Understanding genetic association studies

Peter Kamerman
Outline

CONCEPTS UNDERLYING GENETIC ASSOCIATION STUDIES

Genetic concepts:
- Underlying principals
- Genetic variants
- Linkage disequilibrium
- Population genetics and stratification

Study design:
- Types of studies
- Assessing a genetic association study
  - Quality control
  - Analysis and statistical power
What is the purpose of a genetic association study?
Genetic variation

Mutation occurs

Phenotypic variation

Phenotype frequency

Pain sensitivity
Phenotype variability

Genetic variability

Phenotype variability

Genetic variability
Genetic contribution to extreme phenotypes

Rare, catastrophic monogenic mutations

Insensitivity

Hyper-sensitivity

Phenotype frequency

Pain sensitivity
Genetic contribution to extreme phenotypes

Rare, catastrophic monogenic mutations

Insensitivity

Hyper-sensitivity

Phenotype frequency

Pain sensitivity

GENETIC LINKAGE ANALYSIS
Genetic contribution to “non-extreme” phenotypes

Complex phenotype = multigenic

Phenotype frequency

Pain sensitivity
Genetic variation forms the basis of genetic epidemiology.

Complex phenotype = multigenic

Genetic association analysis

- Family-based
- Population-based

Phenotype frequency

Pain sensitivity
Genetic contribution to “non-extreme” phenotypes

A few loci of moderate effect and intermediate (5-20%) to high (>20%) minor allele frequency, each explaining several percent of disease risk in a population

Gibson, 2012
Genetic contribution to “non-extreme” phenotypes

Many common alleles of intermediate (5-20%) to low (1-5%) frequency with moderate to low effect, respectively (infinitesimal model)

Gibson, 2012; Park et al., 2011
Genetic contribution to “non-extreme” phenotypes

A few rare alleles (<1%) with moderate to high effects (GRR > 2) (rare allele model)

Gibson, 2012
Allele frequency vs. effect size

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

Manolio et al., 2009
Genetic variants
Genetic variation

*Single-nucleotide polymorphisms (SNPs)*

Mutation occurs (single-nucleotide substitution)

New allele

Ancestral

Human genome: ~3 billion base-pairs (haploid)

10-30 million SNPs in the human population

~0.1% variability in SNPs between unrelated individuals (~3 million SNPs)
Only 1-5% of SNPs are functional (causative) mutations

Non-synonymous SNP = missense or nonsense
Promotor region SNPs

The rest are non-functional mutations
Synonymous SNP = no amino acid change
Single nucleotide polymorphisms

Mutation occurs (single-nucleotide substitution)

Ancestral

New allele

...So

Functional SNPs are rare
Single nucleotide polymorphisms

Mutations occur (single-nucleotide substitution)

C -> A

New allele

C

Ancestral

...So

Functional SNPs are rare

...and are hard to identify
Single nucleotide polymorphisms

Mutation occurs (single-nucleotide substitution)

...AND

Non-functional SNPs are abundant
Non-functional SNPs are abundant

...but are they useful?
Linkage disequilibrium
Linkage disequilibrium

Non-random association between alleles at two or more loci

Can use the location of one SNP to “TAG” the presence of another or many other SNPs
Linkage disequilibrium
Linkage disequilibrium

- Locus A far from locus M/N
- Locus B close to locus M/N

Recombination
Linkage disequilibrium

locus A far from locus M/N

locus B close to locus M/N

many generations

linkage equilibrium with alleles at locus A

linkage disequilibrium with B1 allele at locus B

recombination
Correlations decay with number of recombination events:

- Distance
- Population history
Linkage disequilibrium (LD) plots

- Shows correlations between SNPs
  - High LD ($r^2$)
  - Low LD ($r^2$)

Slide courtesy of Chris Spencer, Wellcome Trust Centre for Human Genetics
Population genetics
Population-based genetic variation
Population-based genetic variation

Founder effect

Tishkoff and Williams, 2002
Population-based genetic variation

Adapted from McVean et al., 2005
GTP cyclohydrolase gene (GCH1) tagSNPs

- 25 SNPs MAF > 0.1
- 8 tagSNPs ($r^2 > 0.8$)
GTP cyclohydrolase gene (GCH1) tagSNPs

YRI
- 32 SNPs MAF > 0.1
- 19 tagSNPs ($r^2 > 0.8$)
Population-based genetic variation

High LD: European

Low LD: African

Palmer and Cardon, 2005
Population-based genetic variation

Palmer and Cardon, 2005
Population-based genetic variation

Palmer and Cardon, 2005
Population stratification

Population A

Cases

Controls

Population B

Cases

Controls

Genotype

aa
Aa
AA

(p = 0.34)

(p < 0.001)

(p = 0.46)

Slide courtesy of Dominic Kwiatkowski, Wellcome Sanger Institute
Population stratification

*Types of population structure*

**Discrete:** Populations that are remotely related (e.g., Europeans, Africans and Asians).

**Admixed:** Admixed ancestry (e.g., African Americans, Cape Coloureds) with different individuals having different degrees of admixture.

