Genetic Association Analysis

Clement Ma

Sequence Analysis Workshop

December 11, 2014

Lecture slides adapted from Hyun Min Kang and Gonçalo Abecasis
Outline

• Introduction

• Data overview

• Analysis of common variants

• Analysis of low-frequency variants
Genetic Association Analysis

INTRODUCTION
Genetic association studies

• Goal: Identify genetic variants associated with diseases and traits

• Why?
  – Improve understanding of genetic mechanisms underlying diseases and traits
  – Identify potential drug targets for new therapies
  – Screen individuals with high risk for disease
Genetic architecture of complex traits
Genetic architecture of complex traits

- Rare alleles causing Mendelian disease
- Low-frequency variants with intermediate effect
- Few examples of high-effect common variants influencing common disease
- Array-based genotyping
- Common variants implicated in common disease by GWA

- Effect size
  - High
  - Intermediate
  - Modest
  - Low

- Allele frequency
  - Very rare
  - Rare
  - Low frequency
  - Common
Genetic architecture of complex traits

- **Population sequencing**
- **Dense reference imputation into GWAS**
- **Specialized array genotyping**

- **Effect size**
  - High
  - Intermediate
  - Modest
  - Low

- **Allele frequency**
  - Very rare
  - Rare
  - Low frequency
  - Common

**Key points:**
- **Rare** alleles causing Mendelian disease
- **Low-frequency** variants with intermediate effect
- **Common** variants implicated in common disease by GWA

Few examples of high-effect **common** variants influencing common disease
Genetic architecture of complex traits

- **Array-based GWAS?**
  - Few examples of high-effect common variants influencing common disease

- **Family-based Sequencing**
  - Rare alleles causing Mendelian disease

- **Deep Genome with Very Large Samples?**
  - Rare variants of small effect very hard to identify by genetic means

- **Low-frequency variants with intermediate effect**

- **Common variants implicated in common disease by GWA**

Effect size

- High
- Intermediate
- Modest
- Low

Allele frequency

- Very rare
- Rare
- Low frequency
- Common
Genotype array-based GWAS identified thousands of associated variants

Published G-W significant associations
$(p \leq 5 \times 10^{-8})$ as of 12/2012

NHGRI GWA Catalog: http://www.genome.gov/GWAStudies/
5,783 SNPs from GWAS Catalog with $p \leq 5 \times 10^{-8}$
5,783 SNPs from GWAS Catalog with p ≤ 5 × 10^{-8}
Genome-wide significant SNPs by MAF

5,783 SNPs from GWAS Catalog with $p \leq 5 \times 10^{-8}$

**Common associated variants:**
- Have small effect sizes
- Explain modest proportion of total genetic heritability
Genome-wide significant SNPs by MAF

Low-frequency and rare associated variants may:
• Have larger effect sizes
• Explain larger proportion of trait heritability

5,783 SNPs from GWAS Catalog with p≤5x10^{-8}
DATA OVERVIEW
Phenotypes: binary trait

$n$ individuals

Y
1
0
1
1
0
...
0

Cases = 1
Controls = 0
Phenotypes: quantitative trait (QT)

\[
\begin{array}{c|c|c}
Y & \text{Y Transf.} & \\
0.08 & -2.5 & \\
0.84 & -0.2 & \\
0.17 & -1.8 & \\
0.19 & -1.7 & \\
0.35 & -1.0 & \\
... & ... & \\
1.39 & 0.3 & \\
\end{array}
\]

Transformation:
- Logarithm
- Inverse normal
Genotypes: hard genotypes

$n$ individuals

$m$ markers (SNPs)

<table>
<thead>
<tr>
<th>$g_1$</th>
<th>$g_2$</th>
<th>$g_3$</th>
<th>$g_4$</th>
<th>...</th>
<th>$g_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Genotype imputation

• **Goal**: to increase power by using previously genotyped GWAS samples

• **Problem**: GWAS samples genotyped at fewer or different variant sites

• **Method**: Use genotype imputation to fill in missing genotypes

Using genotype imputation to fill in missing genotypes

1. Starting Data

Genotyped sample

.. C .. G C

Reference haplotypes

A G A T C T C C T
A G C T C T C A T
A G A T C G C C T
A G A T C T A C T
Using genotype imputation to fill in missing genotypes

