Introduction to Genome-Wide Association Studies (GWAS) for complex disease

Dr. Emile Chimusa

Course materials: http://web.cbio.uct.ac.za/~emile/AGe/Data_Sciences/
Principles of Genome wide Association Studies
1. Genome-wide Association Studies
2. Population Stratification
3. Visualizing GWAS Results
4. Phasing and Imputation
<table>
<thead>
<tr>
<th>Case 1</th>
<th>A/T</th>
<th>C/G</th>
<th>A/G</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA</td>
<td>CC</td>
<td>AA</td>
</tr>
<tr>
<td>Case 2</td>
<td>AT</td>
<td>CG</td>
<td>AA</td>
</tr>
<tr>
<td>Case 3</td>
<td>AA</td>
<td>CG</td>
<td>AA</td>
</tr>
<tr>
<td>Control 1</td>
<td>TT</td>
<td>GG</td>
<td>GG</td>
</tr>
<tr>
<td>Control 2</td>
<td>TT</td>
<td>CC</td>
<td>GG</td>
</tr>
<tr>
<td>Control 3</td>
<td>TA</td>
<td>CG</td>
<td>GG</td>
</tr>
</tbody>
</table>
## Encoded data

<table>
<thead>
<tr>
<th></th>
<th>A/T</th>
<th>C/G</th>
<th>A/G</th>
<th></th>
<th>A/T</th>
<th>C/G</th>
<th>A/G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case1</td>
<td>AA</td>
<td>CC</td>
<td>AA</td>
<td>Case2</td>
<td>AT</td>
<td>CG</td>
<td>AA</td>
</tr>
<tr>
<td>Case2</td>
<td>AT</td>
<td>CG</td>
<td>AA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case3</td>
<td>AA</td>
<td>CG</td>
<td>AA</td>
<td>=&gt;</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Con1</td>
<td>TT</td>
<td>GG</td>
<td>GG</td>
<td>Con2</td>
<td>TT</td>
<td>CC</td>
<td>GG</td>
</tr>
<tr>
<td>Con2</td>
<td>TT</td>
<td>CC</td>
<td>GG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Con3</td>
<td>TA</td>
<td>CG</td>
<td>GG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A good ranking strategy would produce SNP3, SNP1, SNP2
We compute correlation (association statistic) between SNP and a disease

- Association statistic is based on allele freq. difference ($\hat{p}_X^+ - \hat{p}_X^-$)
- The larger the difference, the higher the correlation

GWAS: Further Tests

400 cases have A
600 cases have T

$\hat{p}_X^+ = 0.4$

300 controls have A
700 controls have T

$\hat{p}_X^- = 0.3$
Direct and indirect association testing

Hirschhorn and Daly: Nature Reviews Genetics 6: 95 (2005)

- **Direct association**
  - Functional SNP is genotyped and an association is found

- **Indirect association**
  - Functional SNP (blue) is not genotypred, but a number of other SNPs (red), in LD with the functional SNP, are genotyped, and an association is found for these SNPs
GWAS

\( \hat{p}_A^+ \sim N(p_A^+, p_A^+(1-p_A^+)/N) \)
\( \hat{p}_A^- \sim N(p_A^-, p_A^-(1-p_A^-)/N) \)
\( \hat{p}_A^+ - \hat{p}_A^- \sim N(p_A^+ - p_A^-, (p_A^+(1-p_A^+)+p_A^-(1-p_A^-))/N) \)

We approximate (what about rare variants?)

\[ p_A^+(1-p_A^+)+p_A^-(1-p_A^-) \approx 2 \hat{p}_A(1-\hat{p}_A) \]

then if \( p_A^+ = p_A^- \)
\[ S_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2/N \sqrt{\hat{p}_A(1-\hat{p}_A)}}} \sim N(0,1) \]
Association Statistic

\[ S_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2/N} \sqrt{\hat{p}_A (1 - \hat{p}_A)}} \sim N(0,1) \]

- Under the null hypothesis \( p^+_A - p^-_A = 0 \)
- We compute the statistic \( S_A \).
- If \( S_A < \Phi^{-1}(\alpha/2) \) or \( S_A > -\Phi^{-1}(\alpha/2) \) then the association is significant at level \( \alpha \).
GWAS: Application

Chimusa et al. 2013 Hum Mol Gen
GWAS: Application

Chimusa et al. 2013 Hum Mol Gen
GWAS: Application

Chimusa et al. 2013 Hum Mol Gen

![GWAS Application Diagram]
Genome-wide Association Studies and Population stratification

