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

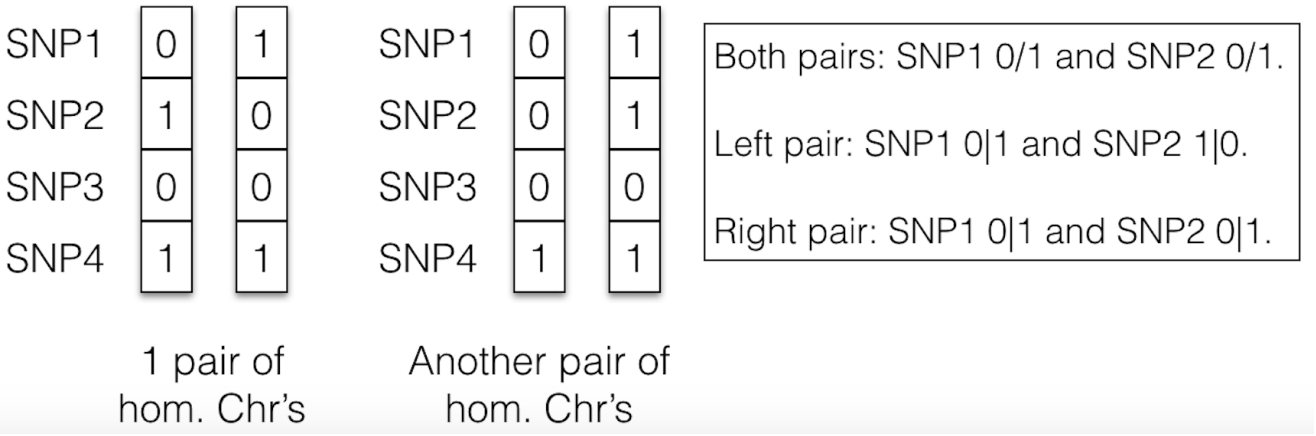
July 12, 2016

Some commentary is typed in, in red like this.

Genetic Terminology

Human genomes are packaged into pairs of homologous chromosomes. 2 pairs shown below. Rows are SNPs.

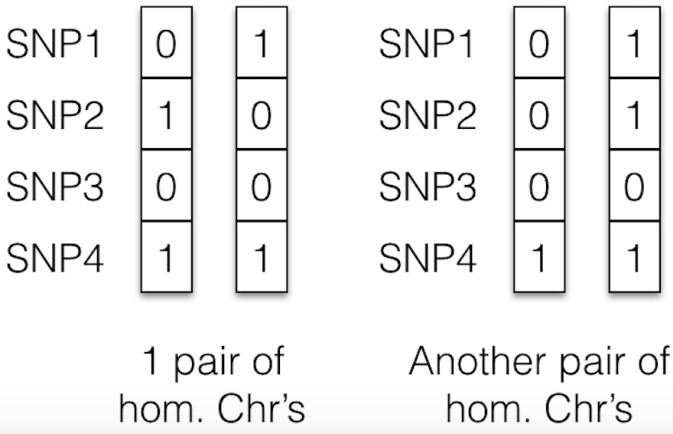
- "SNP" := Point along a chromosome where genomes differ. Often there are two variants or "alleles" at a SNP, labeled 0 and 1.



Genetic Terminology

Human genomes are packaged into pairs of homologous chromosomes. 2 pairs shown below. Rows are SNPs.

- "SNP" := Point along a chromosome where genomes differ. Often there are two variants or "alleles" at a SNP, labeled 0 and 1.
- "Genotype" := 2 homologous chromosomes' alleles at a SNP Ex: SNP1's genotype is 0/1, or 0|1 or 1|0. / and | denote phase.



Both pairs: SNP1 0/1 and SNP2 0/1.

