Statistical analysis of genome-wide association (GWAS) data

Jim Stankovich
Menzies Research Institute
University of Tasmania
J.Stankovich@utas.edu.au
Outline

• Introduction
• Confounding variables and linkage disequilibrium
• Statistical methods to test for association in case-control GWA studies
  – Allele counting chi-square test
  – Logistic regression
• Multiple testing and power
• Example: GWAS for multiple sclerosis (MS)
  – Data cleaning / quality control
  – Results
GWA studies have been very successful since 2007

- Prior to the advent of GWA studies, there was very little success in identifying genetic risk factors for complex multifactorial diseases
- GWA studies have identified over 200 separate associations with various complex diseases in the past two years
- “Human Genetic Variation” hailed as “Breakthrough of the Year” by Science magazine in 2007
This talk: case-control GWA studies

- Obtain DNA from people with disease of interest (cases) and unaffected controls
- Run each DNA sample on a SNP chip to measure genotypes at 300,000-1,000,000 SNPs in cases and controls
- Identify SNPs where one allele is significantly more common in cases than controls
  - The SNP is associated with disease

SNP: rs12425791  stroke
This talk: case-control GWA studies

- Obtain DNA from people with disease of interest (cases) and unaffected controls
- Run each DNA sample on a SNP chip to measure genotypes at 300,000-1,000,000 SNPs in cases and controls
- Identify SNPs where one allele is significantly more common in cases than controls
  - The SNP is associated with disease
- Alternative strategy (Peter Visscher’s talk): test for association between SNPs and a quantitative trait that underlies the disease (endophenotype)

SNP: rs12425791

blood pressure

stroke
Association does not imply causation

- Suppose that genotypes at a particular SNP are significantly associated with disease.
- This may be because the SNP is associated with some other factor (a *confounder*), which is associated with disease but is not in the same causal pathway.

SNP near lactase gene | multiple sclerosis (MS)
Association does not imply causation

- Suppose that genotypes at a particular SNP are significantly associated with disease
- This may be because the SNP is associated with some other factor (a *confounder*), which is associated with disease but is not in the same causal pathway

Diagram:
- Northern European ancestry
  - SNP near lactase gene
  - Multiple sclerosis (MS)
Association does not imply causation

• Suppose that genotypes at a particular SNP are significantly associated with disease
• This may be because the SNP is associated with some other factor (a confounder), which is associated with disease but is not in the same causal pathway
• Possible confounders of genetic associations:
  – Ethnic ancestry
  – Genotyping batch, genotyping centre
  – DNA quality
• Environmental exposures in the same causal pathway
  – Nicotine receptors --> smoking --> lung cancer
  – Alcohol dehydrogenase genes --> alcohol consumption --> throat cancer
**Helpful confounding: linkage disequilibrium**

*Linkage disequilibrium (LD)* is the non-independence of alleles at nearby markers in a population because of a lack of recombinations between the markers.

50,000 years ago:

- ~50kb

Today:

- “Haplotype block”
Direct and indirect association testing

Hirschhorn and Daly: Nature Reviews Genetics 6: 95 (2005)

Functional SNP (blue) is not genotyped, but a number of other SNPs (red), in LD with the functional SNP, are genotyped, and an association is found for these SNPs

Functional SNP is genotyped and an association is found
LD is helpful, because not all SNPs have to be genotyped

Allele counting to test for association between SNP genotype and case / control status

<table>
<thead>
<tr>
<th></th>
<th>GG</th>
<th>GT</th>
<th>TT</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$r_0$</td>
<td>$r_1$</td>
<td>$r_2$</td>
<td>$R$</td>
</tr>
<tr>
<td>Controls</td>
<td>$s_0$</td>
<td>$s_1$</td>
<td>$s_2$</td>
<td>$S$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_0$</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

**Observed allele counts**

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>T</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$2r_0+r_1$</td>
<td>$r_1+2r_2$</td>
<td>$2R$</td>
</tr>
<tr>
<td>Controls</td>
<td>$2s_0+s_1$</td>
<td>$s_1+2s_2$</td>
<td>$2S$</td>
</tr>
<tr>
<td>Total</td>
<td>$2n_0+n_1$</td>
<td>$n_1+2n_2$</td>
<td>$2N$</td>
</tr>
</tbody>
</table>
Allele counting to test for association between SNP genotype and case / control status