2. Identify shared regions of chromosome

Genotyped sample

Reference haplotypes

A G A T C T C C T
A G C T C T C A T
A G A T C G C C T
A G A T C T A C T
Using genotype imputation to fill in missing genotypes

3. Fill in missing genotypes

Genotyped sample

```
AGCT
CGCCC
```

Reference haplotypes

```
AGACT
AGCTC
AGACTGC
AGACTG
```
Genotypes: imputed dosages

$m$ markers (SNPs)

<table>
<thead>
<tr>
<th></th>
<th>$g_1$</th>
<th>$g_2$</th>
<th>$g_3$</th>
<th>$g_4$</th>
<th>...</th>
<th>$g_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.99</td>
<td>0.21</td>
<td>0.98</td>
<td>0.01</td>
<td>...</td>
<td>2.00</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>1.4</td>
<td>0.00</td>
<td>0.00</td>
<td>...</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
<td>0.8</td>
<td>1.00</td>
<td>0.00</td>
<td>...</td>
<td>2.00</td>
</tr>
<tr>
<td>4</td>
<td>1.34</td>
<td>1.6</td>
<td>0.03</td>
<td>0.00</td>
<td>...</td>
<td>1.99</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>0.89</td>
<td>1.00</td>
<td>0.03</td>
<td>...</td>
<td>0.99</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$n$</td>
<td>1.01</td>
<td>0.34</td>
<td>2.00</td>
<td>0.00</td>
<td>...</td>
<td>0.01</td>
</tr>
</tbody>
</table>

$n$ individuals

Imputation Quality Score

<table>
<thead>
<tr>
<th>$r^2_1$</th>
<th>$r^2_2$</th>
<th>$r^2_3$</th>
<th>$r^2_4$</th>
<th>...</th>
<th>$r^2_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.4</td>
<td>0.98</td>
<td>0.99</td>
<td>...</td>
<td>0.97</td>
</tr>
</tbody>
</table>

### Additional covariates

<table>
<thead>
<tr>
<th>$z_1$</th>
<th>$z_2$</th>
<th>...</th>
<th>$z_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Age</td>
<td>...</td>
<td>BMI</td>
</tr>
<tr>
<td>1</td>
<td>54</td>
<td></td>
<td>24.5</td>
</tr>
<tr>
<td>0</td>
<td>36</td>
<td></td>
<td>23.7</td>
</tr>
<tr>
<td>1</td>
<td>72</td>
<td></td>
<td>30.2</td>
</tr>
<tr>
<td>0</td>
<td>66</td>
<td></td>
<td>26.0</td>
</tr>
<tr>
<td>0</td>
<td>65</td>
<td></td>
<td>27.0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>0</td>
<td>55</td>
<td></td>
<td>22.7</td>
</tr>
</tbody>
</table>

$n$ individuals

$c$ covariates
Study individuals: relatedness and population structure

• Unrelated individuals

• Related individuals
  – Identify any relationships between individuals

• Population structure
  – Individuals are from different populations
ANALYSIS OF COMMON VARIANTS
Genetic architecture of complex traits

- Rare alleles causing Mendelian disease
- Low-frequency variants with intermediate effect
- Common variants implicated in common disease by GWA
- Array-based genotyping

Effect size:
- High
- Intermediate
- Modest
- Low

Allele frequency:
- Very rare
- Rare
- Low frequency
- Common
Single variant analysis

Test each variant for association with outcome

<table>
<thead>
<tr>
<th>Y</th>
<th>g₁</th>
<th>g₂</th>
<th>g₃</th>
<th>g₄</th>
<th>...</th>
<th>gₘ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Analysis methods

• Binary traits
  – Contingency table tests cannot adjust for covariates
    • Chi-square Test
    • Cochran-Armitage Trend Test
    • Fisher’s Exact Test
  – Logistic regression can account for covariates

• Quantitative traits
  – Linear regression
Visualizing results: Manhattan Plot

BMI GWAS (Stage 1)

- Previously reported loci
- New replicated loci
- Non replicated loci

(Willer et. al., Nat. Genet., 2009)
Visualizing results: quantile-quantile (QQ) plot