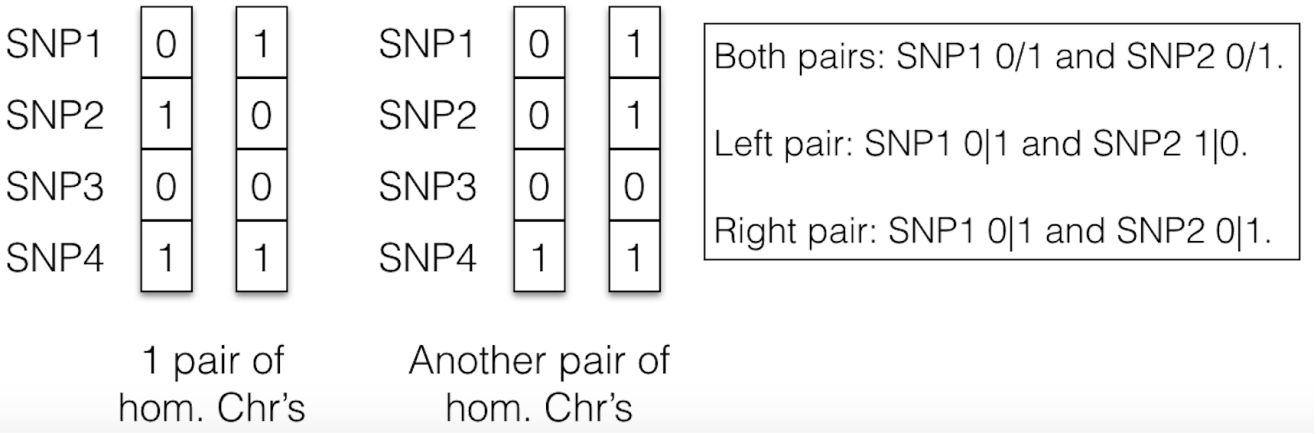
Left pair: SNP1 0|1 and SNP2 1|0.

Right pair: SNP1 0|1 and SNP2 0|1.

Genetic Terminology

Human genomes are packaged into pairs of homologous chromosomes. 2 pairs shown below. Rows are SNPs.

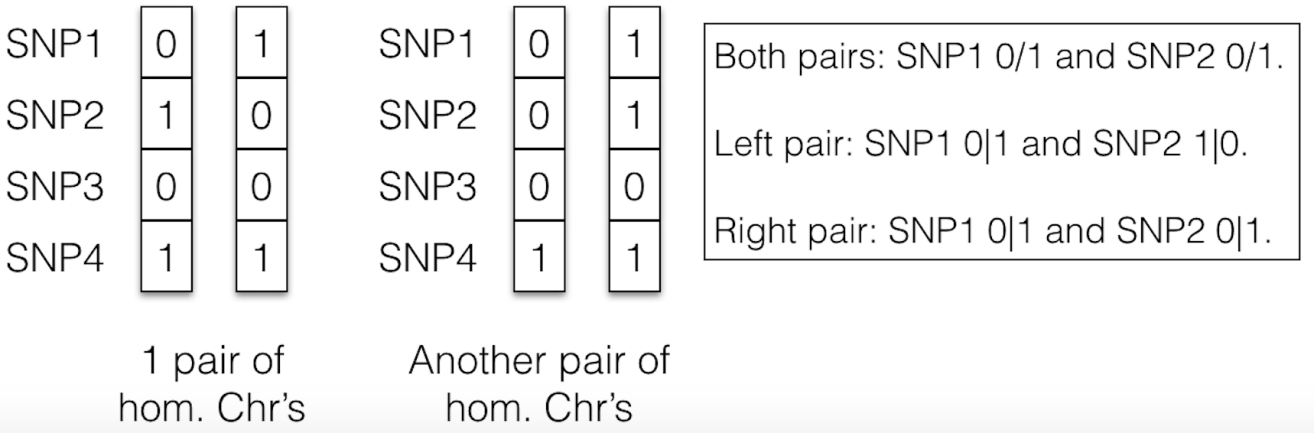
- "Unphased genotype" (/): SNP's genotype is NOT ordered with respect to another SNP's genotype



Genetic Terminology

Human genomes are packaged into pairs of homologous chromosomes. 2 pairs shown below. Rows are SNPs.