<table>
<thead>
<tr>
<th></th>
<th>GG</th>
<th>GT</th>
<th>TT</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$r_0$</td>
<td>$r_1$</td>
<td>$r_2$</td>
<td>$R$</td>
</tr>
<tr>
<td>Controls</td>
<td>$s_0$</td>
<td>$s_1$</td>
<td>$s_2$</td>
<td>$S$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_0$</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

**Observed allele counts**

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>T</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$2r_0 + r_1$</td>
<td>$r_1 + 2r_2$</td>
<td>$2R$</td>
</tr>
<tr>
<td>Controls</td>
<td>$2s_0 + s_1$</td>
<td>$s_1 + 2s_2$</td>
<td>$2S$</td>
</tr>
<tr>
<td>Total</td>
<td>$2n_0 + n_1$</td>
<td>$n_1 + 2n_2$</td>
<td>$2N$</td>
</tr>
</tbody>
</table>

**Expected allele counts**

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$2R(2n_0 + n_1)/(2N)$</td>
<td>$2R(n_1 + 2n_2)/(2N)$</td>
</tr>
<tr>
<td>Controls</td>
<td>$2S(2n_0 + n_1)/(2N)$</td>
<td>$2S(n_1 + 2n_2)/(2N)$</td>
</tr>
<tr>
<td>Total</td>
<td>$2N$</td>
<td></td>
</tr>
</tbody>
</table>
Allele counting to test for association between SNP genotype and case / control status

<table>
<thead>
<tr>
<th></th>
<th>GG</th>
<th>GT</th>
<th>TT</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td>$r_0$</td>
<td>$r_1$</td>
<td>$r_2$</td>
<td>$R$</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td>$s_0$</td>
<td>$s_1$</td>
<td>$s_2$</td>
<td>$S$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$n_0$</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Observed allele counts

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>T</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td>$2r_0+r_1$</td>
<td>$r_1+2r_2$</td>
<td>$2R$</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td>$2s_0+s_1$</td>
<td>$s_1+2s_2$</td>
<td>$2S$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$2n_0+n_1$</td>
<td>$n_1+2n_2$</td>
<td>$2N$</td>
</tr>
</tbody>
</table>

Expected allele counts

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td>$2R(2n_0+n_1)/(2N)$</td>
<td>$2R(n_1+2n_2)/(2N)$</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td>$2S(2n_0+n_1)/(2N)$</td>
<td>$2S(n_1+2n_2)/(2N)$</td>
</tr>
</tbody>
</table>

Chi-square test for independence of rows and columns (null hypothesis):

$$\sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} \sim \chi^2 \text{ with } 1 \text{ df}$$

PLINK --assoc option
Other options (e.g. dominant/recessive models)
--model
The odds ratio: a measure of effect size

Odds of an event occurring = Pr(event occurs) / Pr(event doesn’t occur)  
= Pr(event occurs) / [1 - Pr(event occurs)]

<table>
<thead>
<tr>
<th>Allele counts</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Consider all the G alleles in the sample, and pick one at random.  
The odds that the G allele occurs in a case: \( \frac{a}{c} \)

Consider all the T alleles in the sample, and pick one at random.  
The odds that a T allele occurs in a case: \( \frac{b}{d} \)

\[
\text{odds ratio} = \frac{\text{odds that G allele occurs in a case}}{\text{odds that T allele occurs in a case}} = \frac{a/c}{b/d} = \frac{a}{d} \cdot \frac{c}{b}
\]
### Interpretation of the odds ratio

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Controls</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

The odds ratio (OR) is defined as:

$$ \text{odds ratio (OR)} = \frac{\text{odds that G allele occurs in a case}}{\text{odds that T allele occurs in a case}} = \frac{a}{d} \div \frac{b}{c} $$

OR = increase in odds of being a case for each additional G allele

- OR = 1: no association between genotype and disease
- OR > 1: G allele increases risk of disease
- OR < 1: T allele increases risk of disease

If the disease is rare (e.g. ~0.1% for MS), the odds ratio is roughly equal to the genotype relative risk (GRR):
the increase in risk of disease conferred by each additional G allele

**Example:** If OR = 1.2,

$$ \Pr(\text{MS} \mid TT) = 0.1\% \quad \Pr(\text{MS} \mid GT) = 0.12\% \quad \Pr(\text{MS} \mid GG) = 0.144\% $$
Logistic regression: more flexible analysis for GWA studies

• Similar to linear regression, used for binary outcomes instead of continuous outcomes