BMI GWAS (Stage 1)

b

-\log_{10} P\text{ value}

Expected -\log_{10} P\text{ value}

-\log_{10} P\text{ value}

0 1 2 3 4 5 6

-\log_{10} P\text{ value}

-\log_{10} P\text{ value}

-\log_{10} P\text{ value}

-\log_{10} P\text{ value}

(Willer et. al., Nat. Genet., 2009)
Visualizing results: regional plot

(Willer et. al., Nat. Genet., 2009)
Sources of association

• Causal association
  – Genetic marker alleles influence susceptibility

• Linkage disequilibrium
  – Genetic marker alleles associated with other nearby alleles that influence susceptibility

• Population stratification
  – Genetic marker is unrelated to disease alleles
Example of spurious association due to population stratification

<table>
<thead>
<tr>
<th></th>
<th>Allele 1</th>
<th>Allele 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Affected</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 1</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>(f_1,Aff=0.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 2</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>(f_1,Aff=0.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unaffected</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 1</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>(f_1,Unaff=0.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 2</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>(f_1,Aff=0.8)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\chi^2 = 0.00\)  \(p\)-value = 1.0

<table>
<thead>
<tr>
<th></th>
<th>Allele 1</th>
<th>Allele 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Affected</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>150</td>
<td>225</td>
</tr>
<tr>
<td>(f_1,Aff=0.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unaffected</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>225</td>
<td>150</td>
</tr>
<tr>
<td>(f_1,Aff=0.6)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\chi^2 = 29.2\)  \(p\)-value = 6.5\times10^{-8}
The stratification problem happens..

- If..
  - Phenotypes differ between populations
  - and allele frequencies have drifted apart

- Then..
  - Unlinked markers exhibit association
  - Not very useful for gene mapping!

- For example, Glaucoma has prevalence of ~2% in elderly Caucasians, but ~8% in African-Americans
Possible solutions for population stratification

• Avoid stratification by design
  – Collect a better matched sample by ancestry
  – Use family-based controls
    • E.g. apply Transmission Disequilibrium Test (TDT)

• Analyze association by population groups
  – Using self reported ethnicity or genetic markers
  – Carry out association analysis within each group

• Account for inflated false-positive rate
  1. Apply genomic control
  2. Adjust for population principal components
  3. Variance component model for family-based association test
Genomic control

No stratification

Test locus

Unlinked ‘null’ markers

Stratification → adjust test statistic

(Figure courtesy Shaun Purcell, Harvard, and Pak Sham, HKU)
Genomic inflation factor

• Compute $\chi^2$ statistic for each marker

• Genomic inflation factor ($\lambda$)

$$\lambda = \frac{\text{Median Observed } \chi^2}{\text{Median Expected } \chi^2}$$

– Median expected $\chi^2 = 0.456$
  • Why use median vs. mean?

• Adjust statistic at candidate markers

  – Replace $\chi^2_{\text{biased}}$ with $\chi^2_{\text{fair}} = \chi^2_{\text{biased}}/\lambda$

  – Should be $\lambda \geq 1$
    • Why?

(Devlin & Roeder, *Biometrics*, 1999)
QQ plots: a useful diagnostic

- **Data:** WTCCC Study
- **Phenotype:** T2D status
- **Genotypes:** imputed using GoT2D reference
- **Analysis:** logistic regression

- Classify SNPs as within or outside Known (+/-1Mb) T2D loci
- For all SNPs, $\lambda = 1.095$
  - Some population stratification
- For Known SNPs, $\lambda = 1.127$
  - Very inflated, but under alternative hypothesis
## Genomic control example

<table>
<thead>
<tr>
<th>Y</th>
<th>g₁</th>
<th>g₂</th>
<th>g₃</th>
<th>g₄</th>
<th>...</th>
<th>gₘ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Uncorrected Test Statistics

\[
\chi^2_1, \chi^2_2, \chi^2_3, \chi^2_4, \chi^2_m
\]
Genomic control example

\[
\begin{align*}
\lambda &= \frac{\text{Median Observed } \chi^2}{\text{Median Expected } \chi^2} \\
\text{Calculate Genomic Inflation Factor}
\end{align*}
\]
Genomic control example

<table>
<thead>
<tr>
<th>Y</th>
<th>g₁</th>
<th>g₂</th>
<th>g₃</th>
<th>g₄</th>
<th>...</th>
<th>gₐ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

\[ \chi^2_{1, \text{fair}} = \frac{\chi^2_1}{\lambda} \]

\[ \chi^2_{m, \text{fair}} \]

