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

1. Motivation for ASAFE

2. Available Data

3. Proposed Approach

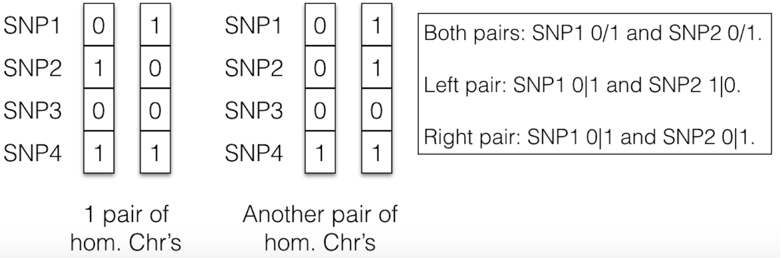
4. Data Simulation

5. Results

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.



SNP1	0	1	SNP1	0	1
SNP2	1	0	SNP2	0	1
SNP3	0	0	SNP3	0	0
SNP4	1	1	SNP4	1	1

1 pair of hom. Chr's

Another pair of hom. Chr's

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

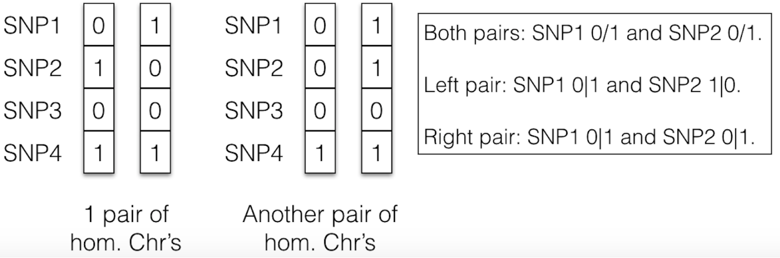
Left pair: SNP1 0|1 and SNP2 1|0.

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.

- "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.



SNP1	0	1
SNP2	1	0
SNP3	0	0
SNP4	1	1

1 pair of
hom. Chr's

SNP1	0	1
SNP2	0	1
SNP3	0	0
SNP4	1	1

Another pair of
hom. Chr's

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

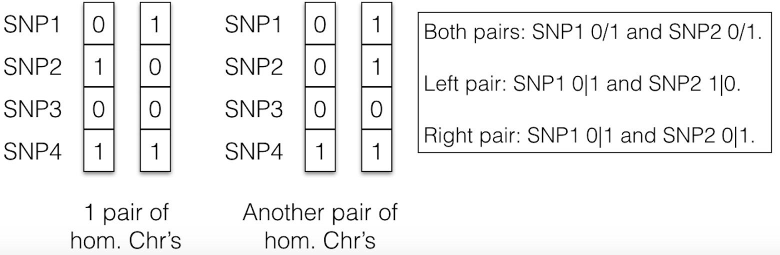
Left pair: SNP1 0|1 and SNP2 1|0.

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



SNP1	0	1
SNP2	1	0
SNP3	0	0
SNP4	1	1

1 pair of hom. Chr's

SNP1	0	1
SNP2	0	1
SNP3	0	0
SNP4	1	1

Another pair of hom. Chr's

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

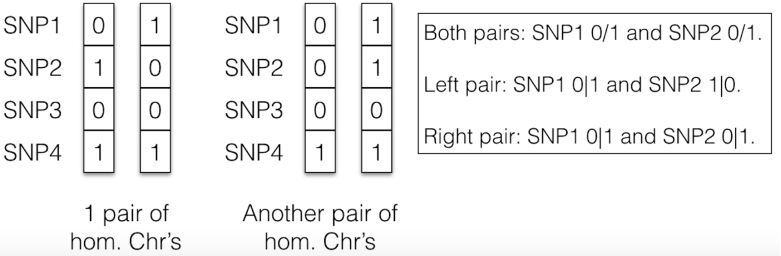
Left pair: SNP1 0|1 and SNP2 1|0.

