Ancestry Specific Allele Frequency Estimation (ASAFE)

Qian Sophia Zhang

Joint work with Dr. Sharon Browning and Dr. Brian Browning
Department of Biostatistics
University of Washington
Slides @ http://biostatqian.github.io/ASAFE/

August 1, 2016

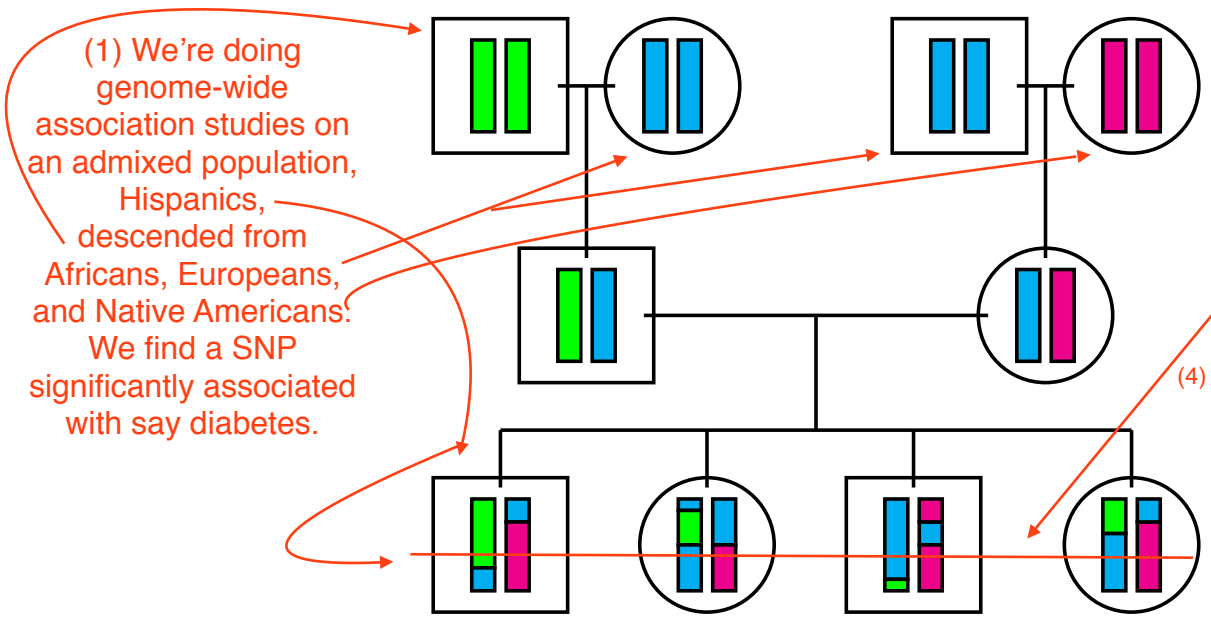
Some commentary is typed in, in red like this.
On each slide, the order in which I say things is numbered.

Outline

- What ASAFE Does
- Data Available
- ASAFE has Low Error on Simulated Data
- Sources for More Info

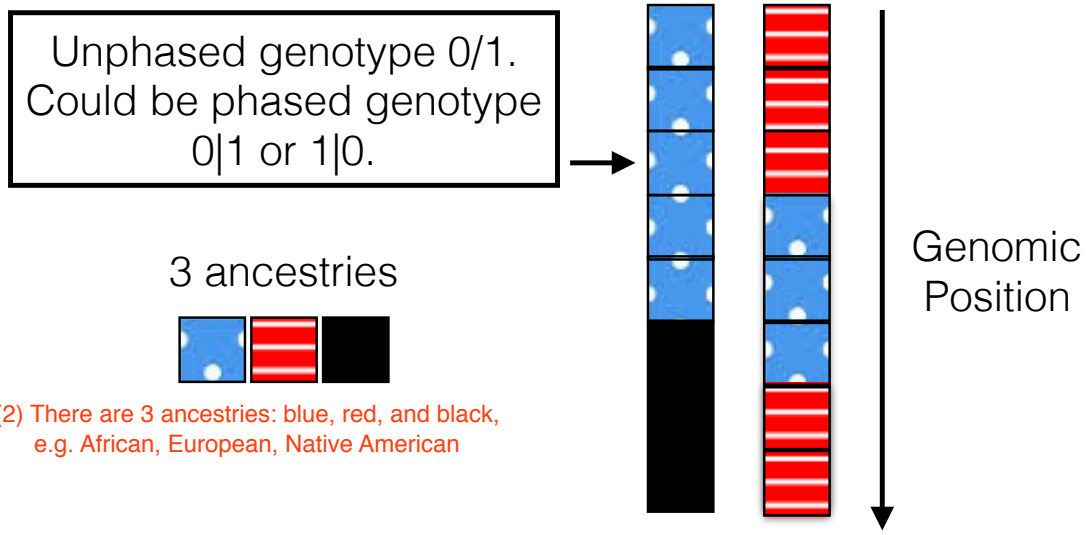
ASAFE: EM Algorithm for Estimating Ancestry Specific Allele Frequencies at a SNP