1. Genome-wide Association Studies
2. Population Stratification
3. Visualizing GWAS Results
3. Phasing and Imputation
Outline

1. Introduction to population stratification

2. Detecting population stratification: $\lambda_{GC}$ and Q-Q plots

3. Correcting for stratification: Genomic Control

4. Correcting for stratification: STRUCTURE and STRAT
Population stratification refers to ancestry differences between cases and controls in a genetic association study.
Detecting stratification: Q-Q plots

Distribution of ranked association P-values for 9 SNPs:

<table>
<thead>
<tr>
<th>Expected:</th>
<th>Actual:</th>
<th>Matches uniform (null) distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>0.40</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.93</td>
<td></td>
</tr>
</tbody>
</table>
Detecting stratification: Q-Q plots

Distribution of ranked association $-\log_{10} P$-values for 9 SNPs:

<table>
<thead>
<tr>
<th>Expected $-\log_{10} P$:</th>
<th>Actual $-\log_{10} P$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.10</td>
</tr>
<tr>
<td>0.70</td>
<td>0.68</td>
</tr>
<tr>
<td>0.52</td>
<td>0.48</td>
</tr>
<tr>
<td>0.40</td>
<td>0.43</td>
</tr>
<tr>
<td>0.30</td>
<td>0.31</td>
</tr>
<tr>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>0.05</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Matches uniform (null) distribution
Detecting stratification: Q-Q plots

Distribution of ranked association $-\log_{10} P$-values for 9 SNPs:

Matches uniform (null) distribution
**Detecting stratification: Q-Q plots**

Distribution of ranked association P-values for 9 SNPs:

<table>
<thead>
<tr>
<th>Expected</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.0001</td>
</tr>
<tr>
<td>0.20</td>
<td>0.0003</td>
</tr>
<tr>
<td>0.30</td>
<td>0.0008</td>
</tr>
<tr>
<td>0.40</td>
<td>0.01</td>
</tr>
<tr>
<td>0.50</td>
<td>0.07</td>
</tr>
<tr>
<td>0.60</td>
<td>0.13</td>
</tr>
<tr>
<td>0.70</td>
<td>0.47</td>
</tr>
<tr>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>0.90</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Detecting stratification: Q-Q plots

Distribution of ranked association P-values for 9 SNPs:

<table>
<thead>
<tr>
<th>Expected</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.0001</td>
</tr>
<tr>
<td>0.20</td>
<td>0.0003</td>
</tr>
<tr>
<td>0.30</td>
<td>0.0008</td>
</tr>
<tr>
<td>0.40</td>
<td>0.01</td>
</tr>
<tr>
<td>0.50</td>
<td>0.07</td>
</tr>
<tr>
<td>0.60</td>
<td>0.13</td>
</tr>
<tr>
<td>0.70</td>
<td>0.47</td>
</tr>
<tr>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>0.90</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Does not match uniform (null) distribution
Detecting stratification: Q-Q plots

Distribution of ranked association $-\log_{10} P$-values for 9 SNPs:

<table>
<thead>
<tr>
<th>Expected $-\log_{10} P$:</th>
<th>Actual $-\log_{10} P$:</th>
<th>Does not match uniform (null) distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>4.00</td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>3.52</td>
<td></td>
</tr>
<tr>
<td>0.52</td>
<td>3.10</td>
<td></td>
</tr>
<tr>
<td>0.40</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>1.15</td>
<td></td>
</tr>
<tr>
<td>0.22</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>
Detecting stratification: $\lambda_{GC}$

Chisq statistic = (#SAMPLES) correlation(genotype,phenotype)$^2$

$$\lambda_{GC} = \frac{\text{Median chisq statistic across all SNPs}}{0.455}$$

$\lambda_{GC} \approx 1 \Leftrightarrow$ No population stratification

$\lambda_{GC} > 1 \Leftrightarrow$ Inflation due to stratification (or other reasons)

[Note: $\chi^2(1\text{ dof}) = 0.455 \Leftrightarrow P\text{-value} = 0.5000$]

[Note: median is more robust to outliers than average]

Devlin & Roeder 1999 Biometrics; also see Reich & Goldstein 2001 Genet Epidemiol
Detecting stratification: Q-Q plots