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
- "Phased genotype" (|): SNP's genotype IS ordered with respect to another SNP's genotype



SNP1	0	1
SNP2	1	0
SNP3	0	0
SNP4	1	1

1 pair of hom. Chr's

SNP1	0	1
SNP2	0	1
SNP3	0	0
SNP4	1	1

Another pair of hom. Chr's

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

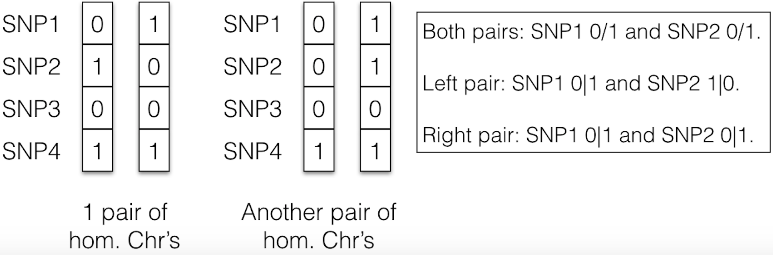
Left pair: SNP1 0|1 and SNP2 1|0.

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
- "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 /



SNP1	0	1
SNP2	1	0
SNP3	0	0
SNP4	1	1

1 pair of hom. Chr's

SNP1	0	1
SNP2	0	1
SNP3	0	0
SNP4	1	1

Another pair of hom. Chr's

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

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

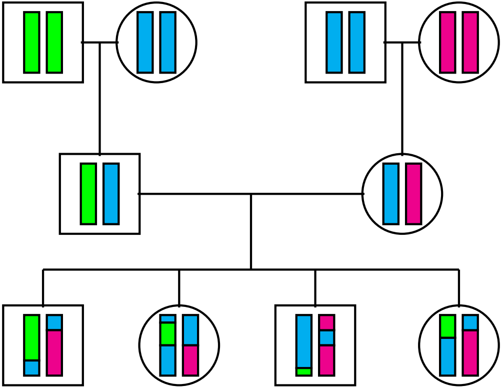
Right pair: SNP1 0|1 and SNP2 0|1.

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

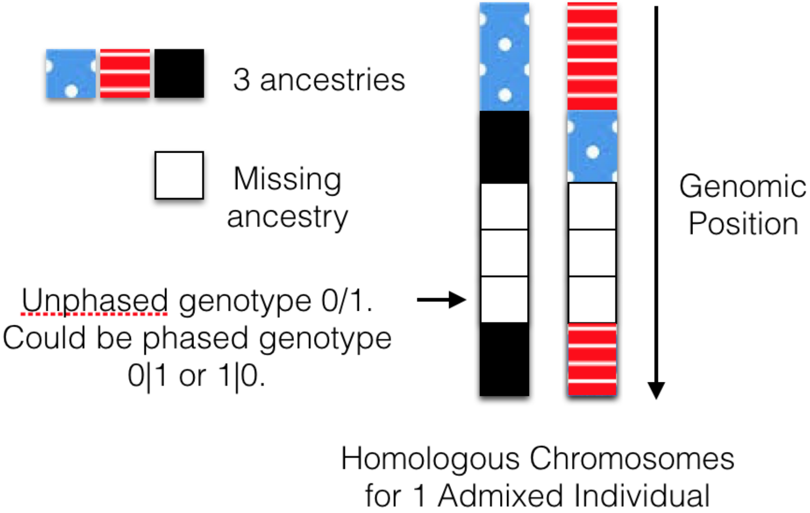
- These frequencies inform the design of a replication study: If allele 1 were more common amongst the African chromosomes than amongst the other two ancestries' chromosomes, then one would want to recruit a population of predominantly African descent for the replication study

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

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



3 ancestries

Missing ancestry

Unphased genotype 0/1. Could be phased genotype 0|1 or 1|0.

