Today's Outline

- Single variant association analysis.
- Single variant association analysis for genome-wide association studies (GWAS).
- Effect size estimation and winner's curse.

Single Variant Test

- In GWAS, single variant test is the most popular approach to investigating associations.
- Y_i: outcomes for $i = 1, \ldots, n$
- $X_i' = (1, x_{i1}, \cdots, x_{iq})$: 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:

$$\frac{\widehat{\beta}}{s.e.(\widehat{\beta})} \sim N(0,1)$$

  - Score test:

$$\left[\frac{\partial}{\partial\beta}\log L(\beta,\widehat{\alpha}_0)\right]\Big|_{\beta=0} \mathsf{I}(\beta=0|\widehat{\alpha}_0)^{-1} \left[\frac{\partial}{\partial\beta}\log L(\beta,\widehat{\alpha}_0)\right]\Big|_{\beta=0} \sim \chi^2_1$$

$$\mathsf{I}(\beta=0|\widehat{\alpha}_0) = \left\{\mathsf{I}_{\beta\beta} - \mathsf{I}_{\beta\alpha}\mathsf{I}^{-1}_{\alpha\alpha}\mathsf{I}_{\alpha\beta}\right\}\Big|_{\beta=0,\widehat{\alpha}_0}$$

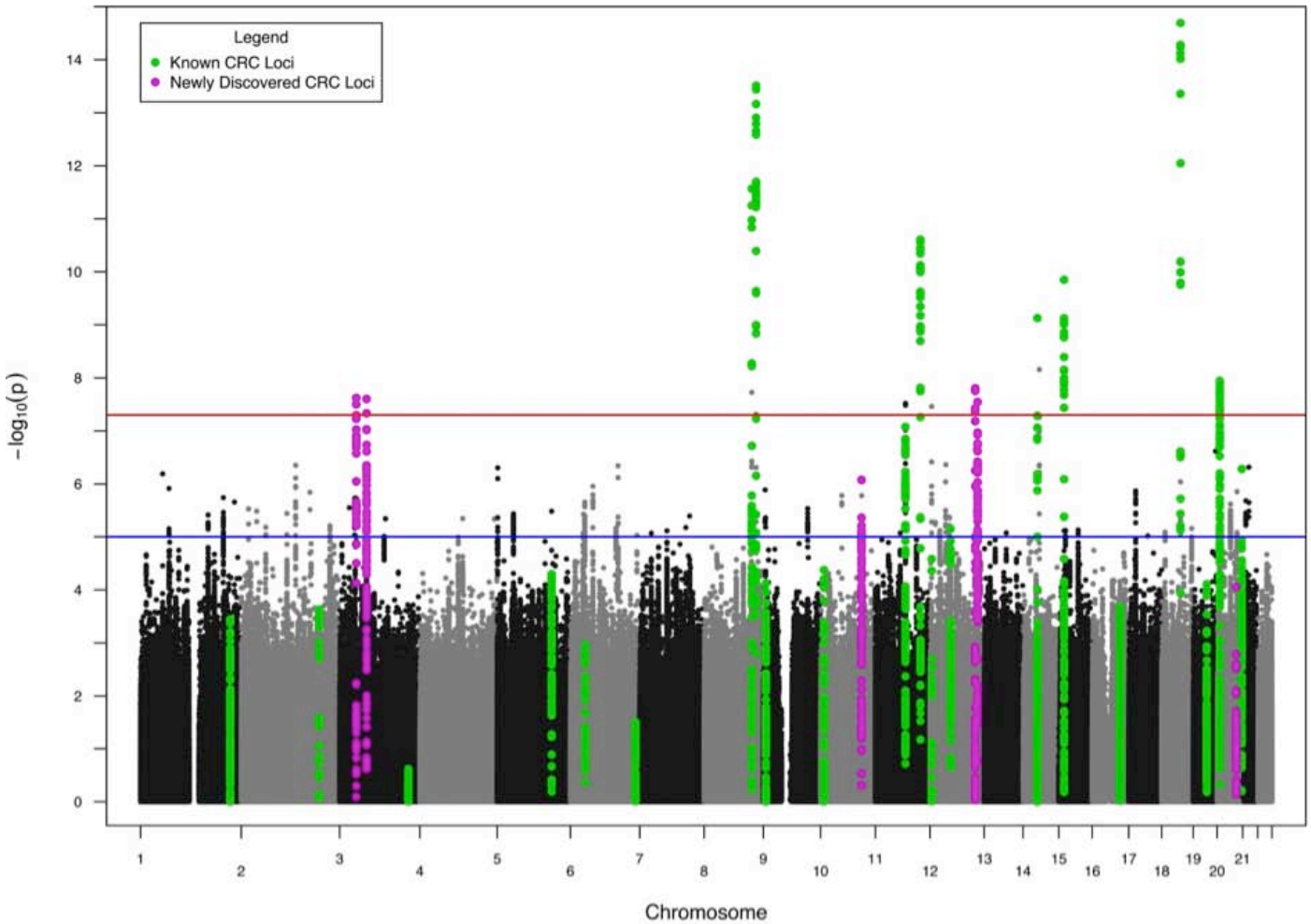
  - Likelihood ratio (LR) test:

$$2\{\log L(\widehat{\alpha},\widehat{\beta}) - \log L(\widehat{\alpha}_0,0)\} \sim \chi^2_1$$

- 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 ($n \approx 40,000$).



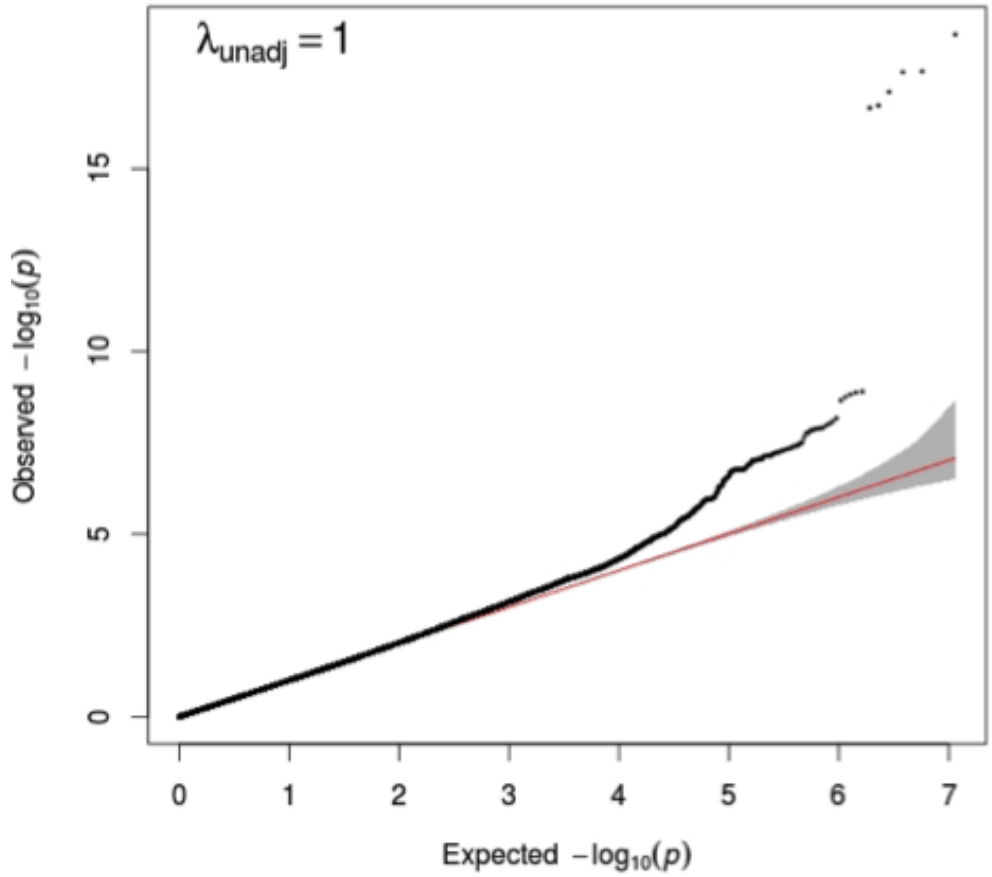
Schumacher FR et al. (2015). GWAS of colorectal cancer identifies six new susceptibility loci. *Nat Commun* DOI:10.1038