Distribution of ranked association $-\log_{10} P$-values for 9 SNPs:

$\lambda_{GC} \approx 1 \Leftrightarrow$ no stratification
Detecting stratification: Q-Q plots

Distribution of ranked association $-\log_{10} P$-values for 9 SNPs:

Does not match uniform (null) distribution

$\lambda_{GC} > 1 \Leftrightarrow$ inflation due to stratification (or other reasons)
Detecting stratification: Q-Q plots

T2D GWAS
500K SNPs

Approximately matches uniform (null) distribution

Diabetes Genetics Initiative 2007 Science
Detecting stratification: $\lambda_{GC}$

T2D GWAS
500K SNPs

Approximately matches uniform (null) distribution

$\lambda_{GC} = 1.05$
Detecting stratification: $\lambda_{GC}$

T2D GWAS
500K SNPs

<table>
<thead>
<tr>
<th>Observed (-logP)</th>
<th>Expected (-logP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$\lambda_{GC} = 1.05 \iff P\text{-value} = 0.49 (\approx 0.50)$
$log_{10}P = 0.31 (\approx 0.30)$

median
chisq = 0.48 ($\approx 0.455$)
Detecting stratification: Q-Q plots

RA GWAS
100K SNPs

Does not match uniform (null) distribution

Plenge et al. 2007 Nat Genet
Detecting stratification: $\lambda_{GC}$

RA GWAS
100K SNPs

Does not match uniform (null) distribution

$\lambda_{GC} = 1.19$

Plenge et al. 2007 Nat Genet
Detecting stratification: $\lambda_{GC}$

RA GWAS
100K SNPs

$\lambda_{GC} = 1.19 \Leftrightarrow \text{P-value} = 0.46$ (not 0.50)
$log_{10}P = 0.34$ (not 0.30)

Plenge et al. 2007 Nat Genet
How much inflation in $\lambda_{GC}$ is OK??

- Short answer: $\lambda_{GC} \leq 1.05$ is usually considered OK
Genomic Control

Detecting stratification: compute $\lambda_{GC}$ (Genomic Control) statistic.

And now for something completely different …

Correcting for stratification: apply Genomic Control correction.
Methods to correct for stratification

Genomic Control: Devlin & Roeder 1999 Biometrics
Reich & Goldstein 2001 Genet Epidemiol

STRAT: Pritchard et al. 2000 Am J Hum Genet
(STRUCTURE)

EIGENSTRAT: Price et al. 2006 Nat Genet
(PCA)

PLINK: Purcell et al. 2007 Am J Hum Genet

Genetic matching: Luca et al. 2008 Am J Hum Genet

EMMAX: Kang et al. 2010 Nat Genet
(Mixed models)

Price et al. 2010 Nat Rev Genet
(Or, use family-based association tests)

**TDT:** Spielman et al. 1993 Am J Hum Genet

**FBAT:** Lange et al. 2002 Am J Hum Genet

**QTD:** Abecasis et al. 2000 Am J Hum Genet

Laird & Lange et al. 2006 Nat Rev Genet; also see Price et al. 2010 Nat Rev Genet
STRUCTURE (from previous slides)

General case: $M$ SNPs ($m = 1$ to $M$), $N$ populations ($n = 1$ to $N$), unknown allele frequency $p_{mn}$ for SNP $m$ in population $n$, observed genotype counts $g_{im}$ for SNP $m$ in many individuals $x_i$.

Which ancestries $\alpha_{in}$ and allele frequencies $p_{mn}$ maximize

$$
\prod_{i=1}^{I} \prod_{m=1}^{M} \left( \sum_{n=1}^{N} \alpha_{in} p_{mn} \right) g_{im} \left( \sum_{n=1}^{N} \alpha_{in} (1 - p_{mn}) \right)^{2 - g_{im}}
$$

• Approach #2: Place Bayesian priors on $\alpha_{in}$ and $p_{mn}$, then sample from posterior via Markov Chain Monte Carlo (MCMC) (STRUCTURE program)

Pritchard et al. 2000 Genetics
STRAT method (Structured Association)

Step 1: Use STRUCTURE to infer subpopulation clusters.
Step 2: Evaluate evidence of association within each cluster.

