Today’s Outline

- Single variant association analysis.
- Single variant association analysis for genome-wide association studies (GWAS).
- Effect size estimation and winner’s curse.
In GWAS, single variant test is the most popular approach to investigating associations.

- $Y_i$: outcomes for $i = 1, \ldots, n$
- $X'_i = (1, x_{i1}, \ldots, x_{iq})$: covariates including the intercept.
- Regression model

\[
g\{E(Y_i)\} = X'_i \alpha + G_i \beta. \]

- If $Y$ is continuous, $g(\cdot)$ is a linear link; If $Y$ is binary, $g(\cdot)$ is a logit link, $\log\{\Pr(Y = 1)/\Pr(Y = 0)\}$. 
Single Variant Test

- $G_i$: genotype value. Suppose the locus takes two alleles, A and a
  - Additive:
    \[ AA = 0, \quad Aa = 1, \quad aa = 2 \]
  - Dominant:
    \[ AA = 0, \quad Aa = 1, \quad aa = 1 \]
  - Recessive:
    \[ AA = 0, \quad Aa = 0, \quad aa = 1 \]
Single Variant Test

- Null hypothesis \( H_0 : \beta = 0 \)

- Three asymptotically equivalent tests
  - Wald test:
    \[
    \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})} \sim N(0, 1)
    \]
  - Score test:
    \[
    \left[ \frac{\partial}{\partial \beta} \log L(\beta, \hat{\alpha}_0) \right]_{\beta=0} I(\beta = 0|\hat{\alpha}_0)^{-1} \left[ \frac{\partial}{\partial \beta} \log L(\beta, \hat{\alpha}_0) \right]_{\beta=0} \sim \chi^2_1
    \]
    \[
    I(\beta = 0|\hat{\alpha}_0) = \left\{ I_{\beta\beta} - I_{\beta\alpha} I^{-1}_{\alpha\alpha} I_{\alpha\beta} \right\}_{\beta=0, \hat{\alpha}_0}
    \]
  - Likelihood ratio (LR) test:
    \[
    2\left\{ \log L(\hat{\alpha}, \hat{\beta}) - \log L(\hat{\alpha}_0, 0) \right\} \sim \chi^2_1
    \]

- Wald test is most intuitive. LR test is directly related to Neyman-Pearson lemma. The score test can be very fast, as it doesn’t require fitting the model under the alternative.
Single Variant Analysis for GWAS Data

- Manhattan plot of GWAS (genome-wide association studies) association analysis ($n \approx 40,000$).

Confounding

- Population stratification is a major confounder in genetic association studies

- It occurs in the following scenario:
  - The phenotype is more common in one population
  - Allele frequencies are different between populations

- The effects of stratification increase with sample size, so that even subtle population substructure can yield grossly inflated type I error for large GWAS
Detecting Stratification

- Quantile-Quantile (QQ) plot shows little stratification.
Detection Stratification

- QQ plot shows stratification

Controlling for Stratification

- **Study design**
  - Careful sampling
  - Family-based controls

- **Statistical methods based on largely “null” markers.**
  - Genomic control
  - Structured association
  - Principal component analysis
Genomic Control

- Select unlinked markers (e.g., pairwise distance $\geq 100$ kb)
- Compute $\chi^2$ for each marker
- Inflation $\lambda = \text{Median observed } \chi^2 / 0.456$
- Adjust statistic by
  \[
  \chi^2_{\text{fair}} = \frac{\chi^2_{\text{observed}}}{\lambda}
  \]
- $\lambda$ also provides a convenient way to summarize magnitude of stratification
Why Genomic Control?

- Simple and convenient approach.

However,

- Crude adjustment, especially when the degrees of stratification vary substantially among the SNPs.
- Does stratification inflate the $p$-value to the same extent under the alternative?

---

Structured Association

- Use unlinked markers to assign individuals to subpopulation
  - Suppose Z are the latent subpopulations, P are allele frequencies in K subpopulations, G are observed genotypes
  - Step 1: Sample $P^{(m)}$ from $Pr(P|G, Z^{(m-1)})$
  - Step 2: Sample $Z_i^{(m)}$ from $Pr(Z_i|G, P^{(m)})$ for each $i$
  - All calculations involves $Pr(G|P, Z)$, which assumes Hardy-Weinberg equilibrium

- Test for association within each population or test for association while conditioning on subpopulation
Features

▶ Can be inferred with relatively few SNPs, but computationally intractable for large # of SNPs.
▶ Describing subpopulation can be useful.

