Power of Genomewide Association Studies

Biostatistics 666
A Simple Disease Model

- Risk allele frequency $p$
- Background allele frequency $f$
- Increase in disease risk per allele $r$

- Examples:
  - $HLA$-$C$ risk allele for psoriasis, $p=.15$, $f=.0065$, $r=2.6$
  - $TNIP1$ risk allele for psoriasis, $p=.05$, $f=.0095$, $r=1.8$
  - $TCF7L2$ risk allele for type 2 diabetes, $p=.35$, $f=.08$, $r=1.4$
  - $R1210C$ risk allele for macular degeneration, $p=10^{-4}$, $f=.05$, $r=25$

- $f$ selected so overall risk of disease is about 1%
What Happens in Cases ...

\[
P(\text{case} \& \text{low risk}) = (1 - p)^2 f
\]
\[
P(\text{case} \& \text{med risk}) = 2p(1 - p)fr
\]
\[
P(\text{case} \& \text{high risk}) = p^2 fr^2
\]

\[
P(\text{case}) = ((1 - p)^2 + 2p(1 - p)r + p^2 r^2) f
\]

\[
P(\text{low risk}|\text{case}) = (1 - p)^2 f / P(\text{case})
\]
\[
P(\text{med risk}|\text{case}) = 2p(1 - p)fr / P(\text{case})
\]
\[
P(\text{high risk}|\text{case}) = p^2 fr^2 / P(\text{case})
\]

\[
P(\text{risk allele}|\text{case}) = (p(1 - p)r + p^2 r^2) / P(\text{case})
\]
What Happens in Screened Controls ...

\[ P(\text{control} \& \text{low risk}) = (1 - p)^2(1 - f) \]
\[ P(\text{control} \& \text{med risk}) = 2p(1 - p)(1 - fr) \]
\[ P(\text{control} \& \text{high risk}) = p^2(1 - fr^2) \]

\[ P(\text{control}) = (1 - p)^2(1 - f) + 2p(1 - p)(1 - fr) + p^2(1 - fr^2) \]

\[ P(\text{low risk}|\text{control}) = (1 - p)^2(1 - f)/P(\text{control}) \]
\[ P(\text{med risk}|\text{control}) = 2p(1 - p)(1 - fr)/P(\text{control}) \]
\[ P(\text{high risk}|\text{control}) = p^2(1 - fr^2)/P(\text{control}) \]

\[ P(\text{risk allele}|\text{control}) = (p(1 - p)(1 - fr) + p^2(1 - fr^2))/P(\text{control}) \]
Today

• A simple genetic model: frequency + risk

• A typical genomewide association study

• Power for genomewide association study

• Designing a two stage genomewide study

• Choices for analysis of two stage studies
Genomewide Association Studies

• Survey ~500,000 SNPs in a large set of cases and controls
  – Subset of SNPs is typically followed up in more samples

• Comprehensively survey common variants across genome
  – Via linkage disequilibrium, most common variants assessed

• Successful: many loci implicated in common disorders
  – Especially in contrast to results of candidate gene studies
Collaborative Association Study of Psoriasis: Example of a Successful GWAS

- Examined ~1,500 cases / ~1,500 controls at ~500,000 SNPs
- Examined 20 promising SNPs in extra ~5,000 cases / ~5,000 controls
- Outcome: 7 regions of confirmed association with psoriasis

Green hits have $p < 5 \times 10^{-8}$ in final analysis

Nair et al, 2009
Top psoriasis associated SNPs in **strong linkage disequilibrium with HLA-Cw6**. Evidence for psoriasis associated SNPs that are far from HLA-Cw6.
Previously identified locus, psoriasis associated SNPs also associated with Crohn’s.
Previously identified locus, psoriasis associated SNPs associated with Crohn’s.
New locus, psoriasis associated SNPs **not associated** with Crohn’s.
New locus; other SNPs in the locus are associated with lupus and rheumatoid arthritis.
TNIP1

New locus; note potential evidence for independently associated alleles.
New locus; IL4 and IL13 are excellent functional candidates.
Q-Q Plot

Genomic control = 1.03
Multiple hits within a pathway...

