Genomewide Analyses

Goncalo Abecasis
University of Michigan School of Public Health

Human Genetics This Century ...

• Very rapid technology improvements ...
  • Each year, 4x more data than year before

• Many discoveries, hundreds of genetic variants for many traits
  • Also, some completely new ideas about to make genetic studies work

• Many challenges, mechanism connecting genes and disease rarely understood
  • Focus is shifting from discovering association to methods that help us understand it

• New models for data access, new strategies for collecting data at scale
  • Still, many bits of relevant information are very hard to get

Human Genetics, Sample Sizes over My Time

<table>
<thead>
<tr>
<th>Year</th>
<th>No. of Samples</th>
<th>No. of Markers</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ongoing</td>
<td>~55,000</td>
<td>450 million</td>
<td>NHLBI Precision Medicine Cohorts / TopMed</td>
</tr>
<tr>
<td>Ongoing</td>
<td>~31,000</td>
<td>50 million</td>
<td>HapMap Reference Consortium</td>
</tr>
<tr>
<td>2012</td>
<td>1,002</td>
<td>40 million</td>
<td>The 1000 Genomes Project (Nature)</td>
</tr>
<tr>
<td>2010</td>
<td>Hundreds</td>
<td>16 million</td>
<td>The 1000 Genomes Project (Nature)</td>
</tr>
<tr>
<td>2010</td>
<td>~100,000</td>
<td>2.5 million</td>
<td>Lipid GWAS (Nature)</td>
</tr>
<tr>
<td>2008</td>
<td>~5,000</td>
<td>2.5 million</td>
<td>Lipid GWAS (Nature Genetics)</td>
</tr>
<tr>
<td>2007</td>
<td>Hundreds</td>
<td>5.1 million</td>
<td>HapMap (Nature)</td>
</tr>
<tr>
<td>2005</td>
<td>Hundreds</td>
<td>1 million</td>
<td>HapMap (Nature)</td>
</tr>
<tr>
<td>2003</td>
<td>Hundreds</td>
<td>10,000</td>
<td>Chr 19 Variation Map (Nature Genetics)</td>
</tr>
<tr>
<td>2002</td>
<td>Hundreds</td>
<td>1,500</td>
<td>Chr 22 Variation Map (Nature)</td>
</tr>
<tr>
<td>2001</td>
<td>Thousands</td>
<td>127</td>
<td>Three Region Variation Map (Am J Hum Genet)</td>
</tr>
<tr>
<td>2000</td>
<td>Hundreds</td>
<td>26</td>
<td>T-cell receptor variation (Hum Mol Genet)</td>
</tr>
</tbody>
</table>

"... of the 166 associations which have been studied 3 or more times, only six have been consistently replicated.“
Hirschhorn et al (2002)

A Genomewide Study of Obesity

Basic Strategy ...

• In individuals who are generally similar, check whether ...
  • ... a particular genotype is associated with individual outcomes.

• In two otherwise similar persons, does...
  • FTO genotype predict body-weight?
  • PCSK9 genotype predict lipid levels and heart disease risk?
  • TCF7L2 genotype predict type 2 diabetes status?

• For quantitative traits, typical to study populations or extreme individuals.
• For disease outcomes, most studies rely on case-control collections.
Typical Quality Control Steps

• What are important quality control steps in a GWAS?

Typical Quality Checks in GWAS

• Assessment of genotyping accuracy
  • Examine individuals genotyped in duplicate
  • Check Mendelian segregation if related individuals available

• Genotyping completeness check
  • Is the amount of missing data small?
  • Are missing data rates similar in cases and controls?
  • Data rarely missing at random, too much missing data leads to spurious signals.

• Check for Hardy-Weinberg Equilibrium
  • Do homozygotes and heterozygotes appear in expected proportions?

• Check for related individuals
  • Verify relationships between individuals match expectations
  • Verify X/Y genotypes match expectations
  • Q-Q plots

Typical GWAS Analyses

• Initial round of association analyses
  • Compare allele frequencies in cases and controls
  • Compare trait values by genotype

• Verify that test statistics are well behaved
  • Most markers should show no evidence for association

• Follow-up promising findings
  • Genotype additional sets of individuals to confirm promising signals from first round analysis

Q-Q Plots: A Useful Diagnostic

In genomewide studies, most markers show no association with the trait and, therefore, very similar observed and expected p-values.
Willer et al, Nature Genetics, 2008

In genomewide studies, only a small subset of markers is expected to show association with any particular trait.
Willer et al, Nature Genetics, 2008
Q-Q Plots: A Useful Diagnostic

![Q-Q Plot](image)

The genomic control value examines markers with little evidence for association. If these large p-values were to deviate from expected, there is a problem! In this case, λ=1.02.