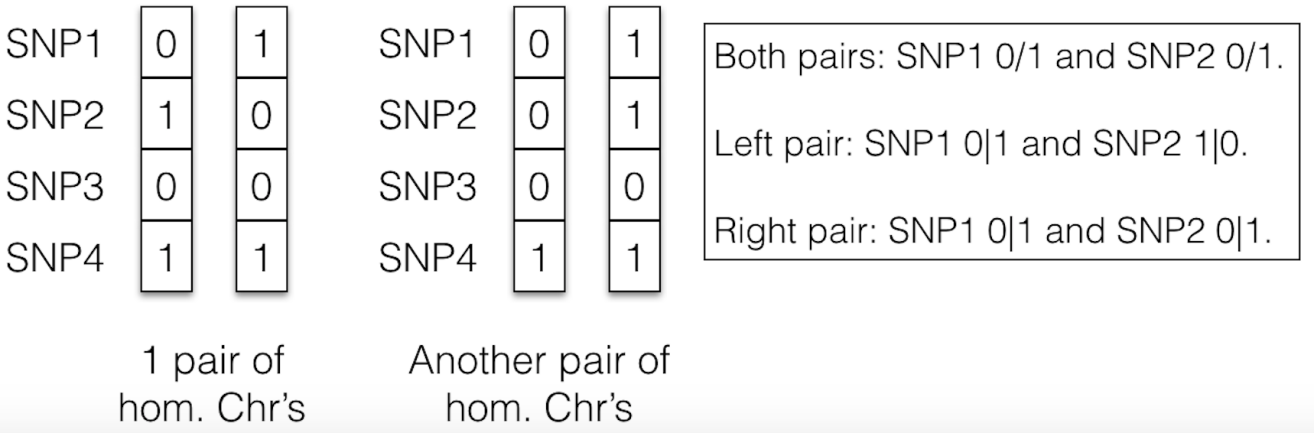
- "Unphased genotype" (/): SNP's genotype is NOT ordered with respect to another SNP's genotype
- "Phased genotype" (|): SNP's genotype IS ordered with respect to another SNP's genotype



Genetic Terminology

Human genomes are packaged into pairs of homologous chromosomes. 2 pairs shown below. Rows are SNPs.

- "Unphased genotype" (/): SNP's genotype is NOT ordered with respect to another SNP's genotype
- "Phased genotype" (|): SNP's genotype IS ordered with respect to another SNP's genotype
- Alleles on the same side of | are on the same chromosome, but not necessarily for /