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

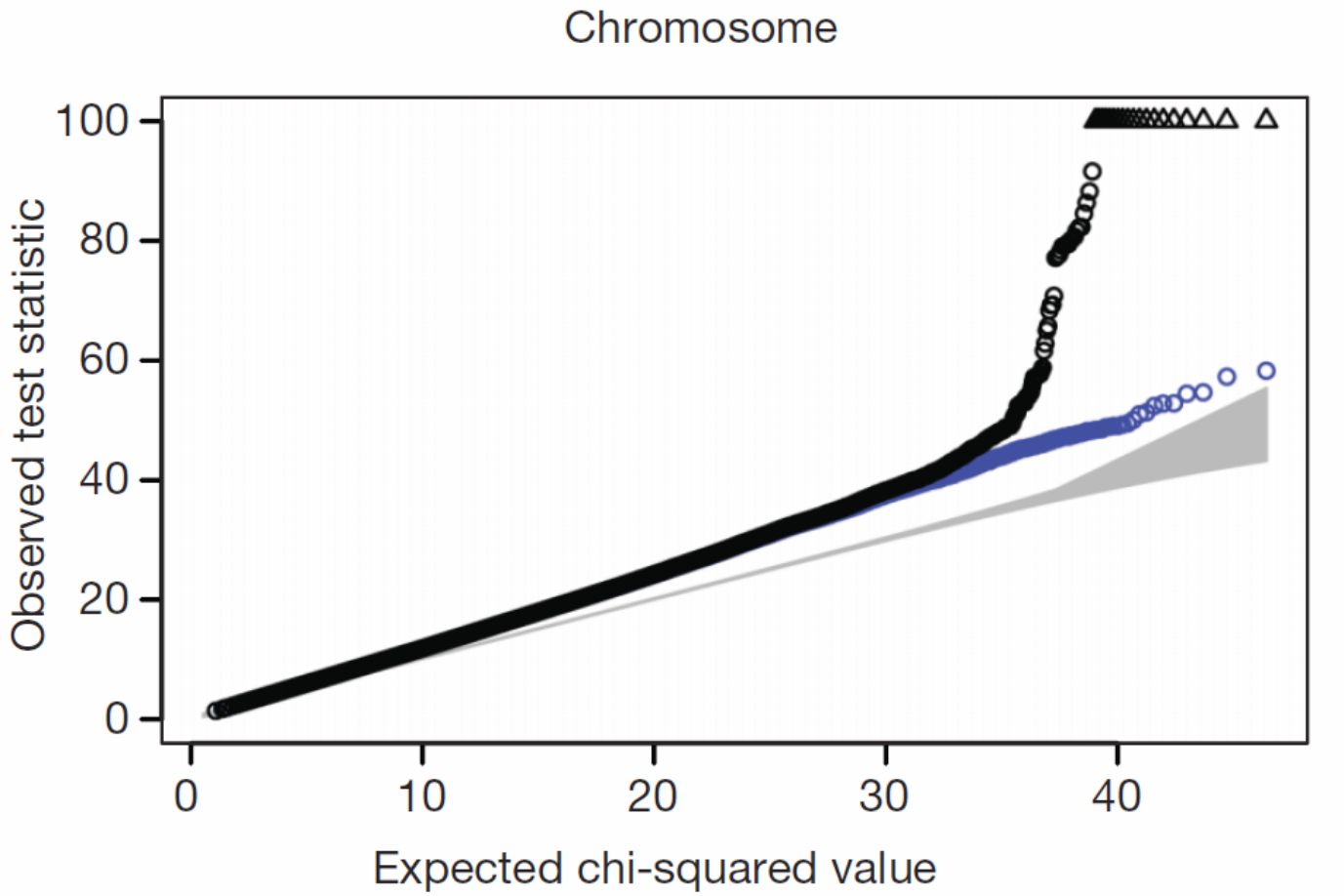
Detecting Stratification

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



Detection Stratification

- QQ plot shows stratification



Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447.7145: 661-678.