However,

▶ Difficult to correctly estimate the population substructure or to correctly assign individuals to subpopulations, especially when the population under study is a continuous mixture of ancestral subpopulation.

Principal Components Analysis

- Infer continuous axes of genetic variation from SNPs.
Model

\( Y \): 1 vs 0, whether or not the subject has the disease of interest.

\( G \): Genotype at a candidate locus.

\( U \): Unknown population structure.

\( Z \): A set of SNPs, which is informative about latent \( U \).

▶ True model

\[
\text{logit}\{\Pr(Y = 1|G, U, Z)\} = G\beta + \gamma(U, Z)
\]

▶ \( \beta \) is parameter of interest, but not identifiable because \( U \) is not observed.
Statistical Framework

- Marginal model

\[
\frac{\Pr(Y = 1 | G, Z)}{\Pr(Y = 0 | G, Z)} = \frac{\Pr(Y = 1, G, Z)}{\Pr(Y = 0, G, Z)} = \int \frac{\Pr(Y = 1, G, Z, u)}{\Pr(Y = 0, G, Z, u)} \frac{\Pr(Y = 0, G, Z, u)}{\Pr(Y = 0, G, Z)} du \\
= \exp(G\beta) \int \exp\{\gamma^*(u, z)\} P(u|Y = 0, G, Z) du
\]

- In order for the second term not to be a function of \( G \)

\[
\Pr(U = u | G, Z, Y = 0) = \Pr(U = u | Z, Y = 0)
\]

- Let \( \xi(Z) \) be an unknown function, we can rewrite

\[
\logit\{\Pr(Y = 1 | G, Z)\} = G\beta + \xi(Z)
\]
A necessary and sufficient condition for (1) to hold is
\[ \Pr(U = u | G, Z, Y = 0) = \Pr(U = u | Z, Y = 0) \]
Or equivalently
\[ \Pr(U = u | G, Z, Y = 1) = \Pr(U = u | Z, Y = 1) \]
This can be seen from
\[ \Pr(U, G | Z, Y = 1) = \Pr(U, G | Z, Y = 0) \exp(\beta G + \gamma(U, Z)) \frac{\Pr(Z, Y = 0)}{\Pr(Z, Y = 1)} \]
\[ Z \text{ dissolves the link between } U \text{ and } G \text{ such that } U \independent G \text{ for each stratum of } Z \text{ in the control (or case) population.} \]
Modeling $\xi(Z)$

- Reduce potentially high dimension $Z \rightarrow \psi(Z)$
- If $\Pr(G = g | Z = z, Y = 0) = \Pr(G = g | \psi(Z) = \psi(z), Y = 0)$ then

  $$\logit\{\Pr(Y = 1 | G = g, \psi(Z) = x)\} = \beta g + \xi(x)$$

- Sketch of proof:

  $$\frac{\Pr(Y = 1, G = g, \psi(Z) = x)}{\Pr(Y = 0, G = g, \psi(Z) = x)} = \frac{\int_{u,z:\psi(z)=x} \Pr(Y = 1, G = g, Z, u) dZdu}{\Pr(Y = 0, G = g, \psi(Z) = x)}$$

  $$= \frac{\int_{u,z:\psi(z)=x} \exp(G\beta + \gamma(u, Z)) \Pr(Y = 0, G = g, Z) dZdu}{\Pr(G = g | Y = 0, \psi(Z) = x) \Pr(Y = 0, \psi(Z) = x)}$$

  $$= \frac{\exp(G\beta) \int_{u,z:\psi(z)=x} \exp(\gamma(u, Z)) \Pr(Y = 0, Z) dZdu}{\Pr(Y = 0, \psi(Z) = x)}$$
Modeling $\xi(Z)$

- Choose lower-dimension $\psi(Z) = \Pr(G = g | Z = z, D = 0)$ by machine learning or linear combination approaches.