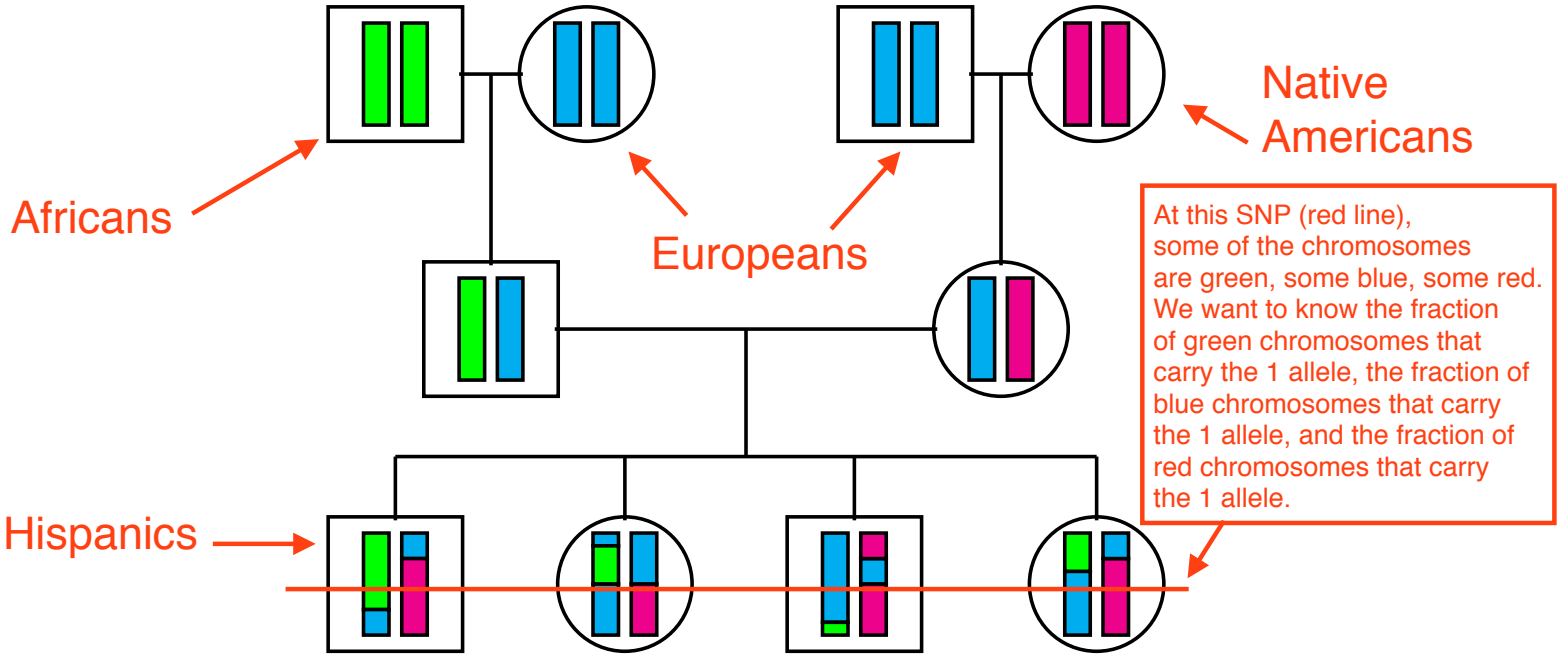


Hispanic Community Health Study (HCHS)

- Cohort study of $13,000$ US Hispanics
- Hispanics are admixed, descended from multiple ancestral populations: Africans, Europeans, and Native Americans
- In this cohort, we test each bi-allelic SNP for association with a trait, say diabetes
- If a SNP is significantly associated with a trait, we want to perform a follow-up study on new individuals to see if we can replicate the association

The Problem that ASAFE Solves

- For a significant SNP, 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



- ASAFE := EM algorithm for estimating these frequencies, for a SNP

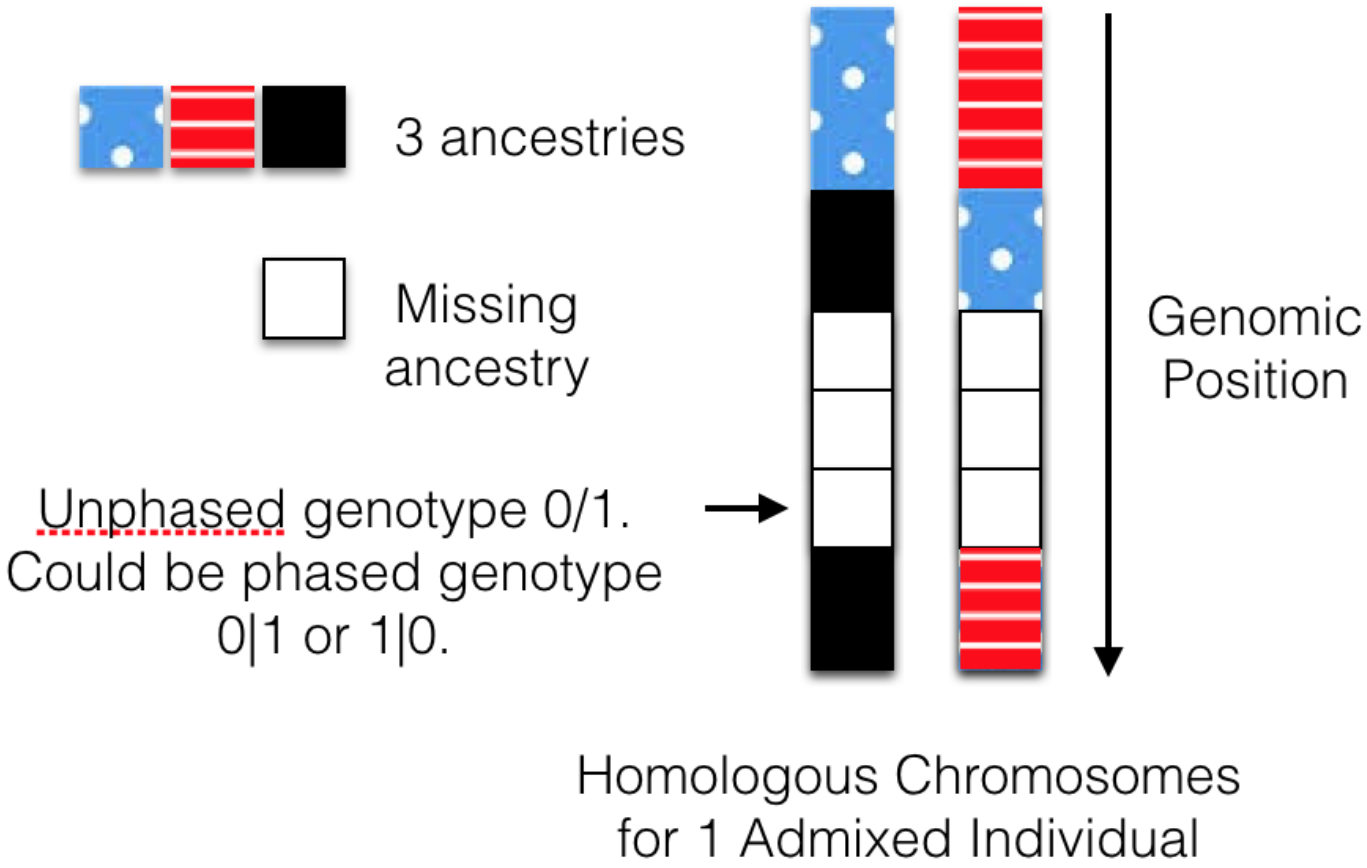
How Ancestry Specific Allele Frequencies Relate to HCHS