**Hierarchical:** Both discrete and admixed population structures. Hierarchical population structures may be seen in studies that involve multi-ethnic cohorts.
Population stratification

*Primary methods for correcting for population structure*

- Self-reported ethnicity
- Genomic control
- Principal components analysis
Population stratification

Assessing for population structure

Q-Q plots

a No stratification

b Stratification without unusually differentiated markers

Price et al., 2010
Population stratification

Self-reported ethnicity

Teo et al., 2010
Population stratification

**Genomic control**

Apply a correction factor ($\lambda_{GC}$) to correct for the difference between observed and expected Chi-square statistic.

Let down by:
- Not conservative enough when few markers used (↑false positives)
- Too conservative if many markers used (↑false negatives)
- Cannot deal with large deviations

Price et al., 2010
Population stratification

**PCA**

PCA transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables (*principal components*).

The PCA method identifies principal components that represent the population structure based on genetic correlations among individuals.

**Notes:**
- Need to prune regions with long-range high LD
- Need to remove related individuals

Teo et al., 2010
Study design
Types of genetic studies for complex traits

*Family-based vs. Population-based*
## Types of genetic studies for complex traits

### Population-based designs

<table>
<thead>
<tr>
<th>Design</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-sectional</td>
<td>Genotype and phenotype (i.e., note disease status or quantitative trait value) a random sample from population</td>
</tr>
<tr>
<td>Cohort</td>
<td>Genotype subsection of population and follow disease incidence for specified time period</td>
</tr>
<tr>
<td>Case-control</td>
<td>Genotype specified number of affected (case) and unaffected (control) individuals. Cases usually obtained from family practitioners or disease registries, controls obtained from random population sample or convenience sample</td>
</tr>
<tr>
<td>Extreme values</td>
<td>Genotype individuals with extreme (high or low) values of a quantitative trait, as established from initial cross-sectional or cohort sample</td>
</tr>
</tbody>
</table>
Types of genetic studies for complex traits

Family-based designs

Transmission disequilibrium test

Untransmitted allele

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>B</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Transmitted allele

Test statistic

\[ T = \frac{(b-c)^2}{b+c} \sim \chi^2_1 \]

Case-parent triads
Genotype affected individuals plus their parents
(affected individuals determined from initial cross-sectional, cohort, or disease-outcome based sample)

Case-parent-grandparent septets
Genotype affected individuals plus their parents and grandparents

Cordell and Clayton, 2005
Types of genetic studies for complex traits

**Family-based vs. Population-based**

### Benefits of family-based

<table>
<thead>
<tr>
<th>Benefits of family-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protection against population substructure</td>
</tr>
<tr>
<td>Improved genotyping quality control</td>
</tr>
<tr>
<td>Allow estimates of parent-of-origin (imprinting) effects</td>
</tr>
<tr>
<td>Can assess whether an allele is inherited or <em>de novo</em></td>
</tr>
</tbody>
</table>

### Drawbacks of family-based designs

<table>
<thead>
<tr>
<th>Drawbacks of family-based designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult sample collection, especially for older onset traits (e.g., diabetes)</td>
</tr>
<tr>
<td>Less power than case-control studies</td>
</tr>
<tr>
<td>Requires more genotyping</td>
</tr>
<tr>
<td>Requires at least one parent to be heterozygote</td>
</tr>
</tbody>
</table>
### Candidate gene vs. Genome-wide

<table>
<thead>
<tr>
<th>Hypothesis driven</th>
<th>Hypothesis-free</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focuses on gene(s) identified through our understanding of a disease</td>
<td>Scans the whole genome for possible associations between polymorphisms and disease of interest</td>
</tr>
<tr>
<td>Efficient use of resources</td>
<td>Only assesses common genetic variants</td>
</tr>
<tr>
<td>Tailored marker selection</td>
<td>Can be semi-tailored</td>
</tr>
</tbody>
</table>
Types of genetic studies for complex traits

**Candidate gene vs. Genome-wide**

<table>
<thead>
<tr>
<th>Hypothesis driven</th>
<th>Hypothesis-free</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focuses on gene(s) identified through our understanding of a disease</td>
<td>Scans the whole genome for possible associations between polymorphisms and disease of interest</td>
</tr>
<tr>
<td>Efficient use of resources</td>
<td>Only assesses common genetic variants</td>
</tr>
<tr>
<td>Tailored marker selection</td>
<td>Can be semi-tailored</td>
</tr>
</tbody>
</table>

The Human660W-Quad Genomic Coverage

The Human660W-Quad BeadChip content covers the majority of common variation in three distinct populations. Graphs are estimated, based on the HapMap release 24 data set of > 2.3 million common SNPs.
Assessing genetic association studies
Assessing genetic association studies

1. How appropriate was the choice of candidate genes/markers?

2. What quality control measures were used?

3. Was the analysis appropriate
Assessing genetic association studies

How appropriate was the choice of candidate genes/markers?

