Population sub-structure
Consider an association test in a recently admixed population

Suppose location A increases susceptibility to a disease common in Africans.

Let locus B associate with skin-color.

As expected, locus A is associated with the disease

Unexpectedly, locus B also shows association

The problem arises because the assumption of a random mating population is no longer true

- The population has structure!
Population sub-structure can increase LD

- Consider two populations that were isolated and evolving independently.
- They might have different allele frequencies in some regions.
- Pick two regions that are far apart (LD is very low, close to 0)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- Pop. A: \( p_1 = 0.1 \), \( q_1 = 0.9 \), \( P_{11} = 0.1 \), \( D = 0.01 \)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Pop. B: \( p_1 = 0.9 \), \( q_1 = 0.1 \), \( P_{11} = 0.1 \), \( D = 0.01 \)
Recent ad-mixing of population

- If the populations came together recently (Ex: African and European population), artificial LD might be created.
- $D = 0.15$ (instead of 0.01), increases 10-fold
- This spurious LD might lead to false associations
- Other genetic events can cause LD to increase, and one needs to be careful

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>$q_1$</td>
<td>$P_{11}$</td>
<td>$D$</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.1</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

$D = 0.1 - 0.25 = 0.15$
Determining population sub-structure

- Given a mix of people, can you sub-divide them into ethnic populations.
- Turn the ‘problem’ of spurious LD into a clue.
  - Find markers that are too far apart to show LD
  - If they do show LD (correlation), that shows the existence of multiple populations.
  - Sub-divide them into populations so that LD disappears.
Determining Population sub-structure

- Same example as before:
- The two markers are too similar to show any LD, yet they do show LD.
- However, if you split them so that all 0..1 are in one population and all 1..0 are in another, LD disappears
Iterative algorithm for population sub-structure

- Define
- \( N = \) number of individuals (each has a single chromosome)
- \( k = \) number of sub-populations.
- \( Z \in \{1..k\}^N \) is a vector giving the sub-population.
  - \( Z_i = k' \Rightarrow \) individual \( i \) is assigned to population \( k' \)
- \( X_{i,j} = \) allelic value for individual \( i \) in position \( j \)
- \( P_{k,j,l} = \) frequency of allele \( l \) at position \( j \) in population \( k \)
Example

- Ex: consider the following assignment
- $P_{1,1,0} = 0.9$
- $P_{2,1,0} = 0.1$
Recall

- Define:
  - \(N\) = number of individuals (each has a single chromosome)
  - \(k\) = number of sub-populations.
  - \(Z \in \{1..k\}^N\) is a vector giving the sub-population.
    - \(Z_i = k'\) => individual \(i\) is assigned to population \(k'\)
  - \(X_{i,j}\) = allelic value for individual \(i\) in position \(j\)
  - \(P_{k,j,l}\) = frequency of allele \(l\) at position \(j\) in population \(k\)
Example

- Ex: consider the following assignment
  - $P_{1,1,0} = 0.9$
  - $P_{2,1,0} = 0.1$
Algorithm: Structure

- Iteratively estimate
  - \((Z^{(0)}, P^{(0)}), (Z^{(1)}, P^{(1)}), \ldots, (Z^{(m)}, P^{(m)})\)
- After ‘convergence’, \(Z^{(m)}\) is the answer.
- Iteration
  - Guess \(Z^{(0)}\)
  - For \(m = 1, 2, \ldots\)
    - Sample \(P^{(m)}\) from \(\Pr(P | X, Z^{(m-1)})\)
    - Sample \(Z^{(m)}\) from \(\Pr(Z | X, P^{(m)})\)
- How is this sampling done?
Choose $Z$ at random, so each individual is assigned to be in one of 2 populations. See example.

Now, we need to sample $P^{(1)}$ from $\Pr(P \mid X, Z^{(0)})$

Simply count

$N_{k,j,l} =$ number of people in population $k$ which have allele $l$ in position $j$

$p_{k,j,l} = N_{k,j,l} / N$

1 0 .. 1
2 0 .. 1
2 0 .. 0
1 1 .. 1
1 0 .. 1
2 0 .. 1
1 0 .. 1
2 0 .. 1
1 0 .. 1
Example