- These frequencies inform the design of a replication study: If allele 1 were more common amongst the African chromosomes than amongst the other two ancestries' chromosomes, then one would want to recruit a population of predominantly African descent for the replication study

Available Data

- At some SNPs, the RFMix program takes phased genotypes as input, and outputs admixed individuals' phased ancestries
- Available data on admixed individuals
  - Phased ancestries, Phased genotypes: Some SNPs
  - No ancestry calls, Unphased genotypes: Other SNPs

Available Data

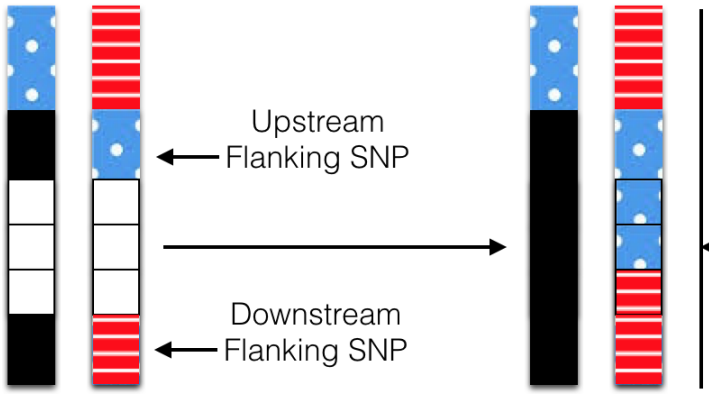


Proposed Approach: Fill in Ancestries

Consider a block of SNPs that have been genotyped in the admixed sample, but that do not have ancestries called. For any SNP in this block:

- Call the SNP's ancestry the nearest flanking ancestry

Then we know all SNPs' unphased genotypes and local ancestry pairs.



Each SNP has unphased genotype 0/1. Could be phased genotype 0|1 or 1|0.

The phased ancestry pair at this SNP is black | blue, and the unphased genotype 0/1. The 1 could go with black and the 0 with the blue, or the 1 could go with the blue and the 0 with the black. This is fundamentally the issue that ASAFE deals with.

Proposed Approach: EM Algorithm to Deal with Unknown Phase of Genotype Relative to Ancestry Pair (ASAFE)

You can view this 2-vector as a new kind of allele, and a (g,a)/(g,a) genotype as a new kind of genotype.

Consider an ancestry-specific allele (allele, ancestry) = (g,a):

- Allele g = 0 or 1
- Ancestry a = African (A), European (E), Native American (N)

There are 6 possibles (g,a) alleles, so 21 values for unordered (g,a)/(g,a) genotype. We call these values unordered (g,a)-genotype categories.

1 complete observation = The (g,a)-genotype category that an individual belongs to at a SNP.

Complete, Unobserved Data Categories

Entry C_i is the name of the i-th complete, unobserved category.

Table: Complete Data Unordered (g,a)-genotype Categories.

(g,a)\(g',a')	(0, A)	(0, E)	(0,N)	(1, A)	(1, E)	(1, N)
(0, A)	C_1					
(0, E)	C_2	C_4				
(0, N)	C_3	C_5	C_6			
(1, A)	C_7	C_9	C_{11}	C_{16}		
(1, E)	C_8	C_{12}	C_{14}	C_{17}	C_{19}	
(1, N)	C_{10}	C_{13}	C_{15}	C_{18}	C_{20}	C_{21}

If an individual is in category C9, that means the individual has (g,a) alleles (1, A) and (0,E).