If \( \lambda \geq 1 \)

Calculate Genomic Inflation Factor

\[ \lambda = \frac{\text{Median Observed } \chi^2}{\text{Median Expected } \chi^2} \]
Principal components analysis (PCA)

• Use PCA to determine “axes of genotype variation” for a selected set of genotypes
  – Principal components mirror European geography
• Include PC’s as covariates in regression model to adjust for stratification

(Figure from Novembre et. al., Nature, 2008)

(Price et. al., Nat. Genet., 2006)
Correcting for population structure using principal components

(Kang et. al., *Nat. Genet.*, 2010)
Variance component model for family-based association test

- Population-based analysis assumes uncorrelated phenotypes between individuals under the null

\[ y \sim \mathcal{N}(X\beta, \sigma^2 I) \]
Variance component model for family-based association test

- Population-based analysis assumes uncorrelated phenotypes between individuals under the null
  \[ y \sim \mathcal{N}(X\beta, \sigma^2 I) \]

- Family-based analysis assumes phenotypes are correlated with relatives’ phenotypes
  \[ y \sim \mathcal{N}(X\beta, \sigma^2_g K + \sigma^2_e I) \quad K_{ij} : \text{kinship coefficient} \]
Variance component model for family-based association test

- Population-based analysis assumes uncorrelated phenotypes between individuals under the null
  \[ y \sim \mathcal{N}(X\beta, \sigma^2 I) \]

- Family-based analysis assumes phenotypes are correlated with relatives’ phenotypes
  \[ y \sim \mathcal{N}(X\beta, \sigma_g^2 K + \sigma_e^2 I) \]
  \[ K_{ij} : \text{kinship coefficient} \]

- Similar model for population-based analysis to account for distant relationship inferred from dense SNP arrays
  \[ y \sim \mathcal{N}(X\beta, \sigma_g^2 \hat{K} + \sigma_e^2 I) \]
  \[ \hat{K}_{ij} : \text{marker-based kinship coeff.} \]

Genome-wide association of human height

- NFBC 1966 birth cohort
- Illumina 370,000 SNPs
- 5,326 unrelated individuals
Uncorrected analysis
- Overdispersion of test statistics -

\[ \lambda_{GC} = \frac{\text{median}\{T_1, T_2, \ldots, T_n\}}{\mathbb{E}[\text{median}\{T\}]} \]

Conditioning on principal components
- Overdispersion still exists -

\[ y = \mu + x\beta + G\gamma + e \]

- G is top k (=100) eigenvectors of kinship matrix K
- \( \lambda_{GC} \) from 1.187 to 1.074
- \( \lambda_{GC} \) is still substantially higher than expected
- Corrects for population structure, but not hidden relatedness

Variance component model
- Overdispersion resolved -

\[ y = \mu + x\beta + u + e \]

\[ \text{Var}(u) = \sigma_g^2 K \]

\[ \text{Var}(e) = \sigma_e^2 I \]

- Using EMMAX reduced \( \lambda_{GC} \) from 1.187 to 1.003
- \( \lambda_{GC} \) falls into 95% confidence intervals

uncorrected \( \lambda_{GC} = 1.187 \)
100 PCs \( \lambda_{GC} = 1.074 \)
EMMAX \( \lambda_{GC} = 1.003 \)

95% CI: 0.992 ~ 1.008
Multiple genetic association studies

• Most associated common variants have small effect sizes (e.g. odds ratios [OR] < 1.2)

• To increase power to detect small genetic effect sizes, combine information across studies using
  – Meta-analysis of study-level association results
  – Joint analysis of all individual-level data
Multiple studies: Data aggregation methods

Joint analysis

- Combine individual-level data and analyze jointly

\[ \hat{\beta}_{joint}, \hat{SE}_{joint}, \text{P-value}_{joint} \]
Multiple studies: Data aggregation methods

**Joint analysis**

- Study 1
- Study 2
- ... Studyn
- All Data

\[
\hat{\beta}_{\text{joint}}, \hat{SE}_{\text{joint}}, P-\text{value}_{\text{joint}}
\]