How well do we know the mechanisms of a disease
- Evidence from animal models and in vitro studies
- Evidence from clinical studies
- Evidence from linkage studies

Markers selection
- “Functional” polymorphisms
- Evidence from the literature of previous association
- TagSNP selection criteria ($r^2 > 0.8$, appropriate population, MAF)
Assessing genetic association studies

What quality control measures were used?

Lab QC
- Assigning cases and controls to each plate
- Running duplicates (same plate and across plates)
Assessing genetic association studies

What quality control measures were used?

**QC per individual** (GWAS)
- Identification of discordant sex information
- Identification of poor DNA quality
- Identification of duplicated or related individuals
- Identification of individuals of divergent ancestry

**QC per marker**
- Identification of SNPs with excessive missing genotypes
- Identification of SNPs showing significant deviation from Hardy-Weinberg equilibrium (HWE)
- Identification of SNPs with different rates of missingness between cases and controls
- Removal of markers with very low minor MAF

Anderson et al., 2010
Assessing genetic association studies

What quality control measures were used?

**QC per individual (GWAS)**
- Identification of discordant sex information
- Identification of poor DNA quality
- Identification of duplicated or related individuals
- Identification of individuals of divergent ancestry

**QC per marker**
- Identification of SNPs with excessive missing genotypes
- Identification of SNPs showing significant deviation from Hardy-Weinberg equilibrium (HWE)
- Identification of SNPs with different rates of missingness between cases and controls
- Removal of markers with very low minor MAF

Anderson et al., 2010
Assessing genetic association studies

What quality control measures were used?

**QC per individual** (GWAS)
- Identification of discordant sex information
- *Identification of poor DNA quality*
- Identification of duplicated or related individuals
- Identification of individuals of divergent ancestry

**QC per marker**
- Identification of SNPs with excessive missing genotypes
- Identification of SNPs showing significant deviation from Hardy-Weinberg equilibrium (HWE)
- Identification of SNPs with different rates of missingness between cases and controls
- Removal of markers with very low minor MAF

Anderson et al., 2010
Assessing genetic association studies

What quality control measures were used?

**QC per individual** (GWAS)
- Identification of discordant sex information
- Identification of poor DNA quality
- Identification of duplicated or related individuals
- Identification of individuals of divergent ancestry

**QC per marker**
- Identification of SNPs with excessive missing genotypes
- Identification of SNPs showing significant deviation from Hardy-Weinberg equilibrium (HWE)
- Identification of SNPs with different rates of missingness between cases and controls
- Removal of markers with very low minor MAF

Anderson et al., 2010
Assessing genetic association studies

**What quality control measures were used?**

**QC per individual** (GWAS)
- Identification of discordant sex information
- Identification of poor DNA quality
- Identification of duplicated or related individuals
- Identification of individuals of divergent ancestry

**QC per marker**
- Identification of SNPs with excessive missing genotypes
- Identification of SNPs showing significant deviation from Hardy-Weinberg equilibrium (HWE)
- Identification of SNPs with different rates of missingness between cases and controls
- Removal of markers with very low minor MAF

Anderson et al., 2010
Assessing genetic association studies

*What quality control measures were used?*

**QC per individual (GWAS)**
- Identification of discordant sex information
- Identification of poor DNA quality
- Identification of duplicated or related individuals
- Identification of individuals of divergent ancestry

**QC per marker**
- *Identification of SNPs with excessive missing genotypes*
- Identification of SNPs showing significant deviation from Hardy-Weinberg equilibrium (HWE)
- Identification of SNPs with different rates of missingness between cases and controls
- Removal of markers with very low minor MAF

Anderson et al., 2010
Assessing genetic association studies

What quality control measures were used?

**QC per individual** (GWAS)
- Identification of discordant sex information
- Identification of poor DNA quality
- Identification of duplicated or related individuals
- Identification of individuals of divergent ancestry

**QC per marker**
- Identification of SNPs with excessive missing genotypes
- *Identification of SNPs showing significant deviation from Hardy-Weinberg equilibrium (HWE)*
- Identification of SNPs with different rates of missingness between cases and controls
- Removal of markers with very low minor MAF

Anderson et al., 2010
Assessing genetic association studies

Identification of SNPs showing significant deviation from Hardy Weinberg equilibrium (HWE)

Under HWE, alleles segregate randomly in the population, allowing expected genotype frequencies to be calculated from allele frequencies:

<table>
<thead>
<tr>
<th>Alleles:</th>
<th>G and T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes:</td>
<td>GG  GT  TT</td>
</tr>
<tr>
<td>Expected frequencies of genotypes:</td>
<td>G^2  2GT  T^2</td>
</tr>
</tbody>
</table>

Departures from HWE in control samples may be caused by the following:
1. Genotyping error. (calling heterozygotic individuals is challenging)
2. Population stratification
3. Selection, therefore assess HWE in controls only

Anderson et al., 2010
Assessing genetic association studies

What quality control measures were used?