• Let $Y_i$ be the phenotype for individual $i$
  $Y_i = 0$ for controls
  $Y_i = 1$ for cases

• Let $X_i$ be the genotype of individual $i$ at a particular SNP
  TT $X_i = 0$
  GT $X_i = 1$
  GG $X_i = 2$
Logistic regression: more flexible analysis for GWA studies

• Similar to linear regression, used for binary outcomes instead of continuous outcomes
• Let $Y_i$ be the phenotype for individual $i$
  - $Y_i = 0$ for controls
  - $Y_i = 1$ for cases
• Let $X_i$ be the genotype of individual $i$ at a particular SNP
  - TT: $X_i = 0$
  - GT: $X_i = 1$
  - GG: $X_i = 2$
• Basic logistic regression model
  - Let $p_i = E(Y_i | X_i)$, expected value of pheno given geno
  - Define $\text{logit}(p_i) = \log_e[p_i/(1 - p_i)]$
Logistic regression: more flexible analysis for GWA studies

- Similar to linear regression, used for binary outcomes instead of continuous outcomes
- Let $Y_i$ be the phenotype for individual $i$
  - $Y_i = 0$ for controls
  - $Y_i = 1$ for cases
- Let $X_i$ be the genotype of individual $i$ at a particular SNP
  - TT: $X_i = 0$
  - GT: $X_i = 1$
  - GG: $X_i = 2$
- Basic logistic regression model
  - Let $p_i = E(Y_i | X_i)$, expected value of pheno given geno
  - Define $\text{logit}(p_i) = \log_e[p_i/(1-p_i)]$
  - $\text{logit}(p_i) \sim \beta_0 + \beta_1 X_i$
Logistic regression: more flexible analysis for GWA studies

- Similar to linear regression, used for binary outcomes instead of continuous outcomes
- Let $Y_i$ be the phenotype for individual $i$
  - $Y_i = 0$ for controls
  - $Y_i = 1$ for cases
- Let $X_i$ be the genotype of individual $i$ at a particular SNP
  - TT $X_i = 0$
  - GT $X_i = 1$
  - GG $X_i = 2$
- Basic logistic regression model
  - Let $p_i = E(Y_i | X_i)$, expected value of pheno given geno
  - Define $\text{logit}(p_i) = \log_e[p_i/(1- p_i)]$

$$\text{logit}(p_i) \sim \beta_0 + \beta_1 X_i$$

Test whether $\beta_1$ differs significantly from zero:
roughly equivalent to allele counting chi-square test

Estimate of odds ratio: $\exp(\beta_1)$
Logistic regression: more flexible analysis for GWA studies

• Similar to linear regression, used for binary outcomes instead of continuous outcomes

• Let $Y_i$ be the phenotype for individual $i$
  $Y_i = 0$ for controls
  $Y_i = 1$ for cases

• Let $X_i$ be the genotype of individual $i$ at a particular SNP
  TT $X_i = 0$
  GT $X_i = 1$
  GG $X_i = 2$

• Add extra terms to adjust for potential confounders: e.g. ethnicity, genotyping batch, genotypes at other SNPs
  Let $p_i = E(Y_i | X_i, C_i, D_i, ...)$

\[
\text{logit}(p_i) \sim \beta_0 + \beta_1 X_i + \beta_2 C_i + \beta_3 D_i + \ldots
\]

PLINK --logistic
Multiple testing

• Suppose you test 500,000 SNPs for association with disease
• Expect around $500,000 \times 0.05 = 25,000$ to have p-value less than 0.05
• More appropriate significance threshold
  \[ p = \frac{0.05}{500,000} = 10^{-7} \]
  *genome-wide significance*

• In our MS GWAS we considered SNPs for follow-up if they had p-values less than 0.001
• To detect a smaller p-value need a larger study
The power to detect an association

• Suppose the G allele of a SNP has frequency 0.2. If each additional G allele increases odds of disease by 1.2, and 1618 cases and 3413 controls are genotyped, what is the power (chance) of detecting an association with significance p<0.001?