**Meta-analysis**

- Study 1
- Study 2
- ... Studyn

\[
\hat{\beta}_1, \hat{SE}_1, P-\text{value}_1
\]
\[
\hat{\beta}_2, \hat{SE}_2, P-\text{value}_2
\]
\[
\ldots
\]
\[
\hat{\beta}_n, \hat{SE}_n, P-\text{value}_n
\]

\[
\hat{\beta}_{\text{meta}}, \hat{SE}_{\text{meta}}, P-\text{value}_{\text{meta}}
\]

- Combine study-level association results using:
  - Inverse-variance weights
  - Sample-size weights
Joint vs. meta-analysis

• For common variants, both joint and meta-analysis are both well-calibrated, and have near-equivalent power

• Meta-analysis is more commonly used
  – Sharing individual-level data is difficult due to logistical and ethical restrictions

• Combining multiple studies is critical to increase power to detect small effect sizes

(Lin & Zeng, Genet. Epidemiol., 2010)
Summary: analysis of common variants

• Single variant analysis with regression-based methods have identified many trait-associated genetic variants

• Important to account for population structure and/or sample relatedness to avoid spurious association
ANALYSIS OF LOW-FREQUENCY AND RARE VARIANTS
Genetic architecture of complex traits

Population sequencing, Dense reference imputation into GWAS, Specialized array genotyping

Rare alleles causing Mendelian disease

Low-frequency variants with intermediate effect

Common variants implicated in common disease by GWA

Few examples of high-effect common variants influencing common disease

Rare variants of small effect very hard to identify by genetic means

Very rare

Rare

Low frequency

Common

Allele frequency

Effect size

High

Intermediate

Modest

Low

50.0

3.0

1.5

1.1

0.001

0.005

0.05
Why study rare variants?

COMPLETE GENETIC ARCHITECTURE OF EACH TRAIT

• Are there additional susceptibility loci to be found?
• What is the contribution of each identified locus to a trait?
  – Sequencing, imputation and new arrays describe variation more fully
  – Rare variants are plentiful and should identify new susceptibility loci

UNDERSTAND FUNCTION LINKING EACH LOCUS TO A TRAIT

• Do we have new targets for therapy?
  What happens in gene knockouts?
  – Use sequencing to find rare human “knockout” alleles
  – Good: Results may be more clear than for animal studies
  – Bad: Naturally occurring knockout alleles are extremely rare
Why study rare variants?

COMPLETE GENETIC ARCHITECTURE OF EACH TRAIT

• Are there additional susceptibility loci to be found?
• What is the contribution of each identified locus to a trait?
  – Sequencing, imputation and new arrays describe variation more fully
  – Rare variants are plentiful and should identify new susceptibility loci

UNDERSTAND FUNCTION LINKING EACH LOCUS TO A TRAIT

• Do we have new targets for therapy?
  What happens in gene knockouts?
  – Use sequencing to find rare human “knockout” alleles
  – Good: Results may be more clear than for animal studies
  – Bad: Naturally occurring knockout alleles are extremely rare

Coding Variants Especially Useful!
Lots of rare functional variants to discover

<table>
<thead>
<tr>
<th>SET</th>
<th># SNPs</th>
<th>Singletons</th>
<th>Doubletons</th>
<th>Tripletons</th>
<th>&gt;3 Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>270,263</td>
<td>128,319</td>
<td>29,340</td>
<td>13,129</td>
<td>99,475</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(47%)</td>
<td>(11%)</td>
<td>(5%)</td>
<td>(37%)</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>410,956</td>
<td>234,633</td>
<td>46,740</td>
<td>19,274</td>
<td>110,309</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(57%)</td>
<td>(11%)</td>
<td>(5%)</td>
<td>(27%)</td>
</tr>
<tr>
<td>Nonsense</td>
<td>8,913</td>
<td>6,196</td>
<td>926</td>
<td>326</td>
<td>1,465</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(70%)</td>
<td>(10%)</td>
<td>(4%)</td>
<td>(16%)</td>
</tr>
<tr>
<td>Non-Syn / Syn Ratio</td>
<td>1.8 to 1</td>
<td>1.6 to 1</td>
<td>1.4 to 1</td>
<td>1.1 to 1</td>
<td></td>
</tr>
</tbody>
</table>

There is a very large reservoir of extremely rare, likely functional, coding variants.

