Week 8: Genome-wide Association Studies (GWAS)

Dr. Mengling FENG
Institute for Infocomm Research
Massachusetts Institute of Technology
mfeng@mit.edu
PLANS FOR WEEK 7 AND WEEK 8

- Week 7, 1\textsuperscript{st} Oct 2015
  - 2 hours class: Single (Simple) Nucleotide Polymorphism
  - 1 hour briefing on project and forming of project teams

- Week 8, 7\textsuperscript{th} Oct 2015
  - Definition of SNP
  - Q & A
  - 2 hours class: Genome-wide Association Study (GWAS)
  - 30 mins Q&A on the lectures and project
**Week 8’s Learning Objectives**

- After the class, students should be able to
  - Define Gene-disease association studies
  - Appreciate the motivations and applications of GWAS
  - Explain the differences between GWAS and Candidate Gene Studies
  - Explain the typical method and workflow for GWAS studies and, more importantly, considerations and limitations for each step
  - Understand the concepts of
    - Linkage Disequilibrium
    - Hypothesis testing
    - Multiple testing correction
    - Population stratification bias
  - Get to know the online resources
GENETIC ASSOCIATION STUDIES

- Investigate how genotypes may associate or cause particular phenotypes
- Genome-Wide Association Study (GWAS)
  - A type of genetic association study
  - Focus on common SNPs
  - Involves large amount of SNPs
- A overview
  - https://www.youtube.com/watch?v=mblaqn4yU70
MOTIVATIONS AND VISION

- Preventive Medication

"Superior Doctors Prevent the Disease.
Mediocre Doctors Treat the Disease Before Evident.
Inferior Doctors Treat the Full Blown Disease."

-Huang Dee: Nai - Ching (2600 B.C. 1st Chinese Medical Text)
THE VISION: PREVENTIVE MEDICATION

- Prevent disease from occurring
  - SNPs as bio-markers to estimate personalized disease risk
  - Inspire behavioral and environmental changes
  - Some preventive intervention
THE VISION: PREVENTIVE MEDICATION

- Prevent disease from occurring
- Identify the cause of the disease
  - Genomics identifies the cause of disease
  - “All medicine may become pediatrics” Paul Wise, Professor of Pediatrics, Stanford Medical School, 2008
  - Treat the cause of the disease rather than the symptoms
- Health care costs can be greatly reduced if
  - Invests in preventive medicine
  - One targets the cause of disease rather than symptoms
- Challenges and limitations:
  - Penetrance and environmental factors
Penetrance and Environmental Factors

- **Penetrance**
  - Is the proportion of individuals carrying a particular variant of a gene (allele or genotype) that also expresses an associated trait (phenotype).

- **Highly penetrant Mendelian single gene diseases**
  - Huntington’s Disease caused by excess CAG repeats in huntingtin’s protein gene
  - Autosomal dominant, 100% penetrant, invariably lethal

- **Reduced penetrance, some genes lead to a predisposition to a disease**
  - BRCA1 & BRCA2 genes can lead to a familial breast or ovarian cancer
  - Disease alleles lead to 80% overall lifetime chance of a cancer, but 20% of patients with the rare defective genes show no cancers

- **Complex diseases requiring alleles in multiple genes**
  - Many cancers (solid tumors) require somatic mutations that induce cell proliferation, mutations that inhibit apoptosis, mutations that induce angiogenesis, and mutations that cause metastasis
  - Cancers are also influenced by environment (smoking, carcinogens, exposure to UV)

- **Some complex diseases have multiple causes**
  - Genetic vs. spontaneous vs. environment vs. behavior
  - Some complex diseases can be caused by multiple pathways
  - Type 2 Diabetes can be caused by reduced beta-cells in pancreas, reduced production of insulin, reduced sensitivity to insulin (insulin resistance) as well as environmental conditions (obesity, sedentary lifestyle, smoking etc.).
CANDIDATE GENE STUDIES VS GWAS
CANDIDATE GENE STUDIES vs GWAS
TYPES OF GWAS DISEASE ASSOCIATIONS (BINARY)
TYPES OF GWAS QUANTITATIVE TRAITS
Typical Steps of GWAS

- **Sampling** (Case-Control method)
- **Genotyping** (Data generation & collection)
- **Quality Control** (Data pre-processing)
- **Statistical Testing** (Data analysis)
- **Replication** (Verification)
Sampling (Case & Control)

- Primer on Causal Inference
  - Definition of Causation
DEFINITION OF CAUSATION
**Definition of Causation**