- $\xi$ is an unknown function and a nonparametric function may be desired (e.g., B-splines)

- Theoretical justification for $\hat{\beta}$ in the presence of nonparametric function $\xi(\cdot)$ with estimated $\psi(Z)$

---

In practice, $\Psi(Z)$ are the leading principal components and $\xi(\cdot)$ is a linear function.

Potential pitfalls in the principal components analysis
- SNPs are correlated
- Individuals may be related

Including individuals of known geographic origin can help interpretation.

Outliers distort (smaller) eigenvectors. Analysis should be performed twice: once to detect outliers and a second time to infer structure in the remaining samples.
Summary

- Principal components can be used to visualize population substructure and as covariates in association analysis.

- Even if the interest is in the single variant association looking at all of the variants can help identify potential confounding issues (e.g., batch effect, population substructure).
Effect Size Estimation

- **Model**
  \[ g\{E(Y_i)\} = X'_i \alpha + G_i \beta. \]

- If \( y \) is a continuous trait: linear regression model
  \[ Y_i = X'_i \alpha + G_i \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \]

- \( X_i = (1, x_{i1}, \cdots, x_{iq}) \): covariates including the intercept.

- \( G_i \): Genotype value.
Likelihood: Estimation of $\beta$

- **Likelihood**

$$ L(\beta, \alpha, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{(Y - \tilde{X}\gamma)'(Y - \tilde{X}\gamma)}{2\sigma^2} \right\} $$

- $\gamma = (\alpha, \beta)$
- $\tilde{X} = [X, G]$
Estimation of $\beta$

- **Score functions**
  \[
  S(\gamma) = \frac{\partial \log L}{\partial \gamma} = \frac{1}{\sigma^2} \tilde{X}'(Y - \tilde{X}\gamma)
  \]
  \[
  S(\sigma^2) = \frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{\sigma^4}(Y - \tilde{X}\gamma)'(Y - \tilde{X}\gamma)
  \]

- **Fisher information**
  \[
  I(\gamma, \sigma^2) = \frac{1}{\sigma^2} \begin{pmatrix}
    \tilde{X}'\tilde{X} & 0 \\
    0 & \frac{n}{2\sigma^2}
  \end{pmatrix}
  \]
Estimation of $\beta$

- MLE of $\hat{\gamma} = (\hat{\alpha}, \hat{\beta}) = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$

$$\tilde{\gamma} \sim N(\gamma, \sigma^2(\tilde{X}'\tilde{X})^{-1})$$

- Unbiased estimators of $\sigma^2$

$$\hat{\sigma}^2 = (Y - \tilde{X}\tilde{\gamma})'(Y - \tilde{X}\tilde{\gamma})/(n - q - 1)$$
Estimation of $\beta$

- If $Y$ is a binary trait, logistic regression model
  \[
  \log \left\{ \frac{\Pr(Y = 1)}{\Pr(Y = 0)} \right\} = X_i' \alpha + G_i \beta
  \]
  or
  \[
  \Pr(Y = 1) = \frac{\exp(X_i' \alpha + G_i \beta)}{1 + \exp(X_i' \alpha + G_i \beta)}
  \]

- MLE of $(\alpha, \beta)$ by maximizing
  \[
  L = \prod_{i=1}^{n} \left\{ \Pr(Y_i = 1) \right\}^{Y_i} \left\{ \Pr(Y_i = 0) \right\}^{1-Y_i}
  \]
  \[
  = \prod_{i=1}^{n} \frac{\exp\{(X_i' \alpha + G_i \beta) Y_i\}}{1 + \exp(X_i' \alpha + G_i \beta)}
  \]
Winner’s Curse

- ‘Winner’s Curse’ = the phenomenon whereby winners at competitive auctions are likely to pay in excess of the item’s worth
- In genetic association studies the winner’s curse is the phenomenon that the disease risk of a newly identified genetic association is overestimated
- It occurs particularly when the statistical power of original study is not sufficient, which is common in GWAS because they are often underpowered to detect small genetic effects at a stringent genome-wide significant level.
- The consequence is that the sample size required for confirmatory study will be underestimated, resulting failure of replication study to corroborate the association.
Bias