- [2] For a significant SNP (with alleles 0 and 1), want ancestry-specific allele frequencies := P(Allele 1 | African), P(Allele 1 | European), and P(Allele 1 | Native American), i.e. frequencies of allele 1 amongst chromosomes of African, European, or Native American origin at the SNP



Available Data: ASAFE assumes we know bi-allelic unphased genotypes and phased ancestries, but not the genotype order relative to ancestry pair

(3) Now consider this SNP. It has phased ancestries blue bar red.
The unphased genotype is 0 slash 1, so could be 0 bar 1 or 1 bar 0. We don't know if the 1 goes with the blue and the 0 with the red, or the 0 with the blue and the 1 with the red. This is fundamentally the issue that ASAFE deals with in estimating frequencies.



(2) There are 3 ancestries: blue, red, and black,
e.g. African, European, Native American

(1) Consider 2 chromosomes for 1 Hispanic individual.

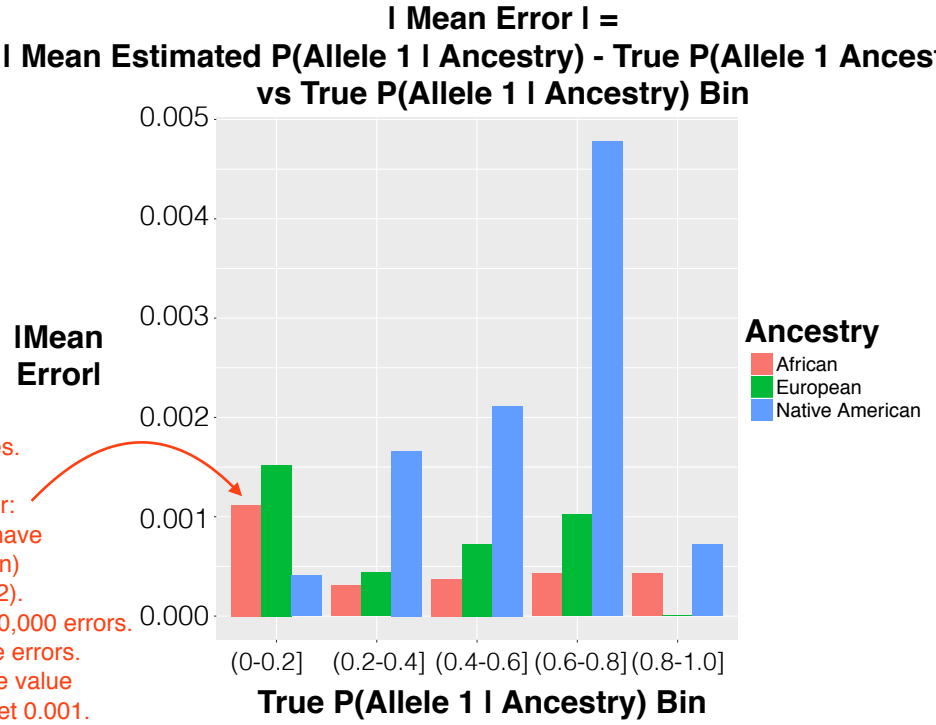
Error Calculation on Simulated Data

- Applied ASAFE to simulated genotypes and ancestries for $56,003$ SNPs in a Hispanic sample $\rightarrow$ Got ancestry-specific allele 1 frequencies for each ancestry (African, European, Native American), at each SNP
- For each SNP and ancestry $a \in \{A, E, N\}$, calculated error $=$ Estimated $p_{1|a}$ - True $p_{1|a}$

Low Error on Simulated Data

- For each a $\in \{A, E, N\}$, grouped SNPs errors by True P(Allele 1 | ancestry) bin, and took the mean of errors within the same bin
- (4) Largest $|\text{Mean of Errors}| = 0.005$. Largest SD of Errors $= 0.03$.

(1) For each ancestry, African, European, or Native American, I grouped SNPs' errors by True P(Allele 1 I ancestry) bin, and took the mean of errors within the same group



(2) Name the axes.

(3) Interpret a bar:
Say 10,000 SNPs have
P(Allele 1 I African)
in the range (0-0.2).
Those 10,000 SNPs have 10,000 errors.
Take the mean of those errors.
Then take the absolute value
of the mean error and get 0.001.

More Info: Paper, Presentation, and Code

- **Poster**: Poster Number 13 at Session Number 213225 at 10:30 AM in CC-Hall F1 West.
- **Paper**: Qian S. Zhang, Brian L. Browning, and Sharon R. Browning. Asafe: ancestry-specific allele frequency estimation. Bioinformatics, 32(14):2227 2229, 2016.
- **R package "ASAFE"**: On Bioconductor
- **Presentation slides, code to reproduce analysis**: http://biostatqian.github.io/ASAFE/