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 $>$ 100 kb)
- Compute χ^2 for each marker
- Inflation $\lambda =$ Median observed $\chi^2/0.456$
- Adjust statistic by

$$\chi^2_{\text{fair}} = \chi^2_{\text{observed}}/\lambda$$

- λ 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?

Delvin B & Roeder K. (1999) Genomic control for association studies. *Biometrics* 55:997–1004.

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.

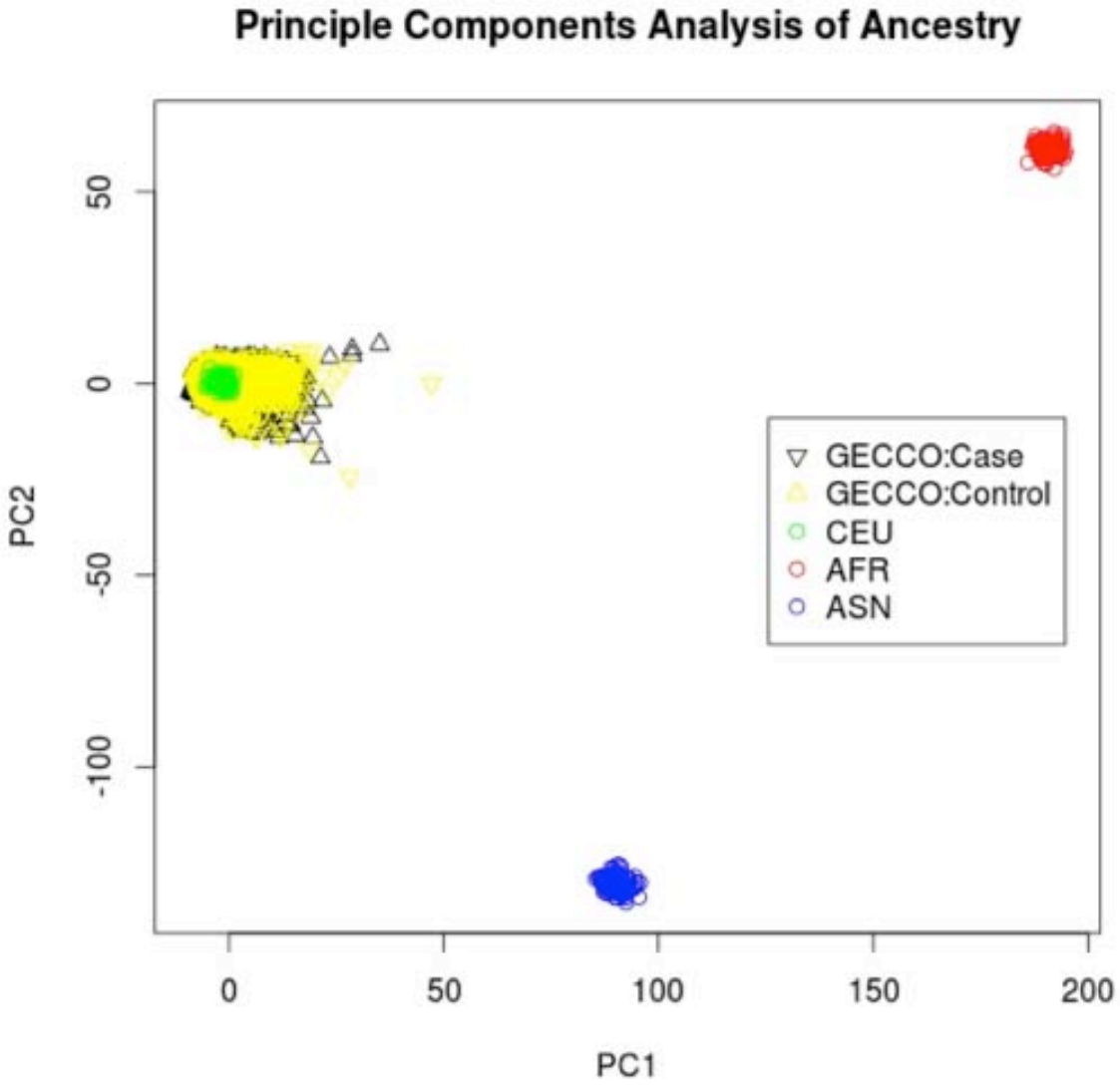
However,

- 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.