Willer et al, Nature Genetics, 2008

When Q-Q Plots Identify Problems...

- Check for quality metrics for poor performing samples.
- Samples with low call rates, unusual heterozygosity, etc.
- As a rule, avoid mixing samples genotyped on different platforms.
- Perhaps the study includes samples with varying genetic backgrounds?
  - Try principal component analysis to characterize the ancestry of each sample.
- Perhaps the study includes related samples?
  - Try linear mixed models to account for similarity between samples.

The Stratification Problem

- If phenotypes differ between populations
- And allele frequencies have drifted apart
- Unlinked markers exhibit association
- Not very useful for gene mapping!

Genomic Control

![Genomic Control Diagram](image)

Genomic Control

- Compute chi-squared for each marker
- Inflation factor λ
  - Average observed chi-squared
  - Median observed chi-squared / 0.456
  - Should be >= 1
- Adjust statistic at candidate markers
  - Replace $\chi^2$biased with $\chi^2$fair = $\chi^2$biased/λ.

Questions

- When defining the inflation factor $\lambda$ ...
- Why do we use a lower bound of 1?
- What might be the advantages of using the median rather than the mean?
Applying Genomic Control

• Simple and convenient approach...
  • Easily adapted to other test statistics, such as those for quantitative trait and haplotype tests

• Under the null, stratification always inflates evidence for association...
  • Is this also true under the alternative?
  • What might be the consequences?

Structured Association

• Use unlinked markers to assign individuals to subpopulations then...
  • Test for association within each subpopulation
  • Test for association while conditioning on subpopulation

Some Attractive Features

• Allows for flexibility in association test

• Describing subpopulations can be useful

• Does not assume constant population differentiation across the genome

Principal Components For a GWAS

Age-Related Macular Degeneration

• Common cause of blindness among the elderly

• Affects >2 million individuals in the United States

• Prevalence increases with old age:
  • ~4% at age 75
  • ~12% at age 80
Side Note from QC: Age-dependent Y-Chromosome Loss


**Macular Degeneration, Comparison of Case and Control Genomes**

**Comparison around VEGFA gene**

**Comparison in a region of chromosome 22**

**TIMP3 Variants on AMD Array**

8 of 10 Sorby's mutations cause unpaired cysteine residues
- Polymerization of TIMP3 protein
- Accumulation in extracellular matrix

**Rare TIMP3 variants and AMD**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Allele Count</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ser38Cys</td>
<td>14</td>
<td>Cystine</td>
</tr>
<tr>
<td>Gly58Cys</td>
<td>1</td>
<td>Disrupting</td>
</tr>
<tr>
<td>Tyr109Cys</td>
<td>1</td>
<td>Variant</td>
</tr>
<tr>
<td>Arg132Cys</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Gly173Cys</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Glu162Cys</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>His181Arg</td>
<td>5</td>
<td>Reported</td>
</tr>
<tr>
<td>Ser204Cys</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

OR = 30  
\(p = 10^{-6}\)
Carrier Status and Age of Onset

<table>
<thead>
<tr>
<th>Advanced AMD Cases</th>
<th>Age of Onset (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrier (N = 29)</td>
<td>Average</td>
</tr>
<tr>
<td></td>
<td>65</td>
</tr>
<tr>
<td>Non-carriers (N = 16,115)</td>
<td>77</td>
</tr>
</tbody>
</table>

→ Carriers statistically significant younger: $p_{diff} = 4 \times 10^{-7}$

Major AMD Risk Variants in TIMP3 carriers ...

Different AMD risk profile than
• Advanced AMD cases without TIMP3 mutations
• Controls

Comparison in a region of chromosome 22

Useful Resources to Know ...

• LocusZoom
  - http://locuszoom.sph.umich.edu/
  - http://portaldev.sph.umich.edu/lzplug/demo.html

• PLINK
  - http://pngu.mgh.harvard.edu/~purcell/plink

How to Run Large Genetic Studies?
• Exploring new ways to engage populations in research
• Continuous Engagement, Web, Mobile Devices
• Currently >8,500 participants

Details ...