Pritchard et al. 2000 Am J Hum Genet
STRAT method (Structured Association)

NULL MODEL:
Assume ancestries $\alpha_{in}$ known (STRUCTURE or ADMIXTURE).
For candidate SNP $m$, use EM algorithm to maximize $L_{\text{NULL}}(\text{data} \mid \text{allele frequencies } p_{mn}) =$

$$
\prod_{i=1}^{I} \left( \sum_{n=1}^{N} \alpha_{in} p_{mn} \right) g_{im} \left( \sum_{n=1}^{N} \alpha_{in} \left( 1 - p_{mn} \right) \right)^{2 - g_{im}}
$$

Pritchard et al. 2000 Am J Hum Genet
STRAT method (Structured Association)

CAUSAL MODEL:
Assume ancestries $\alpha_{in}$ known (STRUCTURE).
For candidate SNP $m$, use EM algorithm to maximize $L_{CAUSAL}(\text{data} \mid \text{allele frequencies } p_{mn,\text{case}}, p_{mn,\text{control}}) =$

$$ \prod_{i=1}^{I_{\text{case}}} \left( \sum_{n=1}^{N} \alpha_{in} p_{mn,\text{case}} \right) g_{im} \left( \sum_{n=1}^{N} \alpha_{in} \left( 1 - p_{mn,\text{case}} \right) \right) \left( \sum_{n=1}^{N} \alpha_{in} \left( 1 - p_{mn,\text{control}} \right) \right)^2 - g_{im} $$

$$ \prod_{i=1}^{I_{\text{control}}} \left( \sum_{n=1}^{N} \alpha_{in} p_{mn,\text{control}} \right) g_{im} \left( \sum_{n=1}^{N} \alpha_{in} \left( 1 - p_{mn,\text{control}} \right) \right) \left( \sum_{n=1}^{N} \alpha_{in} \left( 1 - p_{mn,\text{control}} \right) \right)^2 - g_{im} $$

Pritchard et al. 2000 Am J Hum Genet
STRAT method (Structured Association)

Association statistic: \[ \Lambda = \frac{L_{\text{CAUSAL}}}{L_{\text{NULL}}} \]

To assess statistical significance: OR JUST \(2\log(\Lambda) = \chi^2(1\text{dof})?\)

- Simulate data sets under NULL MODEL with same \(\alpha_{in}, p_{mn}\)
- Compare \(\Lambda\) in real data to \(\Lambda\) in simulated data to get P-value

Relationship between \(p_{mn,\text{case}}\) and \(p_{mn,\text{control}}\) in CAUSAL MODEL:
- Use relative risk \(R\) (same for all populations \(n\))
STRAT method (Structured Association)

Association statistic: \[ \Lambda = \frac{L_{\text{CAUSAL}}}{L_{\text{NULL}}} \]

To assess statistical significance: OR JUST \[ 2\log(\Lambda) = \chi^2(1\text{dof})? \]

- Simulate data set s under NULL MODEL with same \( \alpha, p_{mn} \)
- Compare \( \Lambda \) in real data to \( \Lambda \) in simulated data to get P-value

Relationship between \( p_{mn,\text{case}} \) and \( p_{mn,\text{control}} \) in CAUSAL MODEL:

- Use relative risk \( R \) (same for all populations \( n \))
STRAT method (Structured Association)

Association statistic: \[ \Lambda = \frac{L_{\text{CAUSAL}}}{L_{\text{NULL}}} \]

To assess statistical significance:
• Simulate data sets under NULL MODEL with same \( \alpha_{in}, p_{mn} \)
• Compare \( \Lambda \) in real data to \( \Lambda \) in simulated data to get P-value

Relationship between \( p_{mn,\text{case}} \) and \( p_{mn,\text{control}} \) in CAUSAL MODEL:
• Use relative risk \( R \) (same for all populations \( n \))
• Or, use different relative risk \( R_n \) for each ancestry \( n \) (at locus \( m \))
• Or, let relative risk \( R_\alpha \) depend on genome-wide ancestry \( \alpha = (\alpha_n) \)

Pritchard et al. 2000 Am J Hum Genet
Correcting for stratification: PCA
# Methods to correct for stratification

**Genomic Control:** Devlin & Roeder 1999 Biometrics  
Reich & Goldstein 2001 Genet Epidemiol

**STRAT:** Pritchard et al. 2000 Am J Hum Genet  
(STRUCTURE)

**EIGENSTRAT:** Price et al. 2006 Nat Genet  
(PCA)

**PLINK:** Purcell et al. 2007 Am J Hum Genet

**Genetic matching:** Luca et al. 2008 Am J Hum Genet

**EMMAX:** Kang et al. 2010 Nat Genet  
(Mixed models)

Price et al. 2010 Nat Rev Genet
PCA on individual genotypes (from last week)

