1 Introduction

Population Structure is defined by the genetic differences between individuals due to ancestry differences. Methods used to learn population structure include

- use existing labels (self-reported ancestry)
- learn from genetic data

1.1 Admixture Models

Given $k$ populations: each individual has ancestries proportions from each of these $k$ populations

$$x_i \rightarrow q_i(....)$$

$q_i$ is a $k$-vector


note that mixture model assigns each individual to only one population; admixture model allows multi-population assignment
1.2 PCA

Each individual genotype can be represented as a linear combination of 'eigen' genotypes (Principal Components). The assumption is that PCs represent ancestry.

\[ x_i \rightarrow z_i \text{ mapping function} \]
\[ x_i \in \mathbb{R}^m \]
\[ z_i \in \mathbb{R}^l \]

1.3 Trees (Hierarchical Clustering)

2 Genetic Ancestry Inference

2.1 Supervised Genetic Ancestry Inference

Example 1. POP1 and POP2 with known allele frequencies

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
<th>SNP3</th>
<th>SNP4</th>
<th>....</th>
</tr>
</thead>
<tbody>
<tr>
<td>POP1</td>
<td>0.25</td>
<td>0.57</td>
<td>0.29</td>
<td>0.38</td>
</tr>
<tr>
<td>POP2</td>
<td>0.40</td>
<td>0.32</td>
<td>0.84</td>
<td>0.22</td>
</tr>
<tr>
<td>Individual X</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Does individual x belong to POP1 or POP2?

Ans:

P(Data | x is in POP1) is proportional to
\[ (0.25)^2(0.75)^0(0.57)^0(0.43)^2(0.29)^1(0.71)^1(0.38)^1(0.62)^1 = 0.0006 \]

P(Data | x is in POP2) is proportional to
\[ (0.40)^2(0.60)^0(0.32)^0(0.68)^2(0.84)^1(0.16)^1(0.22)^1(0.78)^1 = 0.0017 \]

Each SNP is considered independently.

We can extend the model to the admixture setting

If individual x has ancestry \( \alpha \) from POP1 and \((1 - \alpha)\) from POP2, then what is the most likely value of \( \alpha \) ?

Ans:

P(Data | \alpha) is proportional to
\[ [0.25\alpha + 0.40(1-\alpha)]^2[0.75\alpha + 0.60(1-\alpha)]^0 \\
[0.57\alpha + 0.32(1-\alpha)]^0[0.43\alpha + 0.68(1-\alpha)]^2 \\
[0.29\alpha + 0.84(1-\alpha)]^1[0.71\alpha + 0.16(1-\alpha)]^1 \\
[0.38\alpha + 0.22(1-\alpha)]^1[0.62\alpha + 0.78(1-\alpha)]^1 \]

max. value 0.0020 attained at \( \alpha = 0.22 \)

General case: \( M \) SNPs (\( m = 1 \) to \( M \)). \( N \) populations (\( n = 1 \) to \( N \)), known allele frequency \( p_{mn} \) for SNP \( m \) in population \( n \), observed genotype counts \( g_{mn} \) for SNP \( m \) in individual x.

If individual x has fractional ancestry \( \alpha_n \) from each population \( n \), subject to \( \sum_n \alpha_n = 1 \), then what are the most likely values of \( \alpha_n \)?
\[ P(DATA \mid x \sim \alpha_1, \ldots, \alpha_N) \text{ is proportional to} \]
\[
\prod_{m=1}^{M} \left( \sum_{n=1}^{N} \alpha_n p_{nm} \right)^{g_m} \left( \sum_{n=1}^{N} \alpha_n (1 - p_{nm}) \right)^{2-g_m} \tag{1}
\]
Answer: find the values of \( \alpha_n \) which maximize this expression.

### 2.2 Unsupervised Genetic Ancestry Inference

General case: \( M \) SNPs \((m = 1 \text{ to } M)\), \( N \) populations \((n = 1 \text{ to } N)\), unknown allele frequency \( p_{nm} \) for SNP \( m \) in population \( n \), observed genotype counts \( g_m \) for SNP \( m \) in many individual \( x_i \).

If individual \( x_i \) has fractional ancestry \( \alpha_{in} \) from each population \( n \), subject to \( \sum_{n} \alpha_{in} = 1 \), then what are the most likely values of \( \alpha_{in} \)?

\[ P(DATA \mid x_i \sim \alpha_{i1}, \ldots, \alpha_{iN}) \text{ is proportional to} \]
\[
\prod_{i=1}^{I} \prod_{m=1}^{M} \left( \sum_{n=1}^{N} \alpha_{in} p_{nm} \right)^{g_{im}} \left( \sum_{n=1}^{N} \alpha_{in} (1 - p_{nm}) \right)^{2-g_{im}} \tag{2}
\]
Answer: find values of \( \alpha_{in}, p_{nm} \) which maximize this expression.