Genomic Position

Homologous Chromosomes for 1 Admixed Individual

1. Motivation for ASAFE

2. Available Data

3. Proposed Approach

4. Data Simulation

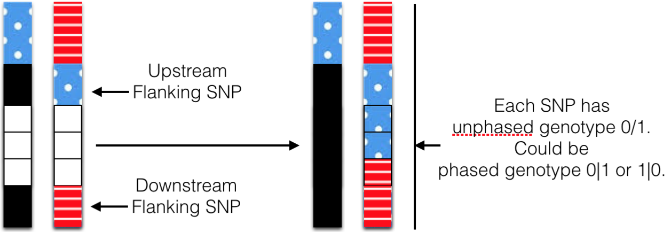
5. Results

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.



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

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}

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

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

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}

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:
$\hat{p}_{1|a} = \hat{p}_{1a}/\hat{p}_a = \hat{p}_{1a}/(\hat{p}_{1a} + \hat{p}_{0a}), a \in \{A, E, N\} \leftarrow \text{THE GOAL!}$

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

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

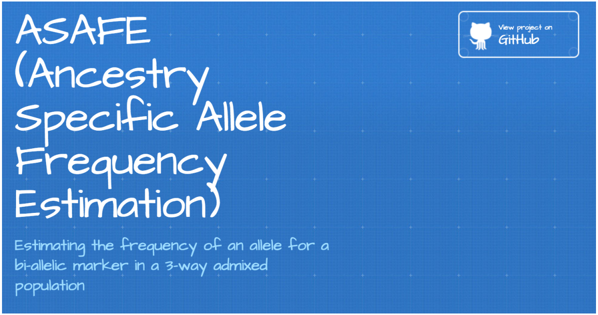
		True Allele 1 Frequency Bins				
Ancestry	Statistic	(0-0.2]	(0.2-0.4]	(0.4-0.6]	(0.6-0.8]	(0.8-1]
African	Mean	-0.0011	-0.0003	-0.0004	0.0004	-0.0004
African	SD	0.0065	0.0185	0.0233	0.0186	0.0118
European	Mean	-0.0015	-0.0004	-0.0007	-0.0010	<0.0001
European	SD	0.0077	0.0209	0.0249	0.0220	0.0122
Nat. 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
- Code to reproduce analysis at http://biostatqian.github.io/ASAFE/



ASAFE (Ancestry Specific Allele Frequency Estimation)

Estimating the frequency of an allele for a bi-allelic marker in a 3-way admixed population

6 Appendix: Extra Slides

Available Data

- Unphased bi-allelic SNP genotypes from three reference panels proxying ancestral Africans, Europeans, and Native Americans. Unphased admixed sample genotypes.
  → Phase all SNP genotypes
  → Program RFMix [Maples et al. (2013)] takes phased SNP genotypes, internally re-phases admixed genotypes, and outputs ancestry calls for each admixed person's chromosomes only at SNPs typed in all 3 reference panels
- Using a program called RFMix, we obtain admixed individuals' phased ancestries at some SNPs
- Available data on admixed individuals
  - Phased ancestries, Unphased genotypes: For SNPs that RFMix did not output re-phased genotypes for
  - No ancestry calls, Unphased genotypes: For SNPs that RFMix could NOT make ancestry calls for (e.g. not typed in a reference panel)

Inadequate Alternative Approaches

Goal: Estimate ancestry-specific allele frequencies, for each marker typed in the admixed sample, with available data.

Approaches:

- (1) Find the allele 1 frequency in each reference panel
- (2) Find allele 1 frequencies in populations sequenced by the 1000 Genomes project [Via García et al. (2012)]
- (3) ADMIXTURE, STRUCTURE

Weaknesses of Approaches:

- (1), (2), (3): If a marker is not typed in a reference panel or sequenced by the 1000 Genomes project, these approaches cannot be taken
- (3) assume linkage equilibrium (independence amongst SNPs), e.g. do not make use of linkage disequilibrium from local ancestry calls

ASAFE EM Algorithm

We approximate the MLE $\hat{\vec{p}}$ using an EM algorithm:

(0) Start with an initial estimate of

$\vec{p} = \vec{p}_0 = [p_{0,ga}, g \in \{0,1\}, a \in \{A,E,N\}] = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6]$

Iterate E-M steps (1 E-M = 1 iteration) until $||\vec{p}_{k+1} - \vec{p}_k||_2 < \epsilon = 10^{-8}$, where $\vec{p}_{k+1}$ is the latest estimate of $\vec{p}$ and $\vec{p}_k$ is the 2nd to latest estimate.