**QC per individual** (GWAS)
- Identification of discordant sex information
- Identification of poor DNA quality
- Identification of duplicated or related individuals
- Identification of individuals of divergent ancestry

**QC per marker**
- Identification of SNPs with excessive missing genotypes
- Identification of SNPs showing significant deviation from Hardy-Weinberg equilibrium (HWE)
- *Identification of SNPs with different rates of missingness between cases and controls*
- Removal of markers with very low minor MAF

Anderson et al., 2010
Assessing genetic association studies

**What quality control measures were used?**

**QC per individual (GWAS)**
- Identification of discordant sex information
- Identification of poor DNA quality
- Identification of duplicated or related individuals
- Identification of individuals of divergent ancestry

**QC per marker**
- Identification of SNPs with excessive missing genotypes
- Identification of SNPs showing significant deviation from Hardy-Weinberg equilibrium (HWE)
- Identification of SNPs with different rates of missingness between cases and controls
- *Removal of markers with very low minor MAF*

Anderson et al., 2010
Assessing genetic association studies

Was the analysis appropriate?

Basic analysis of bi-allelic markers (SNPs)

Genotypic model

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>aA</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>c</td>
<td>e</td>
</tr>
<tr>
<td>Controls</td>
<td>b</td>
<td>d</td>
<td>f</td>
</tr>
</tbody>
</table>
Assessing genetic association studies

Was the analysis appropriate?

Allelic (multiplicative)

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>aA</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>c</td>
<td>e</td>
</tr>
<tr>
<td>Controls</td>
<td>b</td>
<td>d</td>
<td>f</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>2a+c</td>
<td>2e+c</td>
</tr>
<tr>
<td>Controls</td>
<td>2b+d</td>
<td>2F+d</td>
</tr>
</tbody>
</table>

![Graph showing AA, aA, and aa categories with data points.](attachment:image.png)
Assessing genetic association studies

Was the analysis appropriate?

Dominant model

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>aA</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>c</td>
<td>e</td>
</tr>
<tr>
<td>Controls</td>
<td>b</td>
<td>d</td>
<td>f</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>aA+aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>c+e</td>
</tr>
<tr>
<td>Controls</td>
<td>b</td>
<td>d+f</td>
</tr>
</tbody>
</table>
Assessing genetic association studies

Was the analysis appropriate?

Recessive model

<table>
<thead>
<tr>
<th>AA</th>
<th>aA</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>c</td>
<td>e</td>
</tr>
<tr>
<td>b</td>
<td>d</td>
<td>f</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AA+aA</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>a+c</td>
<td>e</td>
</tr>
<tr>
<td>b+d</td>
<td>F</td>
</tr>
</tbody>
</table>
Assessing genetic association studies

Was the analysis appropriate?

Cochrane-Armitage test for trend

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>aA</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>c</td>
<td>e</td>
</tr>
<tr>
<td>Controls</td>
<td>b</td>
<td>d</td>
<td>f</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>aA</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>c</td>
<td>e</td>
</tr>
<tr>
<td>Controls</td>
<td>b</td>
<td>d</td>
<td>f</td>
</tr>
</tbody>
</table>
Assessing genetic association studies

Was the analysis appropriate?

Did they correct for multiple comparisons?

Type I error is the probability of rejecting the null hypothesis when it is true.

The significance level (e.g., $\alpha \leq 0.05$) indicates the proportion of false positives that an investigator is willing to tolerate in his or her study.

1 million comparisons at an $\alpha \leq 0.05 \rightarrow 50\,000$ false positives
Assessing genetic association studies

*Was the analysis appropriate?*

Did they correct for multiple comparisons?

Agreed upon significance threshold: GWAS $1.5 \times 10^{-7}$

Bonferroni correction ($\alpha/n$) and Sidak $(1 - (1 - \alpha)^{1/n})$: too conservative, and assume SNPs are independent

False discover rate (FDR): control for the expected proportion of false positives among SNPs declared significant, and assumes SNPs are independent

Permutation testing
Assessing genetic association studies

*Was the analysis appropriate?*

Did they analyze individual markers or haplotypes?

Adapted from Tegeder et al., 2006
Assessing genetic association studies

Was the analysis appropriate?

Did they correct for covariates?
And finally, what are the *three most important* issues in genetic association studies?
Phenotype!
Phenotype!
Phenotype!