ACCURACY AND ANALYSIS OF GWAS (GENOME WIDE ASSOCIATION STUDIES)

Karen Chau
CSC 334
Fall 2015
GWAS BACKGROUND

- GWAS identifies genetic variations associated with diseases
  - Benefits
    - Detection, treatment, prevention
    - Personalized medicine
  - Ethics
    - Prenatal testing
  - Refinement
    - Increase efficiency and decrease cost
1
in
11
Americans
have
DIABETES

*American Diabetes Association (2012)
Americans have 
PARKINSON'S DISEASE

1 million

Americans have PARKINSON'S DISEASE

*Parkinson’s Disease Foundation (2015)*
American dies from HEART DISEASE every 90 seconds.

*American Heart Association (2015)*
GWAS METHOD

- Compares SNPs between those with and without the disease
GWAS METHOD

- Compares SNPs between those with and without the disease

SNP 2
- No association to disease

SNP 3
- Associated to disease
GWAS METHOD

- Compares SNPs between those with and without the disease

- Each disease
  - Many experiments
  - Each experiment
    - Series of associated SNP
      - Each SNP has a p-value
Hypothesis testing

- Null hypothesis
  - Observations are due to chance
- Alternative hypothesis
  - Observations are influenced

- Test statistic
  - Standardized value calculated from sample

- P-value
  - Pr(observing the test statistic, assuming $H_o$ is true)
WHAT DOES THE DATA MEAN?

- **Accuracy**
  1) How accurate were the studies compared to each other?

- **Analysis**
  1) Given a SNP:
     - What disease(s) are associated with it?
  2) Given a disease:
     - Which SNP(s) are associated with it?
     - Which chromosome(s) are more associated with it?
DETERMINING ACCURACY OF GWAS DATA

- Consolidate SNPs from all studies and filter out >0.05 p-values

[First experiment]
**DETERMINING ACCURACY OF GWAS DATA**

- Consolidate SNPs from all studies and filter out >0.05 p-values

---

Top 35 associations for Coronary Artery Disease (HGVM3648) from data set HGVRS3891 (35 in database)

Export as Microsoft Excel ➔ Go

<table>
<thead>
<tr>
<th>Rank</th>
<th>Identifier</th>
<th>Accession</th>
<th>Region</th>
<th>p-value</th>
<th>-log p-value</th>
<th>Effect size</th>
<th>Risk allele</th>
<th>Related data</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HGVM1880766</td>
<td>rs1558902</td>
<td>chr16:53803574..53803574</td>
<td>0.04317</td>
<td>1.365</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
<tr>
<td>2</td>
<td>HGVM5089460</td>
<td>rs12444979</td>
<td>chr16:19933600..19933600</td>
<td>0.06869</td>
<td>1.163</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
<tr>
<td>3</td>
<td>HGVM3473352</td>
<td>rs10767664</td>
<td>chr11:27725986..27725986</td>
<td>0.1325</td>
<td>0.878</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
<tr>
<td>4</td>
<td>HGVM1405127</td>
<td>rs2531995</td>
<td>chr16:4013467..4013467</td>
<td>0.1373</td>
<td>0.862</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
<tr>
<td>5</td>
<td>HGVM118737</td>
<td>rs887912</td>
<td>chr2:69302877..69302877</td>
<td>0.1735</td>
<td>0.761</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
<tr>
<td>6</td>
<td>HGVM2862188</td>
<td>rs4929949</td>
<td>chr11:8604593..8604593</td>
<td>0.1765</td>
<td>0.753</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
<tr>
<td>7</td>
<td>HGVM1872195</td>
<td>rs1514175</td>
<td>chr1:74991644..74991644</td>
<td>0.1856</td>
<td>0.731</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
<tr>
<td>8</td>
<td>HGVM759087</td>
<td>rs543874</td>
<td>chr1:177889480..177889480</td>
<td>0.2111</td>
<td>0.676</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
<tr>
<td>9</td>
<td>HGVM3675423</td>
<td>rs109668576</td>
<td>chr9:28414339..28414339</td>
<td>0.2118</td>
<td>0.674</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
<tr>
<td>10</td>
<td>HGVM498801</td>
<td>rs571312</td>
<td>chr18:57839769..57839769</td>
<td>0.2363</td>
<td>0.627</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>[Details]</td>
<td>Not supplied</td>
</tr>
</tbody>
</table>
Consolidate SNPs from all studies and filter out >0.05 p-values

[Second experiment]
**DETERMINING ACCURACY OF GWAS DATA**

- Consolidate SNPs from all studies and filter out >0.05 p-values

<table>
<thead>
<tr>
<th>Rank</th>
<th>Identifier</th>
<th>Accession</th>
<th>Region</th>
<th>p-value</th>
<th>Jon p-value</th>
<th>Effect size</th>
<th>Risk allele</th>
<th>Related data</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HGVM498801</td>
<td>rs571312</td>
<td>chr18:57839769..57839769</td>
<td>0.0002495</td>
<td>3.603</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>Details</td>
<td>Not supplied</td>
</tr>
<tr>
<td>2</td>
<td>HGVM1405127</td>
<td>rs2531995</td>
<td>chr16:4013467..4013467</td>
<td>0.002647</td>
<td>2.577</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>Details</td>
<td>Not supplied</td>
</tr>
<tr>
<td>3</td>
<td>HGVM1880766</td>
<td>rs1556902</td>
<td>chr16:53803574..53803574</td>
<td>0.002729</td>
<td>2.564</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>Details</td>
<td>Not supplied</td>
</tr>
<tr>
<td>4</td>
<td>HGVM1682064</td>
<td>rs2867125</td>
<td>chr2:622827..622827</td>
<td>0.00452</td>
<td>2.345</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>Details</td>
<td>Not supplied</td>
</tr>
<tr>
<td>5</td>
<td>HGVM673226</td>
<td>rs987237</td>
<td>chr6:50803050..50803050</td>
<td>0.005968</td>
<td>2.224</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>Details</td>
<td>Not supplied</td>
</tr>
<tr>
<td>10</td>
<td>HGVM498801</td>
<td>rs571312</td>
<td>chr18:57839769..57839769</td>
<td>0.2363</td>
<td>0.627</td>
<td>Not supplied</td>
<td>Not supplied</td>
<td>Details</td>
<td>Not supplied</td>
</tr>
</tbody>
</table>

- Compute frequency of same SNPs
EXPECTED ACCURACY OF GWAS DATA

Common SNPs Across GWAS Experiments

<table>
<thead>
<tr>
<th>Number of Studies</th>
<th>Number of SNPs in Common</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
Given a SNP, what disease(s) are associated with it?

Hash Table

- `<KEY>,<VALUE>`
  - Example: dictionary with `<word>, <definition>`
  - O(1) retrieval and insertion

Question

- `<SNP#>,<DISEASE(S)>`
Given a disease, which SNP(s) are associated with it?
- Hash Table
  - \(<\text{DISEASE}>,<\text{SNP#(s)}>\)

Given a disease, which chromosome(s) are more associated with it?
- Hash Table
  - \(<\text{DISEASE}>,<\text{CHROMOSOME#(s)}>\)
EXTENSIONS

- Use R to plot the results visually
  - For particular disease
    - Look at the frequency of SNPs on each chromosome

Parkinson’s Disease

<table>
<thead>
<tr>
<th>Chromosome #</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Use R to plot the results visually

- For particular disease
  - Look at the frequency of SNPs on each chromosome
Explore more diseases
- Anorexia
- Sickle Cell Anemia
- Parkinson’s
- Schizophrenia
- and more!
REFERENCES

THANK YOU!

Questions/Comments?