- Asymptotic distribution for $\hat{\beta}$ after selection $|\hat{\beta}/\hat{\sigma}| > c$, where $c$ is a cutpoint selected to control the family wise error rate

$$f_{\hat{\beta}|\{|\hat{\beta}|>c\hat{\sigma}\}}(x) = \frac{1}{\sigma} \frac{\phi(\frac{x-\beta}{\sigma})}{\Phi(\frac{\beta}{\sigma} - c) + \Phi(-\frac{\beta}{\sigma} - c)} I\left(\frac{|x|}{\sigma} \geq c\right).$$

- $\phi$: standard normal density.
- $\Phi$: standard cumulant density function

- The expectation of $\hat{\beta}$ for the selected SNP is

$$E(\hat{\beta}) = \beta + \sigma \frac{\phi(\frac{\beta}{\sigma} - c) + \phi(-\frac{\beta}{\sigma} - c)}{\Phi(\frac{\beta}{\sigma} - c) + \Phi(-\frac{\beta}{\sigma} - c)}$$
Bias
Solution

- Large GWAS (or a meta-analysis).
- An independent replication study.
- Statistical methods to correct the bias of estimators and confidence intervals.
Resampling Technique

- Bootstrap method
  - Randomly draw samples with replacement, mimic the original procedure to identify markers, and estimate, $\hat{\beta}_D$
  - The ‘validation’ sample consists of subjects that are not selected in the bootstrap sample, estimate, $\hat{\beta}_E$
  - $\hat{\text{Bias}} = \hat{\beta}_D - \hat{\beta}_E$
  - A more refined resampling-based estimator that accounts for negative covariance between training and validation samples and the difference in allele frequency can be found in Faye et al. (2011, Stat in Med, 30:1898–1912)

---

The maximum likelihood estimator

$$\hat{\beta}_{\text{MLE}} = \arg\max_{\beta} f_{\beta | \{|\hat{\beta}| > c\hat{\sigma}\}}(\hat{\beta}; \beta)$$
Adjusted Confidence Interval (CI)

- The likelihood ratio test
  \[ T = 2\{\log L(\hat{\beta}_{MLE}) - \log L(\beta_0)\} \]

- A 95% CI for \( \hat{\beta}_{MLE} \) consists of those values of \( \beta \) for which the test is non-significant at significance level 0.05.
Adjusted Confidence Interval (CI)

- $T \leq 3.84 = \chi^2_{1,0.95}$
- Hence, the CI consists of the $\beta_0$ values for which

$$\log L(\beta_0) \geq \log L(\hat{\beta}_{MLE}) - \frac{3.84}{2}$$

$$= \log L(\hat{\beta}_{MLE}) - 1.92$$

![Graph showing the confidence interval for a parameter $\lambda$. The log-likelihood function is plotted against $\lambda$, with the confidence interval marked by dashed vertical lines.]
Practice

- $\hat{\beta}$ has upward bias; however, $\hat{\beta}_{\text{MLE}}$ tends to overcorrect and to underestimate $\beta$.

- Combine these two estimators with a weight

$$\hat{\beta}_w = \omega \hat{\beta} + (1 - \omega) \hat{\beta}_{\text{MLE}}$$

$$\omega = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + (\hat{\beta} - \hat{\beta}_{\text{MLE}})^2}$$

- The lower bound of CI

$$\hat{\beta}_{\omega;\alpha/2} = \omega_{\alpha/2} \beta_{\alpha/2} + (1 - \omega_{\alpha/2}) \beta_{\text{MLE};\alpha/2}$$

- The upper bound of CI

$$\hat{\beta}_{\omega;1-\alpha/2} = \omega_{1-\alpha/2} \beta_{1-\alpha/2} + (1 - \omega_{1-\alpha/2}) \beta_{\text{MLE};1-\alpha/2}$$
The discovery set includes 4,878 cases and 4,914 controls, and the replication set includes 13,114 cases and 14,304 controls.