<table>
<thead>
<tr>
<th>Name</th>
<th>$Y_{red=0}$</th>
<th>$Y_{red=1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rheia</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Kronos</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Demeter</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hades</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Hestia</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Poseidon</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hera</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zeus</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Artemis</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Apollo</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Leto</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Ares</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Athena</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hephaestus</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Aphrodite</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cyclope</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Persephone</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hermes</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hebe</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dionysus</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$P[Y_{red=0} = 1] = 8/20 = 0.4$

$P[Y_{red=1} = 1] = 12/20 = 0.6$

$P[Y_{red=0} = 1] \neq P[Y_{red=1} = 1]$

indicates sign of causation

**Red Pill** => **Cure**
**Sampling (Case & Control)**

- A matched cohort
  - Age
  - Gender
  - Other demographics
  - Ancestry profile
POTENTIAL BIASES
POPULATION STRATIFICATION

Control

Case (disease)

Northen

Southen

Northen

Southen

18
Detection of Population Stratification Genomic Control

Observed Armitage Trend Statistics

\[ Y^2 = \frac{N(N(r_1 + 2r_2) - R(n_1 + 2n_2))^2}{R(N - R)(N(n_1 + 4n_2) - (n_1 + 2n_2)^2)} \]

Expected Chi2 Statistics

\[ \chi^2 \sim X^2_A = \frac{2N(2N(r_1 + 2r_2) - R(n_1 + 2n_2))^2}{4R(N - R)(2N(n_1 + 2n_2) - (n_1 + 2n_2)^2)} \]

Measure of population stratification \( \lambda \)

\[ Y^2 \sim \lambda \chi^2_1 \]

\( \lambda = 1 \quad \Rightarrow \text{No population stratification} \)

\( \lambda > 1.05 \quad \Rightarrow \text{Significant population stratification} \)
**Population Stratification QQ Plot & Corrections**

- **Solutions**
  - Remove the deviating SNPs
  - Conduct separate studies for different subpopulations
Typical Steps of GWAS

- **Sampling** (Case-Control method)
- **Genotyping** (Data generation & collection)
- **Quality Control** (Data pre-processing)
- **Statistical Testing** (Data analysis)
- **Replication** (Verification)
GENOTYPING
NAÏVE APPROACH

- Identify all 10 million common SNPs
- Collect 1,000 cases and 1,000 controls
- Genotype all DNAs for all SNPs
- That adds up to 20 billion genotypes

This won’t work in practice:
  - Cost:
    - In 2002, this approach cost 50 cents a genotype.
    - That was $10 billion for each disease – completely out of the question
    - Nowadays, 50 cents/2000 genotypes => $500K per disease
  - Statistical:
    - Multiple test correction => lead to lower power => high rate of false negative
**Haplotype:** Genetic Home Reference

- A set of SNPs (mutations) on the same chromosome that tend to be inherited together
- SNPs can be highly correlated => sub-sampling is possible
Detection of Haplotype: Linkage Disequilibrium

- Linkage disequilibrium:
  - Nonrandom association of alleles at two or more loci
**Linkage Disequilibrium (LD)**

At Equilibrium (independence)

$$P_{AB} = P_A P_B$$

At Disequilibrium (dependence)

$$P_{AB} \neq P_A P_B$$

Linkage Disequilibrium Coefficient $D$

$$D = P_{AB} - P_A P_B$$

Linkage Disequilibrium $r^2$

$$r^2 = \frac{D^2}{P_A P_a P_B P_b} = \frac{D^2}{P_A (1 - P_A) P_B (1 - P_B)}$$

<table>
<thead>
<tr>
<th>Locus B</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>b</td>
</tr>
<tr>
<td>p_{AB}</td>
<td>p_{Ab}</td>
</tr>
<tr>
<td>p_{aB}</td>
<td>p_{ab}</td>
</tr>
<tr>
<td>p_A</td>
<td>p_a</td>
</tr>
<tr>
<td>p_B</td>
<td>p_b</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Locus A</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>a</td>
</tr>
<tr>
<td>p_{AB}</td>
<td>p_{Ab}</td>
</tr>
<tr>
<td>p_{aB}</td>
<td>p_{ab}</td>
</tr>
<tr>
<td>p_A</td>
<td>p_a</td>
</tr>
<tr>
<td>p_B</td>
<td>p_b</td>
</tr>
</tbody>
</table>