NHLBI Exome Sequencing Project
Challenges for association testing of low-frequency variants

• Low minor allele count (MAC)

• Stringent $\alpha = 5 \times 10^{-8}$ (multiple testing)

• For binary traits:
  – Unbalanced numbers of cases and controls (e.g. population-based studies)
Logistic Wald test has low power* for low-frequency and rare variants in balanced studies

Recently noted by Xing et al. (2012) *Ann Hum Genet* 76:168-77
Logistic score test is anti-conservative for low-frequency and rare variants in unbalanced studies

QQ Plot
80 Cases / 1,768 Controls

MAC < 20

20 ≤ MAC < 200
Recommended single marker tests for low-frequency variants

**Binary Traits**

- For balanced studies (case-control ratio < 3:2)
  - Use Firth bias-corrected*, or score logistic regression
  - Avoid Wald test (low power)
- For unbalanced studies (case-control ratio > 3:2)
  - Use Firth, likelihood ratio logistic regression
  - Avoid score test (inflated false positive rate)

**Quantitative Traits**

- Given normally-distributed QTs
  - Use any linear regression test

*(Firth, *Biometrika*, 1993)

(Ma et. al. *Genet. Epidemiol.*, 2013; Ma et. al., *in preparation*)
Limitations of single marker tests

• Single marker tests have low power for rare variants unless sample size very large

• For binary traits, variants require minimum MAC ≥ 26 to have p-values < $5 \times 10^{-8}$:
(No covariates; $N_{cases} = N_{ctrls}$)

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Cases</th>
<th>Ctrls</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>975</td>
<td>1000</td>
</tr>
<tr>
<td>Aa</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

Fisher’s Exact Test $p = 5.1 \times 10^{-8}$

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Cases</th>
<th>Ctrls</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>974</td>
<td>1000</td>
</tr>
<tr>
<td>Aa</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

Fisher’s Exact Test $p = 2.5 \times 10^{-8}$
Gene-based tests

- Gene-based tests jointly analyze multiple rare variants in genetic region (e.g. gene)
- Increases power by:
  - Combining information across rare variants
  - Requiring less stringent $\alpha$, e.g. $\alpha = 2.5 \times 10^{-6}$ for 20K genes

**Burden Test** (with weights $w_j$)

(Madsen & Browning, *PLoS Genet.*, 2009)
Selecting variants for gene-based tests

- If include variants of all frequencies, non-causal and common variants will dilute signal
- Commonly used filters or “masks”:
  - Include variants MAF ≤ 0.05 or 0.01
  - Weight variants by MAF
    - E.g. $w_j \sim \text{Beta}(\text{MAF}, 1, 25)$
  - Select variants based on functional annotation:
    - E.g. Protein Truncating Variants only, nonsynonymous, missense, etc.
- If mask is too restrictive, will reduce to single variant test, and no gain in power
Categories of aggregation tests

• **Burden tests** test association between (weighted) sum of rare alleles with disease or QT
  – CMC (Li & Leal, 2008), WSS (Madsen & Browning, 2009)

• **Dispersion tests** measure deviations from expected distribution
  – SKAT (Wu et al., 2011), C-alpha (Neale et al., 2011)

• **Combined tests** combine strengths of burden and dispersion tests
  – SKAT-O (Lee et al., 2012)
Power of gene-based tests

• Power of gene-based tests affected by the underlying (unknown) genetic architecture of the analyzed region:
  – Number of associated variants in region
  – Number of neutral variants diluting signals
  – Whether direction of effect is consistent within gene
Power comparison
(All causal variants 100% deleterious)

10% Variants in Region are Causal

50% Variants in Region are Causal

Burden is most powerful when there are many causal variants with same direction of effect

(Ma et al., in preparation)
Power comparison
(Causal variants are 50% deleterious / 50% protective)

10% Variants in Region are Causal

50% Variants in Region are Causal

SKAT is most powerful when there are causal variants with opposite direction of effects

(Ma et al., *in preparation*)
Power comparison

(Causal variants are 50% deleterious / 50% protective)

10% Variants in Region are Causal

50% Variants in Region are Causal

SKAT-O is generally powerful and robust for different genetic architectures

(Ma et al., in preparation)
Summary: analysis of low-frequency variants

• Single marker tests remain useful for low-frequency variants
  – Need to carefully select well-calibrated tests

• Gene-based tests can be more powerful for rare variants
  – Power generally determined by underlying genetic architecture