(1) E-Step : Evaluate

$E_{\vec{c}}(log(P(\vec{c}|\vec{p}))|\vec{o}, \vec{p}_k) = E_{\vec{c}}(\sum_{j=1}^{21}[m_j log(p_j)]|\vec{o}, \vec{p}_k) =$
$\sum_{j=1}^{21}[E_{\vec{c}}(m_j|\vec{p}, \vec{p}_k) log(p_j)]$

(2) M-Step: Set

$\vec{p}_{k+1} = \mathrm{argmax}_{\vec{p}} E_{\vec{c}}(log(P(\vec{c}|\vec{p})|\vec{o}, \vec{p}_k))$

E-Step in More Detail

On the k-th iteration of the algorithm, let the expected value of the number m_j of individuals in complete category j be denoted

$$m_{k,j} = E_{\vec{c}}(m_j | \vec{o}, \vec{p}_k) = \sum_{i=1}^{m'_{j'}} \frac{P(c_i = j | \vec{p}_k)}{P(c_i = \text{Any j that is consistent with } o_i = j' | \vec{p}_k)}$$

where

- $\vec{c} = [c_1, ..., c_n]$ are complete categories for n admixed individuals
- $\vec{o} = [o_1, ..., o_n]$ are observed categories for n admixed individuals
- $\vec{p}_k = [p_{k,0A}, p_{k,0E}, p_{k,0N}, p_{k,1A}, p_{k,1E}, p_{k,1N}]$ is the k-th estimate for the $\vec{p}$ that maximizes the observed data log likelihood
- $m'_{j'}$ is the number of individuals in observed category j' that is consistent with complete category j

M-Step in More Detail

$\hat{\vec{p}}_{k+1} = [\hat{p}_{k+1,0A}, \hat{p}_{k+1,0E}, \hat{p}_{k+1,0N}, \hat{p}_{k+1,1A}, \hat{p}_{k+1,1E}, \hat{p}_{k+1,1N}]$, where

$$\hat{p}_{k+1,0A} = \frac{2m_{k,1} + m_{k,2} + m_{k,3} + m_{k,7} + m_{k,8} + m_{k,10}}{2n}$$
$$\hat{p}_{k+1,0E} = \frac{m_{k,2} + 2m_{k,4} + m_{k,5} + m_{k,9} + m_{k,12} + m_{k,13}}{2n}$$
$$\hat{p}_{k+1,0N} = \frac{m_{k,3} + m_{k,5} + 2m_{k,6} + m_{k,11} + m_{k,14} + m_{k,15}}{2n}$$
$$\hat{p}_{k+1,1A} = \frac{m_{k,7} + m_{k,9} + m_{k,11} + 2m_{k,16} + m_{k,17} + m_{k,18}}{2n}$$
$$\hat{p}_{k+1,1E} = \frac{m_{k,8} + m_{k,12} + m_{k,14} + m_{k,17} + 2m_{k,19} + m_{k,20}}{2n}$$
$$\hat{p}_{k+1,1N} = \frac{m_{k,10} + m_{k,13} + m_{k,15} + m_{k,18} + m_{k,20} + 2m_{k,21}}{2n}$$

where $n =$ Number of individuals.

Chen, G. K., Marjoram, P., and Wall, J. D. (2009). Fast and flexible simulation of dna sequence data. *Genome research*, 19(1):136–142.

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. *The American Journal of Human Genetics*, 93(2):278–288.

Via García, M., Consortium, . G. P., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 2012, vol. 491, p. 56-65*.