25
LINKAGE DISEQUILIBRIUM

LD Map Type: r-square

0 0.2 0.4 0.6 0.8 1

Tag SNP
TAGSNPS (TAGGING SNPS)

- A SNP that can represent a group of SNPs

- Typical steps to identify tagSNPs:
  - Identify the search region
  - Define the metric for assessing the tagging
    - How well the tagSNP/tagSNPs predict their neighbors
  - Select an algorithm
  - Validate the performance of the learned tagSNPS
**Selection of tagSNPs: Principal Component Analysis (PCA)**

- A orthogonal transformation to concise represent a set of data

- Principal components are the eigenvectors of the covariance matrix $X^T X$
**Steps of Principal Component Analysis**

Give data $X$

- **Step 1:** Data normalization
- **Step 2:** Calculate eigenvalues and eigenvectors of $X^TX$
- **Step 3:** Sort the eigenvalues
- **Step 4:** Pick the top $k$ eigenvalues and the corresponding eigenvectors (the principal components); one may adopt a threshold cut off selection strategy
  
  $\Rightarrow W$

- **Step 5:** Project your data onto the principal components
  - $T = XW$

- Alternative Approach
PCA FOR tagSNPs

LD Map Type: r-squared

Tag SNP
Genotyping with tagSNPs

- Identify all 300K (instead of 10 million) tagSNPs
- Collect 1,000 cases and 1,000 controls
- Genotype all DNAs for all SNPs
- That adds up to 600 million genotype
  - Reduction in three magnitudes
  - Genotype costs can be reduced to just thousands of dollars
The HapMap Project

Typical Steps of GWAS

- **Sampling** (Case-Control method)
- **Genotyping** (Data generation & collection)
- **Quality Control** (Data pre-processing)
- **Statistical Testing** (Data analysis)
- **Replication** (Verification)
**Statistical Testing**

- **Effect size**: Odds ratio
  - Ratio of odds

\[
OR = \frac{D_E/H_E}{D_{NE}/H_{NE}}
\]

- **Statistical significance**
  - Chi-2 test to obtain p-value

- **For single hypothesis testing**:
  - To control Type-I error (false positive) to be below 5%
  - p-value cut off at 0.05

- **BUT**, we are not testing only 1 SNP but 300K of them
**Multiple Test Effect**

\[
P(\text{detecting an effect when there is none}) = \alpha = 0.05
\]
\[
P(\text{not detecting an effect when there is none}) = 1 - \alpha
\]
\[
P(\text{not detecting an effect when there is none in every experiment}) = (1 - \alpha)^k
\]
\[
P(\text{detecting an effect when there is none on at least one experiment}) = 1 - (1 - \alpha)^k
\]

---

**Graph:**

- **Y-axis:** Probability of at least one spurious finding.
- **X-axis:** Number of Tests

- **Label:** "Familywise Error Rate" with \(\alpha = 0.05\)
MULTIPLE TEST CORRECTION

- Bonferroni Correction
  - Just divide by the number of hypotheses
    \[ \alpha_c = \frac{\alpha}{k} \]

- Šidák Correction
  - Asserts independence
    \[ \alpha = 1 - (1 - \alpha_c)^k \]
    \[ \alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}} \]
Typical Steps of GWAS

- **Sampling** (Case-Control method)
- **Genotyping** (Data generation & collection)
- **Quality Control** (Data pre-processing)
- **Statistical Testing** (Data analysis)
- **Replication** (Verification)
GOLDEN AGE OF GWAS
THE VISION: PREVENTIVE MEDICATION

- Prevent disease from occurring
- Identify the cause of the disease
  - Genomics identifies the cause of disease
  - “All medicine may become pediatrics” Paul Wise, Professor of Pediatrics, Stanford Medical School, 2008
  - Treat the cause of the disease rather than the symptoms
- Health care costs can be greatly reduced if
  - Invests in preventive medicine
  - One targets the cause of disease rather than symptoms
- Challenges and limitations:
  - Penetrance and environmental factors
WEEK 8’S LEARNING OBJECTIVES

- After the class, students should be able to
  - Define Gene-disease association studies
  - Appreciate the motivations and applications of GWAS
  - Explain the differences between GWAS and Candidate Gene Studies
  - Explain the typical method and workflow for GWAS studies and, more importantly, considerations and limitations for each step
  - Understand the concepts of
    - Linkage Disequilibrium
    - Hypothesis testing
    - Multiple testing correction
    - Population stratification bias
  - Get to know the online resources