Statistical Methodologies for Analyzing Whole Genome Association Data

John P. Rice, Ph.D.
Washington University School of Medicine
Crossing Over During Meiosis

(a)  

(b)  

(c)  

(d)
Gametes $A_1 B_2, A_2 B_1$ are recombinants
$A_1 B_1, A_2 B_2$ are non-recombinants

$\theta = \text{Prob (recombinant)}$

$\theta = .01 \iff \text{A and B are 1 cM apart}$
Genome Arithmetic

- Kb=1,000 bases; Mb=1,000Kb
- 3.3 billion base pairs; 3,300 cM in genome
  \[ \frac{3,300,000,000}{3,300} = 1 \text{ Mb/cM} \]
- 33,000 genes
  \[ \frac{33,000}{3,300} \text{ Mb} = 10 \text{ genes / Mb} \]
- Thus, 20 cM region may have 200 genes to examine
- Erratum – closer to 20,000 genes in humans
Linkage Vs. Association

- **Linkage:**
  - Disease travels with marker within families
  - No association within individuals
  - Signals for complex traits are wide (20MB)

- **Association:**
  - Can use case/control or case/parents design
  - Only works if association in the population
  - Allelic heterogeneity (eg, BRAC1) a problem

- Linkage – large scale; Association fine scale (<200kb)
Example of a LOD Curve
Disequilibrium

Let \( P(A_1) = p_1 \)
Let \( P(B_1) = q_1 \)
Let \( P(A_1B_1) = h_{11} \)

No association if \( h_{11} = p_1q_1 \)

\[ D = h_{11} - p_1q_1 \]
D′ and $r^2$

D tends to take on small values and depends on marginal gene frequencies

\[ D' = \frac{D}{\max(D)} \]
\[ r^2 = \frac{D^2}{(p_1p_2q_1q_2)} = \text{square of usual correlation coefficient (φ)} \]

Note: $r^2 = 0 \iff D' = 0$

D′ = ±1 if one cell is zero

$r^2$ can be small even when $D' = ±1$

Prediction of one SNP by another depends on $r^2$
Basic Idea

- If SNP A is a disease susceptibility gene, and if we genotype SNP B (for example in a whole genome association study), and if A and B are in disequilibrium, then cases and controls will have different frequencies of alleles at B.

- Power to detect A is related to $N/r^2$. 
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>B1</th>
<th>B2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>B</td>
<td>50</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50.00</td>
<td>0.00</td>
<td>50.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>55.56</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>B</td>
<td>40</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40.00</td>
<td>10.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>80.00</td>
<td>20.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>44.44</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>90</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90.00</td>
<td>10.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.00</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

$D' = 1, r^2 = .1$
D ′ = 1, r² = .01
Blocks and Bins

- Predictability of one SNP by another best described by $r^2$ – basic statistics
- Block – set of SNPs with all pair-wise LD high (Please specify measure)
- If one uses $r^2$ – insert a SNP with low frequency in between SNPs with freqs close to 0.5, then block breaks up!
- Perlegen (Hinds et al, Science, 2005) — use bins where a tag SNP has $r^2$ of 0.8 with all other SNPs. Bins may not be contiguous.
Summary (Blocks and Bins)

- Blocks using $D'$ may have a “biological” interpretation (long stretches with $|D'| = 1$)
- Selection of Tag SNPs is a statistical issue, want to predict untyped SNPs from those that are typed – $r^2$ is natural measure
- Phase of SNPs is important – usually ignored
- Most current WGA studies use bins based on $r^2$ (typically $r^2 > 0.8$)
- There is an art to selecting tag SNPs
Statistical Analysis

- **Case/Control Design**
  - Use standard statistical tests (logistic regression) to test whether the distribution of the SNP differs between cases and controls
  - Sensitive to population stratification
- **Family Based Design**
  - Alleles 1 and 4 are transmitted -- CASE
  - Alleles 2 and 3 are non-transmitted –CONTROL
  - NOTE: Genotype 3 people to get 1 case and 1 control
  - NOT sensitive to population stratification
Problem of Multiple Tests

Significant level = $\alpha$

We perform $N$ (independent) tests

We expect to reject $N\alpha$ tests if null hypothesis is true for each test.