- $N_{k,j,l} = \text{number of people in population } k \text{ which have allele } l \text{ in position } j$
- $p_{k,j,l} = N_{k,j,l} / N_{k,j,*}$
- $N_{1,1,0} = 4$
- $N_{1,1,1} = 6$
- $p_{1,1,0} = 4/10$
- $p_{1,2,0} = 4/10$
- Thus, we can sample $P^{(m)}$

$$p_{kl}|X, Z \sim \mathcal{D}(\lambda_1 + n_{kl1}, \ldots, \lambda_{j_1} + n_{klj_1}),$$
where

$$n_{klj} = \#\{(i,a) : x_i^{(l,a)} = j \text{ and } z^{(l)} = k\}$$
Sampling Z

- Pr[Z₁ = 1] = Pr["01" belongs to population 1]?
- We know that each position should be in linkage equilibrium and independent.
- Pr["01" | Population 1] = p₁₁₀ * p₁₂₁
  =(4/10)*(6/10)=(0.24)
- Pr["01" | Population 2] = p₂₁₀ * p₂₂₁ = (6/10)*(4/10)=0.24
- Pr [Z₁ = 1] = 0.24/(0.24+0.24) = 0.5

\[ \Pr(z^{(l)} = k|X, P) = \frac{\Pr(x^{(l)}|P, z^{(l)} = k)}{\sum_{k'}^{K}\Pr(x^{(l)}|P, z^{(l)} = k')}, \quad (A8) \]

where \( \Pr(x^{(l)}|P, z^{(l)} = k) = \prod_{i=1}^{L} p_{klx(i,1)}p_{klx(i,2)}. \)

Assuming, HWE, and LE
Suppose, during the iteration, there is a bias.

Then, in the next step of sampling $Z$, we will do the right thing.

$\Pr[\text{“01”}|\text{pop. 1}] = p_{1,1,0} * p_{1,2,1} = 0.7*0.7 = 0.49$

$\Pr[\text{“01”}|\text{pop. 2}] = p_{2,1,0} * p_{2,2,1} = 0.3*0.3 = 0.09$

$\Pr[Z_1 = 1] = 0.49/(0.49+0.09) = 0.85$

$\Pr[Z_6 = 1] = 0.49/(0.49+0.09) = 0.85$

Eventually all “01” will become 1 population, and all “10” will become a second population.
Allowing for admixture

- Define $q_{i,k}$ as the fraction of individual $i$ that originated from population $k$.

- Iteration
  - Guess $Z^{(0)}$
  - For $m = 1, 2, ..$
    - Sample $P^{(m)}, Q^{(m)}$ from $Pr(P, Q | X, Z^{(m-1)})$
    - Sample $Z^{(m)}$ from $Pr(Z | X, P^{(m)}, Q^{(m)})$
Estimating $Z$ (admixture case)

- Instead of estimating $Pr(Z(i)=k|X,P,Q)$, (origin of individual $i$ is $k$), we estimate $Pr(Z(i,j,l)=k|X,P,Q)$

$$Pr(Z_{i,j,l} = k \mid X,P,Q) = \frac{q_{i,k} \Pr(X_{i,j,l} \mid Z_{i,j,l} = k,P)}{\sum_{k'} q_{i,k'} \Pr(X_{i,j,l} \mid Z_{i,j,l} = k',P)}$$
Results: Thrush data

- For each individual, $q(i)$ is plotted as the distance to the opposite side of the triangle.
- The assignment is reliable, and there is evidence of admixture.
NJ versus Structure: thrush data

- Objective function is different in standard clustering algorithms!
Population Structure in isolated human populations

- 377 locations (loci) were sampled in 1000 people from 52 populations.
- 6 genetic clusters were obtained, which corresponded to 5 geographic regions (Rosenberg et al. Science 2003)
Population sub-structure: research problem

- Systematically explore the effect of admixture. Can admixture be predicted for a locus, or for an individual?
- The sampling approach may or may not be appropriate. Formulate as an optimization/learning problem:
  - (w/out admixture). Assign individuals to sub-populations so as to maximize linkage equilibrium, and hardy weinberg equilibrium in each of the sub-populations
  - (w/ admixture) Assign (individuals, loci) to sub-populations
The Transposed SNP Matrix

Individuals (N)

allele counts

SNPs (M)
Dimensionality reduction

Individuals (N)

SNPs (M)

allele counts

Individuals (N)

k

dimensionality reduction
We get the intrinsic dimensionality of a dataset.
Principle Components Analysis

- Consider the genotype values of 2 SNPs over 6 individuals.
- Clearly, the two SNPs are highly correlated.
- Projecting all the individuals on a single line could explain most of the data.
Reducing to a single dimension

- Goal: Represent each individual (a point in high dimensional space) by a single value.
- If the intrinsic dimensionality is 1, it is possible that the single dimension is sufficient to ‘explain the sample space’
Consider the mean of all points \( m \), and a vector emanating from the mean.