\( \mathbf{G} = M \times N \) 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 \( i \): \( p_i = \text{Avg}_i g_{ij} \), set \( g_{ij} = g_{ij} - p_i \)
  
  (Missing data: set \( g_{ij} = 0 \) if SNP \( i \) in individual \( j \) is missing data)

• Optional: normalize by \( \sqrt{p_i(1-p_i)} \), i.e. set \( g_{ij} = g_{ij} / \sqrt{p_i(1-p_i)} \)

\( \mathbf{\Psi} = N \times N \) covariance matrix of \( \mathbf{G} \)

\( \mathbf{\Psi} = \mathbf{V} \mathbf{D} \mathbf{V}^T \) (Eigen-decomposition)

Columns of \( \mathbf{V} \) are eigenvectors (principal components, PCs) of \( \mathbf{G} \).

Diagonal entries of \( \mathbf{D} \) are eigenvalues of \( \mathbf{G} \).

The hope: Top PCs (PC1, PC2) correspond to genetic ancestry.

Price et al. 2006 Nat Genet, Patterson et al. 2006 PLoS Genet
also see McVean 2009 PLoS Genet, Engelhardt & Stephens 2010 PLoS Genet
Correcting for stratification: EIGENSTRAT

1. Apply principal components analysis to genotype data to infer continuous axes of genetic variation.

2. For each inferred axis $k$:
   Subtract from each genotype and each phenotype an amount attributable to ancestry along that axis.

For genotypes at SNP $i$:

$$g_{ij, adjusted} = g_{ij} - \sum_{k=1}^{K} c_{ik} v_{jk}, \text{ where } c_{ik} = \frac{\sum v_{jk} g_{ij}}{\sum v_{jk}^2}$$

($v_{jk} = \text{PC coordinate of individual } j \text{ on axis } k$)

Price et al. 2006 Nat Genet; also see Zhu et al. 2002 Genet Epidemiol
Correcting for stratification: EIGENSTRAT

1. Apply principal components analysis to genotype data to infer continuous axes of genetic variation.

2. For each inferred axis $k$:
   Subtract from each genotype and each phenotype an amount attributable to ancestry along that axis.

   For phenotypes:
   \[ \pi_{j,\text{adjusted}} = \pi_j - \sum_{k=1}^{K} c_k v_{jk} \]
   \[ c_k = \frac{\sum v_{jk} \pi_j}{\sum v_{jk}^2} \]
   \[(v_{jk} = \text{PC coordinate of individual } j \text{ on axis } k)\]
Correcting for stratification: EIGENSTRAT

1. Apply principal components analysis to genotype data to infer continuous axes of genetic variation.

2. For each inferred axis $k$:
   Subtract from each genotype and each phenotype an amount attributable to ancestry along that axis.

1. Evaluate association between ancestry-adjusted genotypes and ancestry-adjusted phenotypes.

Price et al. 2006 Nat Genet; also see Zhu et al. 2002 Genet Epidemiol
Correcting for stratification: EIGENSTRAT

1. Apply principal components analysis to genotype data to infer continuous axes of genetic variation.

2. For each inferred axis \( k \):
   Subtract from each genotype and each phenotype an amount attributable to ancestry along that axis.

1. Evaluate association between ancestry-adjusted genotypes and ancestry-adjusted phenotypes.

   \[
   \text{Chisq statistic} = (N-K-1) \cdot \text{corr}(g_{ij,\text{adjusted}}, \pi_{j,\text{adjusted}})^2
   \]

Price et al. 2006 Nat Genet; also see Zhu et al. 2002 Genet Epidemiol
1. Apply principal components analysis to genotype data to infer continuous axes of genetic variation.

2. For each inferred axis $k$: Subtract from each genotype and each phenotype an amount attributable to ancestry along that axis.