**Example:**

$N = 100$, $\alpha = .05$, $x =$ # of rejections

$$P(x \geq 1) = 1 - P(x = 0)$$

$$= 1 - (1 - \alpha)^{100}$$

$$= .99408$$

Note: $1 - (1 - \alpha)^N \approx N\alpha$ for $\alpha$ small

Choose $\alpha' = \alpha/N = .0005$

The $1 - (1 - \alpha')^{100} = .0488$

**Bonferroni Correction**

**Problem:** Power goes down as $\alpha$ decreases
Multiple tests for association

- Intuition: LD extents over smaller regions than linkage
- More “independent” tests for LD -- There must be at the equivalent of at least 200,000 independent tests in one experiment (linkage about 2,000 independent tests)
- Multiple testing for whole genome association studies will be problematic
- Practical question – How to correct for multiple tests
Multiple Testing

- Suppose we use 600,000 SNPs, and there are 10 true susceptibility loci. Test at significance level \( p=0.001 \), and power is 60%.
- We expect \( 10 \times 0.6 = 6 \) true positives, and \( 600,000 \times 0.001 = 600 \) false positives. We expect one false positive to be significant at the 0.0000002 level.
- Tests are not independent, so use of Bonferroni correction of \( 0.05/600,000 = 0.000000008 \) is too conservative. Even with appropriate p-value, there would be little power without massive sample sizes. A gene with the effect size needed to be detected would already be known.
False Discovery Rate (FDR)

- $V = \#$ true null hypotheses called significant
- $S = \#$ non-true hypotheses called significant
- $Q = \frac{V}{V + S}$ (false positives/all positives)
- $FDR = E(Q)$

- Benjamini & Hochberg (1995)
  - When testing $m$ hypotheses $H_1, \ldots, H_m$, order $p$-values $p_1, \ldots, p_m$, let $k$ be largest $i$ for which $p_i \leq \frac{i}{m} q^*$
  - Then reject $H_1, \ldots, H_m$

Theorem: Above controls FDR at $q^*$

Computer program: QVALUE
Multiple Testing

- FDR helps and is commonly used
- Question: Should all markers be tested using the same p-value?
  Use a set of weights in the FDR computations. If a small proportion are over-weighted, does not reduce the power to detect the others very much, but helps the detection of the ones to “bet” on.
  Use of prior linkage evidence may be a way to increase power.
Example: Top 10 SNPs from Analysis of 1,500 SNPs

<table>
<thead>
<tr>
<th>Obs</th>
<th>rs ID</th>
<th>Primary p-value (p_V3_MI)</th>
<th>Storey q-value (smoother pi0 = 0.89)</th>
<th>LD Bin</th>
<th>Number of markers in bin</th>
<th>Min r^2 of tag with nontags</th>
<th>MAF</th>
<th>MAF (HapMap)</th>
<th>Chr</th>
<th>Pos (bp)</th>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12334778</td>
<td>1.21E-04</td>
<td>0.113</td>
<td>85</td>
<td>2</td>
<td>0.968</td>
<td>0.484</td>
<td>0.500</td>
<td>8</td>
<td>123,567,696</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>4506214</td>
<td>2.56E-04</td>
<td>0.113</td>
<td>85</td>
<td>2</td>
<td>0.968</td>
<td>0.477</td>
<td>0.492</td>
<td>8</td>
<td>123,564,218</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>4336618</td>
<td>3.47E-04</td>
<td>0.113</td>
<td>68</td>
<td>2</td>
<td>0.982</td>
<td>0.325</td>
<td>0.408</td>
<td>8</td>
<td>123,536,162</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>6986303</td>
<td>3.66E-04</td>
<td>0.113</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.294</td>
<td>0.275</td>
<td>8</td>
<td>134,547,711</td>
<td>ST3GAL1</td>
<td>INTRON</td>
</tr>
<tr>
<td>5</td>
<td>7846137</td>
<td>4.35E-04</td>
<td>0.113</td>
<td>68</td>
<td>2</td>
<td>0.982</td>
<td>0.318</td>
<td>0.400</td>
<td>8</td>
<td>123,532,225</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>6</td>
<td>6470170</td>
<td>5.44E-04</td>
<td>0.118</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.226</td>
<td>0.150</td>
<td>8</td>
<td>124,736,503</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>7</td>
<td>10101440</td>
<td>1.26E-03</td>
<td>0.234</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.127</td>
<td>0.125</td>
<td>8</td>
<td>127,604,613</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>8</td>
<td>1394051</td>
<td>1.71E-03</td>
<td>0.279</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.099</td>
<td>0.100</td>
<td>8</td>
<td>134,927,184</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>9</td>
<td>4870857</td>
<td>2.74E-03</td>
<td>0.397</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.398</td>
<td>0.383</td>
<td>8</td>
<td>124,661,555</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>10</td>
<td>4259418</td>
<td>3.51E-03</td>
<td>0.457</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.318</td>
<td>0.458</td>
<td>8</td>
<td>123,518,309</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
Conclusions

- WGA studies will be done (6 GAIN studies have just been selected) and be in the public domain
- Candidate gene studies have been problematic (the prior probability of selecting the right gene may be 1/10,000), so may be very low power.
- Multiple testing issues a major challenge for WGA studies, but these will be overcome