• Three of the top replicated hits are for:
  – IL23R (IL-23 receptor)  $3 \times 10^{-8}$
  – IL23A (IL-23 subunit)  $9 \times 10^{-10}$
  – IL12B (IL-23/IL-12 subunit)  $1 \times 10^{-28}$

• Two other replicated hits at:
  – TNFAIP3 (TNFα-inducible protein 3)  $9 \times 10^{-12}$
  – TNIP1 (TNFAIP3 interacting protein 1)  $1 \times 10^{-20}$

• Evidence for epistasis among these SNPs?
  – None.
### Summary of Results

<table>
<thead>
<tr>
<th>SNP</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>P-value</th>
<th>Nearby Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$f_{\text{cases}}$</td>
<td>$f_{\text{controls}}$</td>
<td>OR</td>
<td>$f_{\text{cases}}$</td>
</tr>
<tr>
<td>rs12191877</td>
<td>.31</td>
<td>.14</td>
<td>2.79</td>
<td>.30</td>
</tr>
<tr>
<td>rs2082412</td>
<td>.86</td>
<td>.79</td>
<td>1.56</td>
<td>.85</td>
</tr>
<tr>
<td>rs17727338</td>
<td>.09</td>
<td>.06</td>
<td>1.72</td>
<td>.09</td>
</tr>
<tr>
<td>rs20541</td>
<td>.83</td>
<td>.78</td>
<td>1.37</td>
<td>.83</td>
</tr>
<tr>
<td>rs610604</td>
<td>.37</td>
<td>.32</td>
<td>1.28</td>
<td>.36</td>
</tr>
<tr>
<td>rs2066808</td>
<td>.96</td>
<td>.93</td>
<td>1.68</td>
<td>.95</td>
</tr>
<tr>
<td>rs2201841</td>
<td>.35</td>
<td>.29</td>
<td>1.35</td>
<td>.32</td>
</tr>
</tbody>
</table>

Notice how estimated effect size is consistently higher in Stage 1. The “Winner’s Curse” is a common feature of genomewide studies.
Power Calculations

• For a given genetic model, evaluate alternative study designs

• For a given study design, identify genetic models that are likely to be detected

• Typically deal with many uncertainties...
  – What is an appropriate genetic model?
  – What is a desirable level of power?
Test Statistic

\[ z = \frac{\hat{p}' - \hat{p}}{\sqrt{[\hat{p}'(1 - \hat{p}')] + \hat{p}(1 - \hat{p})]/2N} \]

Where:

\( \hat{p}' \) is the observed case allele frequency
\( \hat{p} \) is the observed control allele frequency
\( N \) is the number of cases and controls
Distribution Under the Null

- Under the null hypothesis $p = p'$

- $Z$ is distributed as $\text{Normal}(0, 1)$

- Derive P-value thresholds for target significance level $\alpha$

- Using Inverse Normal Cumulative Distribution Function
  - $\alpha = 0.05$ leads to $C = -\Phi^{-1}\left(\frac{0.05}{2}\right) = 1.96$
  - $\alpha = 5 \cdot 10^{-8}$ leads to $C = -\Phi^{-1}\left(\frac{5 \cdot 10^{-8}}{2}\right) = 5.45$
Distribution Under The Alternative

• For a specific set of expected case and control allele frequencies, ...

• ...we can calculate expected value of test statistic

\[ \mu = \frac{p' - p}{\sqrt{[p'(1 - p') + p (1 - p)]/2N}} \]

• Under the alternative, statistic is Normal(\(\mu, 1\)).
Power

• To calculate power, we first calculate:
  – Significance threshold \( C \)
  – Expected test statistic \( \mu \)

• Use normal cumulative distribution function \( \Phi \)

• \( P(|Z| > C) \)
  \[
  = P(Z > C) + P(Z < -C) \\
  = 1 - \Phi(C - \mu) + \Phi(-C - \mu)
  \]
Example

- Test 1,000,000 independent markers
  - \( \alpha = \frac{0.05}{1,000,000} = 5 \times 10^{-8} \)
  - \( C = 5.45 \)

- Case allele frequency \( p' = 0.55 \)
- Control allele frequency \( p = 0.45 \)
- \( N_{\text{cases}} = N_{\text{controls}} = 1,000 \)
- \( \mu = 6.35 \)

- Power = 81%
  - If \( N = 500 \), power = 17%
  - If \( N = 2000 \), power = 100%
A comprehensive study might examine all $M$ SNPs in all $N$ samples.
Analysis of One Stage Study

Declare significance using p-value threshold of $0.05 / M$. Threshold of $5 \times 10^{-8}$ is typical, assumes 1 million independent tests.
Two Stage Genomewide Association Studies
Two Stage Genomewide Study

A more cost effective study might only examine:

- All SNPs in a fraction of samples, $\pi_{\text{samples}}$
- All individuals for a fraction of markers, $\pi_{\text{markers}}$
Relative Genotyping Effort

• The total number of genotypes required in a two stage study is …

\[ N_{genotypes} = MN \pi_{samples} + MN(1 - \pi_{samples})\pi_{markers} \]

• For example, if we …
  – Genotype 30% of samples in Stage 1
  – Follow-up 0.1% of markers in Stage 2
  – Total number of genotypes will be reduced 69.93%
Relative Cost