Use EM to estimate parameters

### 2.2.1 Human Genome Diversity Project (HGDP)

Structure results on HGDB:

- 1056 individuals from 52 populations
- 377 microsatellite markers (multi-allelic)

EM algorithm (Frappe) on HGDP:

- 938 HGDP individual (118 related individual removed)
- 51 world populations
- illumina 650k chip

### 2.3 PCA

10 points in 1000000-dimensional space

![Decomposition](image)

Example
\[
X = \begin{bmatrix}
2 & -2 \\
1 & -1 \\
0 & 0 \\
-1 & 1 \\
-2 & 2
\end{bmatrix}
\]

\[
\Psi = VDV^T \text{(decomposing covariance matrix)}
\]

\[
\begin{bmatrix}
10 & -10 \\
-10 & 10
\end{bmatrix} = \begin{bmatrix}
1/\sqrt{2} & 1/\sqrt{2} \\
-1/\sqrt{2} & 1/\sqrt{2}
\end{bmatrix} \begin{bmatrix}
20 & 0 \\
0 & 0
\end{bmatrix} \begin{bmatrix}
1/\sqrt{2} & -1/\sqrt{2} \\
-1/\sqrt{2} & 1/\sqrt{2}
\end{bmatrix}
\]

\[
PC1 = \begin{bmatrix}
1/\sqrt{2} \\
-1/\sqrt{2}
\end{bmatrix}; \text{ Eigenvalue 1 = 20}
\]

\[
\Psi_{V1} = d_1v_1 = \begin{bmatrix}
20/\sqrt{2} \\
-20/\sqrt{2}
\end{bmatrix}
\]

2.4 PCA on genotype data

\(G = M \times N\) matrix of individual genotypes
M SNPs, N individuals
\(g_{ij}\) = genotype (0, 1, or 2 alleles) of SNP i in individual j

Subtract off the mean of SNP i, \(p_i = \text{Avg}_{j}(g_{ij}/2)\), set \(g_{ij} = g_{ij} - 2p_i\)
(Missing data: set \(g_{ij} = 0\) if SNP i in individual j is missing data)
Optional: normalize by \(\sqrt{2p_i(1-p_i)}\), i.e. set \(g_{ij} = g_{ij}/\sqrt{2p_i(1-p_i)}\)

\(\Psi = N \times N\) covariance matrix of G
\(\Psi = VDV^T\) (Eigen-decomposition)
Columns of V are eigenvectors (principal components, PCs) of G.
Diagonal entries of D are eigenvalues of G.
The hope: Top PCs (PC1, PC2) correspond to genetic ancestry.
In summary:
genotype is discrete
row data $\rightarrow$ normalized data $\rightarrow$ covariance $\rightarrow$ PCA $\rightarrow$ axes

Results:
The first PC separates African from the rest of the population; the second PC separates European and east Asian.

Other Methods:
- Multi-Dimensional Scale (MDS)
- Kernal PCA
- Isomap SNE (non-linear reduction)
- Random vs Associative

3 Population Admixture

Admixed population is one that has ancestry from multiple distinct populations.
What is a population? Fine structure: admixture is common in human history

3.1 Admixture and Ancestry

Driving force mostly comes from recombination over time
$x_i$: individual
$z_i$: local ancestry
$q_i$: genome-wide ancestry note: trees cannot describe admixture

3.2 Population Structure and GWAS

population structure can lead to false discoveries

3.2.1 Detecting Stratification

Assumption: Most SNPs are not associated with the phenotype.
Distribution of P-values close to uniform
Can be visualized using a Q-Q plot
P-values can look non-uniform for model mis-specification
For example: Batch effects, Family relationships

3.2.2 Approaches to Deal with Stratification

- Cluster individuals into populations and do GWAS in each population

$$Z_i \in [\mathbb{R}]$$

$$m \text{ test} \rightarrow \frac{\alpha}{m}$$
Issue: clustering may not give good result due to admixture
Potential Solution: do local clustering, use k-separated groups:

\[
\begin{align*}
y_1 &= \alpha + X\beta + \epsilon_1 + Z_1 \\
y_2 &= \alpha + X\beta + \epsilon_2 + Z_2 \\
&\vdots \\
y_k &= \alpha + X\beta + \epsilon_k + Z_k
\end{align*}
\]

\(Z_i\) are the indicators

- Include principal components in regression

probabilistic graphical model

Confounding:
Affects both genotype (x) and phenotype (y)