Ancestry-Specific Allele Frequency Estimation (ASAFE) [2]

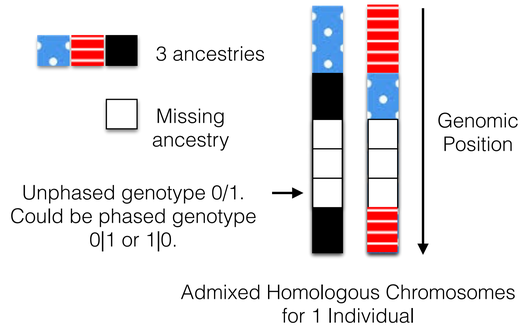
Qian Zhang[1]; Brian Browning, PhD[1,2]; Sharon Browning, PhD[1]
[1]University of Washington, Department of Biostatistics. [2]University of Washington, Department of Medicine, Division of Medical Genetics.

1. Introduction

Hispanic Community Health Study
- Hispanics descended from Africans, Europeans, and Native Americans
-Given a significant, trait-associated single nucleotide polymorphism (SNP) in a genome-wide association study (GWAS) of Hispanics, want to know e.g. allele 1's frequency amongst chromosomes of African, European, or Native American origin at the SNP, to design a follow-up GWAS

2. Data

RFMix [1] calls ancestries for Hispanics only at SNPs typed in African, European, Native American, and Hispanic samples.



Available data: Hispanic individuals'
- Unphased genotypes for all bi-allelic SNPs
- Phased ancestries for some SNPs

3. Methods

(1) Filled in missing ancestries

(2) Used EM algorithm (ASAFE) to get, for each SNP, maximum likelihood estimates of P(Allele 1 | African), P(Allele 1 | European), P(Allele 1 | Native American)

EM algorithm deals with not knowing the genotype order relative to the ancestry pair. Example: If we see an individual with genotype 0/1 and ancestry pair African|European, the ancestry-specific genotype could be (0, African) / (1, European) or (1, African) / (0, European).

Assessing ASAFE

(1) Simulated 56,003 SNPs for Hispanics

(2) Ran ASAFE with unphased admixed genotypes and phased admixed ancestries → Got ancestry-specific allele 1 frequencies for each ancestry (African, European, Native American) at each SNP

4. Results

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

		True P(Allele 1 | Ancestry) 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
European	Mean	-0.0015	-0.0004	-0.0007	-0.0010	<0.0001
European	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

5. Conclusions

- Developed EM algorithm ASAFE to estimate ancestry-specific allele 1 frequencies: P(Allele 1 | Ancestry), where Ancestry can take 3 values
- ASAFE has low error, regardless of true P(Allele 1 | Ancestry) frequency (Table 1)
- R package on GitHub & Bioconductor

6. References

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

2. ASAFE: Ancestry-Specific Allele Frequency Estimation. Qian S. Zhang; Brian L. Browning; Sharon R. Browning. Bioinformatics 2016; doi: 10.1093/bioinformatics/btw220.

7. Acknowledgments

This research was funded in part by:

1) NIH Grant HG005701

2) Biostatistics, Epidemiologic and Bioinformatic Training in Environmental Health Grant ES015459

3) UW Medical Scientist Training Program

SCHOOL OF PUBLIC HEALTH