Algebraically, this projection on \( \beta \) means that all samples \( x \) can be represented by a single value \( \beta^T(x-m) \).
Higher dimensions

- Consider a set of 2 (k) orthonormal vectors $\beta_1, \beta_2...$
- Once projected, each sample means that all samples $x$ can be represented by 2 (k) dimensional vector
  - $\beta_1^T(x-m), \beta_2^T(x-m)$
How to project

- The generic scheme allows us to project an m dimensional surface into a k dimensional one.
- How do we select the k ‘best’ dimensions?
- The strategy used by PCA is one that maximizes the variance of the projected points around the mean
PCA

- Suppose all of the data were to be reduced by projecting to a single line $\beta$ from the mean.
- How do we select the line $\beta$?
Let each point $x_k$ map to $x_k' = m + a_k \beta$. We want to minimize the error

$$\sum_k \left\| x_k - x_k' \right\|^2$$

Claim 1: Each point $x_k$ maps to $x_k' = m + \beta^T(x_k - m)\beta$

- $a_k = \beta^T(x_k - m)$
Proof of Observation 1

\[
\min_{a_k} \|x_k - x'_k\|^2 = \min_{a_k} \|x_k - m + m - x'_k\|^2 \\
= \min_{a_k} \|x_k - m\|^2 + \|m - x'_k\|^2 - 2(x'_k - m)^T (x_k - m) \\
= \min_{a_k} \|x_k - m\|^2 + a_k^2 \beta^T \beta - 2a_k \beta^T (x_k - m) \\
= \min_{a_k} \|x_k - m\|^2 + a_k^2 - 2a_k \beta^T (x_k - m)
\]

Differentiating w.r.t \( a_k \)

\[
2a_k - 2\beta^T (x_k - m) = 0 \\
a_k = \beta^T (x_k - m) \\
\Rightarrow a_k^2 = a_k \beta^T (x_k - m) \\
\Rightarrow \|x_k - x'_k\|^2 = \|x_k - m\|^2 - \beta^T (x_k - m)(x_k - m)^T \beta
\]
Minimizing PCA Error

$$\sum_k \left\| x_k - x'_k \right\|^2$$

$$= C - \sum_k \beta^T (x_k - m)(x_k - m)^T \beta = C - \beta^T S \beta$$

- To minimize error, we must maximize $\beta^T S \beta$
- By definition, $\lambda = \beta^T S \beta$ implies that $\lambda$ is an eigenvalue, and $\beta$ the corresponding eigenvector.
- Therefore, we must choose the eigenvector corresponding to the largest eigenvalue.
PCA steps

1. \( m = \frac{1}{n} \sum_{j=1}^{n} x_j \)

2. \( h^T = [1\cdots 1] \)

3. \( M = X - mh^T \)

4. \( S = MM^T = \sum_{j=1}^{n} (x_j - m)(x_j - m)^T \)

5. \( B^T SB = \Lambda \)

6. Return \( B^T M \)
Population structure within Europe.
We started (and finished) by considering sources of variation

Models of evolution of population under natural assumptions
  1. HW/Linkage equilibrium
  2. Efficient simulation of populations via coalescent theory

Detecting structural variation

Detection of regions under selection

Association testing

Population sub-structure

Evolution under recombinations/recombination hot-spot detection via counting of recombination events (partially)

Haplotype phasing (partially)
1. Phylogeny reconstruction (perfect phylogeny, distance-based methods)
2. Optimization using (Integer-) Linear programming, simulated annealing and other paradigms
3. Greedy algorithms, dynamic programming
4. Stochastic sampling methods (MCMC, Gibbs sampling)
5. Statistical tests for significance
6. Efficient simulation techniques
   1. Coalescent for populations
   2. Simulating genotype/phenotype associations
   3. Pairwise analysis
Final exam date
Population Substructure Conclusions

- Population substructure (violating the assumption of random mating) can create artificial linkage disequilibrium, and false associations.
- One way out of it is to identify and sub-divide the populations into sub-populations.
- Another approach is to ‘correct’ for the effects of structure.