Summary odds ratios and p-values for the SNPs showing association with Colorectal Cancer

<table>
<thead>
<tr>
<th>rsID</th>
<th>Gene</th>
<th>Allele</th>
<th>Chr</th>
<th>Position</th>
<th>Trend p-value</th>
<th>Per Allele OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10411210</td>
<td>RHPN2</td>
<td>C/T</td>
<td>19</td>
<td>38224140</td>
<td>2.0 × 10⁻⁷ 6.9 × 10⁻⁴</td>
<td>0.79 (0.72-0.86) 0.81 (0.72-0.95) 0.90 (0.85-0.96) 0.24 (0.84-0.94)</td>
</tr>
<tr>
<td>rs961253</td>
<td>C/A</td>
<td>7.8 × 10⁻⁷ 3.4 × 10⁻⁵</td>
<td>1.13 (1.08-1.19) 1.10 (1.00-1.18) 1.11 (1.06-1.17)</td>
<td>0.87 (1.06-1.15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs355527</td>
<td>G/A</td>
<td>7.8 × 10⁻⁷ 3.4 × 10⁻⁵</td>
<td>1.13 (1.08-1.19) 1.10 (1.00-1.18) 1.11 (1.06-1.17)</td>
<td>0.87 (1.06-1.15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs9929218</td>
<td>CDH1</td>
<td>G/A</td>
<td>16</td>
<td>67378447</td>
<td>1.1 × 10⁻⁶ 1.5 × 10⁻⁴</td>
<td>0.88 (0.84-0.93) 0.91 (0.84-1.00) 0.93 (0.90-0.97) 0.71 (0.90-0.96)</td>
</tr>
<tr>
<td>rs4444235</td>
<td>BMP4</td>
<td>T/C</td>
<td>14</td>
<td>53480669</td>
<td>5.6 × 10⁻⁶ 1.8 × 10⁻⁴</td>
<td>1.12 (1.07-1.18) 1.03 (0.99-1.17) 1.10 (1.05-1.16) 0.42 (1.04-1.14)</td>
</tr>
<tr>
<td>rs1862748</td>
<td>CDH1</td>
<td>C/T</td>
<td>16</td>
<td>67390444</td>
<td>8.5 × 10⁻⁷ 1.5 × 10⁻⁴</td>
<td>0.88 (0.84-0.93) 0.91 (0.84-1.00) 0.93 (0.90-0.97) 0.64 (0.90-0.96)</td>
</tr>
<tr>
<td>rs4951291</td>
<td>G/A</td>
<td>1</td>
<td>202273161</td>
<td>6.6 × 10⁻⁶ 5.7 × 10⁻¹</td>
<td>0.85 (0.79-0.91) 0.97 (0.80-1.01) 1.02 (0.95-1.09) 0.35 (0.95-1.01)</td>
<td></td>
</tr>
<tr>
<td>rs7259371</td>
<td>RHPN2</td>
<td>G/A</td>
<td>19</td>
<td>38226481</td>
<td>3.4 × 10⁻⁶ 2.1 × 10⁻³</td>
<td>0.86 (0.81-0.92) 0.93 (0.81-1.01) 0.91 (0.86-0.97) 0.84 (0.86-0.97)</td>
</tr>
<tr>
<td>rs4951039</td>
<td>A/G</td>
<td>1</td>
<td>202273220</td>
<td>6.6 × 10⁻⁶ 5.2 × 10⁻²</td>
<td>0.85 (0.79-0.91) 0.97 (0.80-1.01) 1.09 (1.00-1.19) 0.03 (0.96-1.01)</td>
<td></td>
</tr>
</tbody>
</table>

a Major/minor allele;
b From NCBI build 139;
c Significance level (p-value) for testing equality of bias-adjusted and replication odds ratios.
Other Likelihood-based Estimator

- MLE $\hat{\beta}_{\text{MLE}}$ provides no guarantee of unbiasedness or efficiency, because large-sample assumptions are already applied to $\hat{\beta}$ when constructing the conditional likelihood.

- An alternative estimator

$$\tilde{\beta} = \int \beta f^*_\{|\hat{\beta}| > c\hat{\sigma}\} (\hat{\beta}; \beta) d\beta$$

- $\tilde{\beta}$ is a posterior mean with a flat prior on $\beta$ and has favorable MSE properties

- Averaging $\tilde{\beta}$ and $\hat{\beta}_{\text{MLE}}$ to balance out the strengths of the two estimators

---

Summary

- Single variant association
- Use genome-wide SNPs to account for confounding (population substructure)
- Estimation of effect size and winner's curse
Recommended Reading