• The reduction in cost is typically less dramatic ...
• ... but still substantial

• Main limitation is that genotyping is cheaper “in bulk”
  – τ is ratio of stage 1 to stage 2 costs on a per genotype basis

• Cost ratio = \( \pi_{samples} + (1 - \pi_{samples})\pi_{markers}\tau \)

• For example, if we ...
  – Genotype 30% of samples in Stage 1
  – Follow-up 0.1% of markers in Stage 2
  – Relative cost ratio is 100

  – Total cost will be reduced 63.00%
Replication Based Analysis

Select markers to follow-up using p-value threshold of $\pi_{\text{markers}}$.
Declare significance using threshold of $0.05 / (M \cdot \pi_{\text{markers}})$
Final analysis uses only stage 2 samples.
Joint Analysis

Select markers to follow-up using p-value threshold of $\pi_{\text{markers}}$. Declare significance using threshold of approximately 0.05/M. Final analysis uses stage 1 and stage 2 samples.
Power for Replication Based Analysis

• Simplest approach would be to calculate
  – $C_1$ and $C_2$ as the significance thresholds for each stage
  – $\mu_1$ and $\mu_2$ as the expected statistics for each stage
  – $P_1$ and $P_2$ as the power for each stage
  – $P_{\text{replication}} = P_1 P_2$ as the overall power

• Refined analysis might enforce that stage 1 and stage 2 statistics should have the same sign

\[
P_2 = (1 - \Phi[C_2 - \mu_2]) \frac{1 - \Phi[C_1 - \mu_1]}{1 - \Phi[C_1 - \mu_1] + \Phi[-C_1 - \mu_1]} + \Phi[-C_2 - \mu_2] \frac{\Phi[-C_1 - \mu_1]}{1 - \Phi[C_1 - \mu_1] + \Phi[-C_1 - \mu_1]}
\]
Power for Joint Analyses

- Simplest approach would be to calculate
  - $C_1$ and $C$ as stage 1 and overall significance thresholds
  - $\mu_1$ and $\mu$ as stage 1 and overall expected statistics
  - $P_1$ and $P$ as stage 1 and single stage study power
  - $P_{\text{joint}} = P_1 P$ as the overall power

- Refined analysis models joint distribution of stage 1 and overall test statistic

\[
P_{\text{joint}} = P(|z_{\text{joint}} > C_{\text{joint}}| T)
\]
\[
= \int_{-C_1}^{-\infty} [P(z_{\text{joint}} > C_{\text{joint}} | z_1 = x) + P(z_{\text{joint}} < -C_{\text{joint}} | z_1 = x)] f(x|T) dx
\]
\[
+ \int_{C_1}^{\infty} [P(z_{\text{joint}} > C_{\text{joint}} | z_1 = x) + P(z_{\text{joint}} < -C_{\text{joint}} | z_1 = x)] f(x|T) dx
\]

$T: |Z| > C_1$
Replication or Joint Analysis?

• Replication based analysis
  – Requires smaller multiple testing adjustment

• Joint analysis uses more data
  – We expect stronger signal using all available data

• Both analyses are compatible with the same experimental design
Replication of Joint Analysis?

300,000 markers genotyped on 1000 cases, 1000 controls
Multiplicative model, prevalence 10%, GRR = 1.4
Replication or Joint Analysis?
Effect of Varying $\pi_{samples}$

- $\alpha=0.05 / 300,000$
- $\pi_{markers} = 0.01$
- $N = 1,000$
- $p = 0.50$
- $p'= 0.66$
Replication or Joint Analysis?

Effect of Varying $\pi_{\text{markers}}$

- $\alpha = 0.05 / 300,000$
- $\pi_{\text{samples}} = 0.30$
- $N = 1,000$
- $p = 0.50$
- $p' = 0.66$
Refining Calculation

• Instead of setting $p$ and $p'$ arbitrarily, use a genetic model

• Suppose that the relative risk of disease is:
  – Baseline for those with no risk alleles
  – $r_1$ for those with one risk allele
  – $r_2$ for those with two risk alleles

• Then:

$$p' = \frac{p(1 - p)r_1 + p^2 r_2}{(1 - p)^2 + 2p(1 - p)r_1 + p^2 r_2}$$
Refining Calculation II

- Instead of setting $p$ and $p'$ arbitrarily, use a genetic model

- Suppose that controls are known to be free of disease and $K$ is the disease prevalence

- Then:

$$ p_{control} = \frac{p - Kp'}{1 - K} $$
Some Important Messages

• Power calculations can help design study
  – How to best invest limited funds?

• Well designed two stage studies approximate power of more costly studies where all samples genotyped at all markers

• Joint analysis is much more efficient than replication based analyses
Recommended Reading