1. Evaluate association between ancestry-adjusted genotypes and ancestry-adjusted phenotypes.

Note: equivalent to including $v_{jk}$ as covariates

$$\pi_j = \alpha_i g_{ij} + \sum_{k=1}^{K} \beta_k v_{jk}$$

Price et al. 2006 Nat Genet; also see Zhu et al. 2002 Genet Epidemiol
Genome-wide Association Studies and Population stratification

1. Genome-wide Association Studies
2. Population Stratification
3. Visualizing GWAS Results
3. Phasing and Imputation
Visualization of GWAS Results

- PLINK graphical output
- User added custom tracks in the UCSC browser
  - http://genome.ucsc.edu/
  - http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#CustomTracks
- WGAviwer
  - Locus Zoom plots
    - http://csg.sph.umich.edu/locuszoom/
- Homemade graphs
  - http://csg.sph.umich.edu/locuszoom/
Display Results in UCSC Genome Browser
Display Results in WGAvviewer
Genome-wide Association Studies and Population stratification

1. Genome-wide Association Studies
2. Population Stratification
3. Visualizing GWAS Results
3. Phasing and Imputation
Phasing and Imputation

In general, it is not possible to directly infer haplotypes from genotypes:

These two scenarios have different haplotypes but identical genotypes.
Phasing and Imputation

Humans and most eukaryotic model organisms are **diploid**: we have 2 chromosomal copies of most of our DNA.

<table>
<thead>
<tr>
<th>G</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
</tr>
</tbody>
</table>

The collection of alleles (from one or more sites) on a pair of chromosomes is called a **genotype**:

G/G  C/T  A/A  C/T  C/C

The collection of alleles on a single chromosome are called a **haplotype**
Phasing and Imputation

In general, it is not possible to directly infer haplotypes from genotypes:

These two scenarios have different haplotypes but identical genotypes:
Phasing and Imputation

Experimental or computational methods are employed to directly determine or statistically infer haplotypic phase.

Allele-specific PCR can be used to amplify a single haplotype for sequencing.

However, ‘allelic dropout’ or ‘differential amplification’ can lead to miscalls, and nested (redundant) primers must be used to account for this.

Obtaining phased data takes roughly 10 times the amount of effort as obtaining unphased genotype data. Even more expensive/intensive to obtain long-range haplotypes
Phasing and Imputation

- Current algorithms are designed for use with genome-wide data (for the most part, since that’s what we have)
- Examples: fastPHASE (Scheet/Stephens, the first phaser capable of use with genome-wide data), Beagle (Brownings), MACH (Li/Abecasis), SHAPEIT (Delaneau/Marchini), impute2 (Howie et al.)
- All programs do work on chromosomes, but they will take hours to days per chromosome and dataset.
  - All are freely available, though, and should run fine on 64-bit linux. Beagle should run on 32-bit.
Phasing and Imputation

- Phasing also allows you to impute untyped variants

- For imputation, your reference haplotypes can have many variants that aren’t typed on your GWAS chip
# Phasing and Imputation

## Typical imputation scenario

<table>
<thead>
<tr>
<th>HapMap or 1,000 Genomes</th>
<th>Reference haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 1 1 0 0 1 1 0 0 0 1 1 1</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 0 1 1 1 0 1 1 1 0 0 1</td>
<td></td>
</tr>
<tr>
<td>1 1 1 1 1 0 0 0 1 0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1 0 1 1 0 0 0 1 1 1 1 1 0 0 1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cases and controls typed on SNP chip</th>
<th>Study genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ? ? ? 2 ? 0 ? ? ? ? 0 0 ? 0</td>
<td>0 0 ? 0</td>
</tr>
</tbody>
</table>
Phasing and Imputation

- The mosaics of haplotypes are drawn by haplotypic similarity...do you think it is better to err on the side of including more reference individuals or less?

- In general, more. Always use multiple populations. Larger reference panels will in general be more accurate, although imputation will take longer.
Phasing and Imputation

- Best imputation practices:
  - Use high-quality genotypes (potentially worse problem with imputation than GWAS)
  - Ensure everything is on the forward strand...
  - Most algorithms can jointly phase/impute, but...
    - If you impute first, you can substantially speed up your work. Like from weeks to days. With a minimal power tradeoff

- Evaluate output
Phasing and Imputation

- Output will come with a couple of different files, typically.
- Often, estimated untyped sites won’t be perfect, they will include some uncertainty.
- Captured in Dosages…
- Instead of AA, AG, GG:
- Continuous 0-1 counts
- Per genotype
Advantages:
• More SNPs!
• Better coverage of the genome than SNPs alone (using imputation accounts for linkage disequilibrium better than just chip genotypes)

Narrowing of signal... (Fine mapping)
Phasing and Imputation

Fine mapping
Understanding the Biological data:
Work is done, relax on beach?