Hardy-Weinberg Equilibrium assumed in the admixed population to get probability p_j of an individual falling into the j-th complete data category:

The (g,a) alleles combine independently

$$p_j = \begin{cases} p_{ga}p_{g'a'}, & \text{if } (g,a) = (g',a') \\ 2p_{ga}p_{g'a'}, & \text{otherwise} \end{cases}$$

Incomplete, Observed Data Categories

Entry O_i is the name of the i-th incomplete, observed category. Colored entries are observed data categories that map to multiple complete data categories.

Table: Incomplete, Observed Data Categories.

(g,a)\(g',a')	(0, A)	(0, E)	(0,N)	(1, A)	(1, E)	(1, N)
(0, A)	O_1					
(0, E)	O_2	O_4				
(0, N)	O_3	O_5	O_6			
(1, A)	O_7	O_8	O_9	O_{13}		
(1, E)	O_8	O_{10}	O_{11}	O_{14}	O_{16}	
(1, N)	O_9	O_{11}	O_{12}	O_{15}	O_{17}	O_{18}

An individual in observed category O8 has phased ancestry AIE and unphased genotype 0/1. We don't know if the 1 goes with the A or with the E, so two complete categories correspond to this one observed category.

Overlaying complete and observed categories gives their correspondence. This correspondence allows us to express the probability p'_j of an individual being in observed data category $j', j' \in \{1, ..., 18\}$ in terms of complete data category probabilities $p_j, j \in \{1, ..., 21\}$.

Outline Approach to Estimating Ancestry-Specific Allele Frequencies

Because of the connection

- Between p'_j and p_j, and
- Between p_j and (g,a)-allele probabilities $\vec{p} = [p_{ga} : g \in \{0, 1\}, a \in \{A, E, N\}]$,

maximizing the observed data log likelihood (e.g. via EM algorithm ASAFE)

$$log(P(\vec{o} = [o_1, ..., o_n])|\vec{p'} = [p'_1, ..., p'_{18}]) = \sum_{j'=1}^{18} m'_{j'} log(p'_{j'})$$

where o_i = Observed category of the i-th individual, and $m'_{j'}$ = Number of individuals in observed category j'

gives us a maximum likelihood estimate (MLE) $\hat{\vec{p}} = [\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{0N}, \hat{p}_{1A}, \hat{p}_{1E}, \hat{p}_{1N}]$ of $\vec{p}$, from which we obtain ancestry-specific allele frequency estimates:
$\hat{p}_{1|a} = \hat{p}_{1a}/\hat{p}_a = \hat{p}_{1a}/(\hat{p}_{1a} + \hat{p}_{0a}), a \in \{A, E, N\} \leftarrow$ THE GOAL!

Simulated Genetic Data

- Used MaCS [Chen et al. (2009)] to simulate Hispanic individuals' sequence data
- For each of the $56,003$ SNPs in the sequence data, ran ASAFE with inputs: Unphased admixed genotypes (ignoring known phase) and phased admixed ancestries
- Got ancestry-specific allele 1 frequencies for each ancestry (African, European, Native American), at each SNP
- For each SNP, calculated
  error $=$ Estimated $p_{1|a}$ - True $p_{1|a}$, $a \in \{A, E, N\}$

1. Motivation for ASAFE
2. Available Data
3. Proposed Approach
4. Data Simulation
5. Results

Low Error on Simulated Data

Mean and SD of errors $\{\hat{p}_{1|a} - p_{1|a} : a \in \{A, E, N\}\}$, grouped by:

- True allele frequency bin that $p_{1|a}$ falls into: Columns
- Ancestry $a \in \{A, E, N\}$: Rows

How to interpret this cell: Say 10,000 of our 56,003 SNPs have P(1 I African) in (0-0.2]. These 10,000 SNPs have 10,000 associated errors. The mean of those 10,000 errors is -0.0011. Their SD is 0.0065.

		True Allele 1 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

Regardless of true ancestry-specific allele frequency $p_{1|a}$ bin, errors are low: Largest $|\text{Mean}| = 0.005$. Largest SD $= 0.03$.

More Results in paper supplement.

More Info: Paper and Code

- Qian S. Zhang, Brian L. Browning, and Sharon R. Browning. Asafe: ancestry-specific allele frequency estimation. Bioinformatics, 32(14):2227 2229, 2016.
- Package "ASAFE" on Bioconductor

Oops this is wrong! It's been accepted by Bioconductor, but isn't on there as of 7/12/2016.

- Code to reproduce analysis at http://biostatqian.github.io/ASAFE/