Null hypothesis: true OR=1
Observed OR from this distribution

Odds ratio (OR)

p=0.001
The power to detect an association

- Suppose the G allele of a SNP has frequency 0.2. If each additional G allele increases odds of disease by 1.2, and 1618 cases and 3413 controls are genotyped, what is the power (chance) of detecting an association with significance p<0.001?
The power to detect an association

- Suppose the G allele of a SNP has frequency 0.2. If each additional G allele increases odds of disease by 1.2, and 1618 cases and 3413 controls are genotyped, what is the *power* (chance) of detecting an association with significance \( p < 0.001 \)?
Effect of increasing sample size

Observed OR tends to be closer to true OR (narrower distributions)
⇒ Null and alternative distributions become more separate
⇒ Power increases
Multiple sclerosis - degradation of myelin sheath around nerve fibres
Multiple sclerosis

- neurodegenerative disease
- autoimmune attack on myelin sheaths around nerve cells
- more females affected than males (3:1)
- average age-at-onset ~30
- ~16,000 people with MS in Australia ($2 billion p.a.)
- no cure
Risk factors

- Epstein-Barr virus
- Exposure to infant siblings (Ponsonby et al, JAMA, 2005)
- Latitude gradient, childhood sun exposure (van der Mei et al, Lancet, 2003)
- Only genetic risk factor known before 2007 (first GWAS): HLA-DRB1*1501 discovered in 1972 (60% MS and 30% controls)

Timeline:

- 2000
- 2001
- 2002
- 2003
- 2004
- 2005
- 2006
- 2007
- 2008
- 2009

- Human genome project
- The SNP consortium
- The International HapMap Project
- SNP genotyping arrays
- GWA studies

Genes:

- IL7R
- IL2RA
- CLEC16A
- CD58
- EVI5/RPL5
- CD226
- KIF1B
- TYK2
Australian and New Zealand MS GWAS

• Assemble collection of DNA samples (all states + NZ)
• Genotype 1952 MS cases from around Australia and New Zealand with Illumina 370CNV BeadChips (Patrick Danoy, Matt Brown, Diamantina Institute, UQ)
• Analyse GWAS data
  – Quality control (Devindri Perera, Menzies)
  – Impute genotypes at millions of other SNPs (Sharon Browning, Univ of Auckland)
  – Compare case genotypes with >3500 controls from the UK and US (publicly available data)
• Replication genotyping (Justin Rubio’s lab, Howard Florey Institute, Univ of Melbourne)
Quality control - MS samples (PLINK)

- Start with 1952 samples
- Exclusions
  - Samples with >2% of SNPs not called 70 --mind
Genotype call rate

Graph showing the no call rate for Oragene saliva DNA samples.
Quality control - MS samples (PLINK)

• Start with 1952 samples
• Exclusions
  – Samples with >2% of SNPs not called 70
  – Suspect batch of samples 128
  – Uncertain phenotype 10
  – Duplicates / relatives 88
  – Ancestry outliers 35
Quality control - ethnicity

- Principal components analysis: EIGENSTRAT
  Nat Genet 38: 904

- Use an independent set of ~77,000 SNPs
  --indep-pairwise

- 178 outliers removed:
  - 35 MS
  - 143 controls
Quality control - MS samples

• Start with 1952 samples
• Exclusions
  – Samples with >2% of SNPs not called 70 --mind
  – Suspect batch of samples 128
  – Uncertain phenotype 10
  – Duplicates / relatives 88 --genome
  – Ancestry outliers 35
  – Sex discrepancies 3 --check-sex
• Leaves 1618 samples

Quality control - SNPs

• Start with 310,504 SNPs in both case and control datasets
• Exclude SNPs
  – Not called in >5% of samples --geno
  – In Hardy-Weinberg disequilibrium --hwe
  – Where one allele has frequency < 1% --maf
• Leaves 302,098 SNPs
Choice of 5% no-call threshold

• We originally planned to use a 10% threshold, but lots of SNPs with no call rate 5-10% showed deviations from Hardy-Weinberg equilibrium

• Closer look at SNPs with call rates between 5% and 10% suggested that they were unreliable
GWAS - results

Total sample = 1618 MS cases + 3413 controls

HLA

$P = 7 \times 10^{-94}$

$P = 10^{-7}$

$P = 0.001$
Extra QC for associated SNPs: cluster plots

UK controls  ANZ cases  both
The replication phase

- Selected 100 SNPs for replication genotyping

- 2,256 ANZ MS cases + 2,310 ANZ controls

- Two chromosome regions on chr 12 and chr 20 showed (almost) genome-wide significant (p<5 x 10⁻⁷) association with MS after combining GWAS and replication data

- SNPs in 13/53 other regions with replication p-values < 0.1: more than expected by chance (p=0.002)
Chromosome 12 association: the downside of LD

