Genetics and Genomics

3. Association Studies and Gene Expression

ggibson.gt@gmail.com
http://www.cig.gatech.edu

Outline

- General overview of association studies (GWAS)
- Sample Results
- Gene Expression Profiling
Principle of Association Studies

<table>
<thead>
<tr>
<th>Individual</th>
<th>Site</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A T C C G A</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>A C T C G A</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>A C C A - G</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>T T C A G A</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>A T C A G A</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>A C C C - G</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>T C T A - G</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>A T C C G A</td>
<td>8</td>
</tr>
</tbody>
</table>

Are the phenotype scores associated with each class of SNP drawn from the same or different distributions?

Linkage versus Association

Linkage examines recent recombination events in a pedigree:
- over just several generations
- large chromosomal regions detected
- no information on allele frequency

Association examines historical recombination events in a population:
- basically a 10,000 generation pedigree
- resolution to single genes
- estimates effect size and frequency

Why LD happens

When a mutation occurs, by definition it is only on one chromosome and hence “associated” with the genotypes elsewhere on that chromosome.

Over time, the mutation increases in frequency and becomes a polymorphism. It remains in LD with the genotypes on the chromosome it appeared on.

Eventually recombination breaks up the LD, in proportion to genetic distance.

Measurement of LD

LD is the non-random association of genotypes.

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT</td>
<td>24</td>
<td>48</td>
<td>24</td>
</tr>
<tr>
<td>TC</td>
<td>48</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>CC</td>
<td>24</td>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT</td>
<td>24</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TC</td>
<td>48</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>CC</td>
<td>24</td>
<td>0</td>
<td>24</td>
</tr>
</tbody>
</table>
Haplotypes and Tagging SNPs

LPL example
Case-Control and TDT designs

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Allele M</td>
<td>Allele m</td>
</tr>
<tr>
<td>Affected</td>
<td>34</td>
<td>278</td>
</tr>
<tr>
<td>Unaffected</td>
<td>69</td>
<td>256</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>14.0</td>
<td></td>
</tr>
<tr>
<td>$P$</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

Transmitted Allele

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Allele M</td>
<td>Allele m</td>
</tr>
<tr>
<td>M</td>
<td>78</td>
<td>46</td>
</tr>
<tr>
<td>m</td>
<td>62</td>
<td>62</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>8.2</td>
<td></td>
</tr>
<tr>
<td>$P$</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>
Repeatability and Forest Plots

Population Structure

<table>
<thead>
<tr>
<th></th>
<th>Blue subpopulation</th>
<th>Rod subpopulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA</td>
<td>AG</td>
</tr>
<tr>
<td>Case</td>
<td>80</td>
<td>640</td>
</tr>
<tr>
<td>Control</td>
<td>800</td>
<td>6400</td>
</tr>
<tr>
<td>Case/control</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Total population

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>260</td>
<td>1040</td>
<td>1480</td>
</tr>
<tr>
<td>Control</td>
<td>1800</td>
<td>8400</td>
<td>13,800</td>
</tr>
<tr>
<td>Case/control</td>
<td>0.155</td>
<td>0.124</td>
<td>0.107</td>
</tr>
</tbody>
</table>

Odds ratio (AG) = 1.2  \( p = 10^{-8} \)
GWAS first appeared 10 years ago, now several new diseases each month.

Inflammatory diseases show multiple associations, with some common variants (notably the MHC).

Depression and Hypertension show nothing: likely no variants with a relative risk greater than 1.5.

The Genetics of 7 Diseases

700 loci clearly influence height in combined analysis of 250,000 people.

Diverse roles:
- Hedgehog signalling
- Chromatin structure
- Cell cycle regulation
- Extracellular matrix deposition

Possible contribution of some of these loci to osteoarthritis, cancer, athleticism.

Half the genes have at least two independent associations.

But... they only explain 15% of the variation for height (one fifth of the heritability).
Heritability of obesity ~ 60%

2/3 Americans BMI > 25

One gene, FTO, is repeatedly associated with BMI, hip circumference and weight, in most human populations

Homozygote classes differ in weight by up to 2 kg

Study of 230,000 people → 49 loci for WHR, many linked to adipose, insulin biology
20 loci only in women

Study of 340,000 people → 97 loci for BMI, many linked to neuronal function
Little overlap with WHR

Inflammatory Bowel Disease affects ~1% of adults either as Crohn’s or Ulcerative Colitis

Two genes, IL23R and NOD2, each explain 1% of the variance

Another 160 genes each explain less than 0.25% of it

Strongly overlap with PID and other autoimmune diseases

Vast majority are in immune function genes, and ongoing efforts suggest activity in specific immune cell types
The Genetics of Type 2 Diabetes

54 established Fasting Glucose or Fasting Insulin loci tend to have pancreatic islet cell functions
23 replicate in African Americans in Trans-ethnic analyses

TCF7L2 is the strongest risk locus for T2D in Caucasians. The ancestral allele is the risk allele: it is found in 90% of Africans, 40% of Europeans, and just 5% of Asians

Considering the top 18 loci:
1% of CAU have >24 alleles
2% of CAU have <12 alleles
They differ 4-fold in odds of T2D

The Genetics of Schizophrenia

128 independent SNP associations from GWAS of 37,000 cases

Strong enrichment in genes expressed in certain neuronal cell types or implicated in synaptic transmission

But at least 5% of cases attributable to CNV: copy number variation

3 major chromosomal deletions of >100kb at frequency <1% are almost exclusively found in schizophrenics
23andme studies

Other interesting traits:

- Endurance Runner vs Sprinter
  (30% of people change their answer if they know their ACTN3)
- Left vs Right Handedness
  (nothing striking)
- Have you ever needed braces or wisdom teeth surgery?
- Breast size (finds breast cancer risk loci)
- Hand-clasp dominance ...


Some references

- The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-678
The Rationale for Gene Expression Profiling

Question: What makes a muscle cell different from a skin cell different from a liver cell?

Analogy: What makes a living room different from a bathroom different from the kitchen?

Answer: What furniture and appliances and windows are placed where and when
What genes are expressed where and when and how much

Strategy: Measure the abundance of mRNA transcripts in a bunch of samples, and use statistically rigorous approaches to identify differential expression

Methods: Prior to 1995: One gene at a time qPCR
1995-2015: Microarrays
Since 2015: RNAseq

Annotation of Gene Function

Gene Ontology

Expression in a tissue generally implies the gene is active there

Co-expression with similar types of gene may imply “guilt by association”

FISH and sub-cellular imaging of proteins tells us where they act
Differential Expression Analysis

Cancer

Inflammatory Bowel Disease


Clinical Diagnostics

Population Profiling

Gene Regulation
Expression QTL

Expression QTL are polymorphisms that affect the expression of a gene. They are particularly interesting when they overlap with GWAS hits.

Single Cell RNASeq

Peripheral Blood Monocytes

Neuronal nuclei

Additional Tiers of Gene Regulation

Li et al. (2016) Science 352: 600-604

Systems Biology