | SD | 0.0065 | 0.0185 | 0.0233 | 0.0186 | 0.0118 |
| Eur. | Mean | −0.0015 | −0.0004 | −0.0007 | −0.0010 | <0.0001 |
| Eur. | SD | 0.0077 | 0.0209 | 0.0249 | 0.0220 | 0.0122 |
| Nat. Am. | Mean | −0.0004 | −0.0017 | 0.0021 | 0.0048 | 0.0007 |
| Nat. Am. | SD | 0.0083 | 0.0235 | 0.0238 | 0.0257 | 0.0118 |

markers not genotyped in a reference panel, the phasing of the local ancestries relative to the genotypes is unknown and further methodology is needed to obtain ancestry specific allele frequency estimates.

Here, we derive an EM algorithm (ASAFE) to estimate ancestry-specific allele frequencies at a bi-allelic marker in a three-way admixed diploid population, given admixed local ancestries and genotypes at the marker, with each admixed individual's genotype allele order relative to ancestry order unknown. The major advantage of ASAFE over alternative ancestry-specific allele frequency estimation approaches is that ASAFE is applicable to markers in the admixed sample that are absent from a reference panel. Furthermore, ASAFE takes advantage of linkage-disequilibrium based information by using local ancestry calls.

Gravel *et al.* (2013) describe a somewhat similar algorithm to ASAFE, but only for two-way admixture, and in the slightly different setting of estimating the ancestral allele frequency in several closely related ancestral populations (such as the Native American ancestors of Columbians, Puerto Ricans and Mexicans). They do not provide a publicly available implementation of their method.

We provide a publicly available implementation of our algorithm, and a description of its performance on simulated Hispanic data. Our algorithm applies to any diploid admixed population descended from three ancestral populations, so our method is not restricted to human populations.

## 2 Methods

A local ancestry inference program such as RFMix (Maples *et al.*, 2013) can be used to estimate the ancestry across the genome in an admixed sample. Local ancestral segments tend to extend over long genomic distances in recently admixed populations because they are ended only by recombination since the admixture event. Thus although the ancestry may only be estimated at positions that are genotyped in both the admixed and reference individuals, the local ancestry at most intermediate positions can be inferred with confidence.

Given local ancestries and genotypes at each SNP, ancestry-specific allele frequencies can be estimated independently for each SNP, using an EM algorithm that handles unknown genotype phase relative to ancestry phase to estimate three ancestry-specific allele frequencies for a bi-allelic marker. We call this algorithm Ancestry Specific Allele Frequency Estimation (ASAFE). Input data to ASAFE are individuals' ancestry pairs for the SNPs for which one would like estimates, and the same individuals' genotypes at these SNPs. A derivation of the EM algorithm is given in the Supplementary Information.

We simulated 10 Mb of sequence data for individuals from each of three populations representing Europeans, West Africans and Native Americans. We used 250 individuals from each population as reference individuals, and created 250 admixed individuals from additional simulated individuals. Details are in the Supplementary Information.

We then applied ASAFE to the admixed data, treating genotypes as unphased and using the known ancestries of the simulated admixed individuals in the analysis. We gave ASAFE admixed genotypes and ancestries at 56 003 SNPs, and for each SNP, obtained the allele frequencies for each of three ancestries, African, European, and Native American. The computation time was 0.5 h on a Linux server with a Intel Xeon CPU E5-2630L 2.0 GHz processor.

We assessed the accuracy of ASAFE in the following way. For each SNP, we estimated ancestry-specific allele frequencies for the derived allele. For each ancestry, at each SNP, we subtracted the true ancestry-specific allele frequency from the estimated ancestry-specific allele frequency. We considered the true ancestry-specific allele frequency to be the sample frequency of the allele amongst admixed chromosomes descended from that ancestry. To see if the accuracy of ASAFE might differ depending on the true ancestry-specific allele frequency, we binned SNPs by their true ancestry-specific allele frequencies.

## 3 Results

Table 1 gives summary statistics on ASAFE's error in estimating ancestry-specific allele frequencies for 250 admixed individuals. ASAFE has low error across all allele frequency bins and ancestries.

Supplementary Table S1 shows error summary statistics for various combinations of the three ancestral population allele frequencies. The tested frequency combinations cover a broad range of possibilities that might be observed for populations more or less diverged than our simulated Hispanic scenario. Accuracy remains high in each case. Mean errors (bias) are at most 0.001 in absolute value and SDs are at most 0.03, with the highest SDs being observed when the three frequencies are close to each other. Supplementary Table S2 shows error summary statistics when local ancestry is called with error. Adding 7% diploid error to the local ancestry calls leads to mean errors (bias) of at most 0.05 in absolute value while SDs are at most 0.04.

## 4 Conclusion

As more GWASs are performed in admixed populations such as Hispanics, it is becoming increasingly important to estimate ancestry-specific allele frequencies for bi-allelic markers. We derived an EM algorithm to estimate ancestry-specific allele frequencies given data on a 3-way admixed population and provide a publicly available implementation of the algorithm.

## Funding

This study was supported by research grant GM099568 from the National Institutes of Health, USA. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03).

*Conflict of Interest*: none declared.

## References

Alexander,D.H. *et al*. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*., **19**, 1655–1664.

Falush,D. *et al*. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Gravel,S. *et al*. (2013) Reconstructing Native American migrations from whole-genome and whole-exome data. *PLoS Genet*., **9**, e1004023.

Maples,B.K. *et al*. (2013) RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet*., **93**, 278–288.

Pritchard,J.K. *et al*. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.