- rs703842
  - GWAS
    - P = 4.1 x 10^{-6}
  - replication
    - P = 1.4 x 10^{-6}
  - GWAS + rep
    - P = 5.4 x 10^{-11}

- Allele frequency 0.33
- Odds ratio 0.81 (protective)
Chromosome 12 association: the downside of LD

KIF5A SNP associated with rheumatoid arthritis and type 1 diabetes
Chromosome 12 association: the downside of LD

Logistic regression with both SNPs in the same model:
rs10876994 $p = 0.004$
rs703842 $p = 0.08$
Chromosome 12 association: the downside of LD

CYP27B1: most likely candidate??
**CYP27B1**

- Cytochrome p450 gene family (drug metabolizing)
- Encodes 25-hydroxyvitamin D-1 alpha hydroxylase (1α-OHase)
- Converts 25(OH)D\(_3\) to bioactive 1,25(OH)\(_2\)D\(_3\)
- 1,25(OH)\(_2\)D\(_3\) regulates calcium metabolism and the immune system via vitamin D receptor (VDR)

Adorini and Penna (2008)
Nat Clin Prac Rheum 4: 404-12
The chromosome 20 association

- **rs6074022**
  - GWAS
    - $P = 2.5 \times 10^{-5}$
  - Replication
    - $P = 4.6 \times 10^{-4}$
  - GWAS + Rep
    - $P = 1.3 \times 10^{-7}$

- Allele frequency: 0.25
- Odds ratio: 1.20 (increased risk)
**CD40**

- Member of TNF receptor superfamily: regulates many cell- and antibody-mediated immune responses
- SNPs in CD40 are associated with risk of rheumatoid arthritis and Graves’ disease
- Functional SNP rs1883832C>T, 1 base pair upstream of the ATG translation initiation codon
- Allelic heterogeneity
Another use of logistic regression: test for gene-gene interaction

Modest evidence that each risk allele has a bigger effect in the presence of the other risk allele (p = 0.03)

MS risk alleles
Chr 12 = rs703842A
Chr 20 = rs6074022G
Summary

- Case-control GWA studies have been very successful in the past couple of years
- Linkage disequilibrium means that most, but not all, common human genetic variation is captured by genotyping a few hundred thousand SNPs
- Small effect sizes (e.g. OR 1.2) mean that GWA studies need to be large, with thousands of cases and controls --> big collaborations
- Methods of statistical analysis are fairly straightforward, but care is required to clean data
- The ultimate test of any association: replication in an independent population
**Acknowledgments - MS GWAS**

| Hobart: | Devindri Perera  
|         | Bruce Taylor  
|         | Karen Drysdale  
|         | Preethi Guru  
|         | Brendan McMorrnan  
|         | Simon Foote  
| Melbourne: | Justin Rubio  
|           | Melanie Bahlo  
|           | Helmut Butzkueven  
|           | Vicky Perreua  
|           | Laura Johnson  
|           | Judith Field  
|           | Cathy Jensen  
|           | Ella Wilkins  
|           | Caron Chapman  
|           | Mark Marriott  
|           | Niall Tubridy  
|           | Trevor Kilpatrick  
| Gold Coast: | Simon Broadley  
|             | Lyn Griffiths  
|             | Lotfi Tajouri  
|             | Michael Pender  
| Brisbane: | Matthew Brown  
|           | Patrick Danoy  
|           | Johanna Hadler  
|           | Karen Pryce  
|           | Peter Csurshes  
|           | Judith Greer  
| Sydney: | Graeme Stewart  
|         | David Booth  
|         | Robert Heard  
|         | Jim Wiley  
| Adelaide: | Mark Slee  
| New Zealand: | Sharon Browning  
|             | Brian Browning  
|             | Deborah Mason  
|             | Ernie Willoughby  
|             | Glynnis Clarke  
|             | Ruth McCallum  
|             | Marilyn Merriman  
|             | Tony Merriman  
| Newcastle: | Jeanette Lechner-Scott  
|           | Rodney Scott  
|           | Pablo Moscato  
|           | Mathew Cox  
| Perth: | Bill Carroll  
|        | Alan Kermlode  

**Funding:**  
MS Research Australia, John T Reid Charitable Trusts,  
Trish MS Research Foundation, Australian Research Council  

The Australian and NZ MS Genetics Consortium (2009).  
Nat Genet 41: 824