Pritchard JK, Stephens M, Rosenberg NA & Donnelly P. (2000) Association mapping in structured populations. *Am J Hum Genet*, 67(1), 170–181.

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

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

- β is parameter of interest, but not identifiable because U is not observed.

Statistical Framework

- Marginal model

$$\begin{aligned}\frac{\Pr(Y=1|G,Z)}{\Pr(Y=0|G,Z)} &= \frac{\Pr(Y=1,G,Z)}{\Pr(Y=0,G,Z)} \\ &= \int \frac{\Pr(Y=1,G,Z,u)}{\Pr(Y=0,G,Z,u)} \frac{\Pr(Y=0,G,Z,u)}{\Pr(Y=0,G,Z)} du \\ &= \exp(G\beta) \int \exp\{\gamma^*(u,z)\} P(u|Y=0,G,Z) du\end{aligned}$$

- 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

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

Statistical Framework

$$\text{logit}\{\Pr(Y = 1|G, Z)\} = G\beta + \xi(Z) \tag{1}$$

- 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 dissolves the link between U and G such that $U \perp G$ for each stratum of Z 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

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

- Sketch of proof:

$$\begin{aligned}
&\frac{\Pr(Y = 1, G = g, \Psi(Z) = x)}{\Pr(Y = 0, G = g, \Psi(Z) = x)} \\
&= \frac{\int_{u,z:\psi(z)=x} \Pr(Y = 1, G = g, Z, u)dZdu}{\Pr(Y = 0, G = g, \Psi(Z) = x)} \\
&= \frac{\int_{u,z:\psi(z)=x} \exp(G\beta + \gamma(u, Z))\Pr(Y = 0, G = g, Z)dZdu}{\Pr(G = g|Y = 0, \Psi(Z) = x)\Pr(Y = 0, \Psi(Z) = x)} \\
&= \exp(G\beta)\frac{\int_{u,z:\psi(z)=x} \exp(\gamma(u, Z))\Pr(Y = 0, Z)dZdu}{\Pr(Y = 0, \Psi(Z) = x)}
\end{aligned}$$

Modeling $\xi(Z)$

- Choose lower-dimension $\Psi(Z) = \Pr(G = g|Z = z, D = 0)$ by machine learning or linear combination approaches.
- ξ is an unknown function and a nonparametric function may be desired (e.g., B-splines)
- Theoretical justification for $\widehat{\beta}$ in the presence of nonparametric function $\xi(\cdot)$ with estimated $\Psi(Z)$

Lin DY & Zeng D (2011) Correcting for Population Stratification in Genomewide Association Studies. *J Am Stat Assoc* 106:997–1008.

Practice

- 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 β

- Likelihood

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

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

Estimation of β

- Score functions

$$S(\gamma) = \frac{\partial \log L}{\partial \gamma} = \frac{1}{\sigma^2}\widetilde{X}'(Y - \widetilde{X}\gamma)$$

$$S(\sigma^2) = \frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{\sigma^4}(Y - \widetilde{X}\gamma)'(Y - \widetilde{X}\gamma)$$

- Fisher information

$$I(\gamma, \sigma^2) = \frac{1}{\sigma^2}\begin{pmatrix} \widetilde{X}'\widetilde{X} & 0 \\ 0 & \frac{n}{2\sigma^2} \end{pmatrix}$$

Estimation of β

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

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

- Unbiased estimators of σ^2

$$\widehat{\sigma}^2 = (Y - \widetilde{X}\widehat{\gamma})'(Y - \widetilde{X}\widehat{\gamma})/(n - q - 1)$$

Estimation of β

- 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 (α, β) by maximizing

$$\begin{aligned} L &= \prod_{i=1}^{n} \{\Pr(Y_i=1)\}^{Y_i}\{\Pr(Y_i=0)\}^{1-Y_i} \\ &= \prod_{i=1}^{n} \frac{\exp\{(X_i'\alpha + G_i\beta)Y_i\}}{1+\exp(X_i'\alpha + G_i\beta)} \end{aligned}$$

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

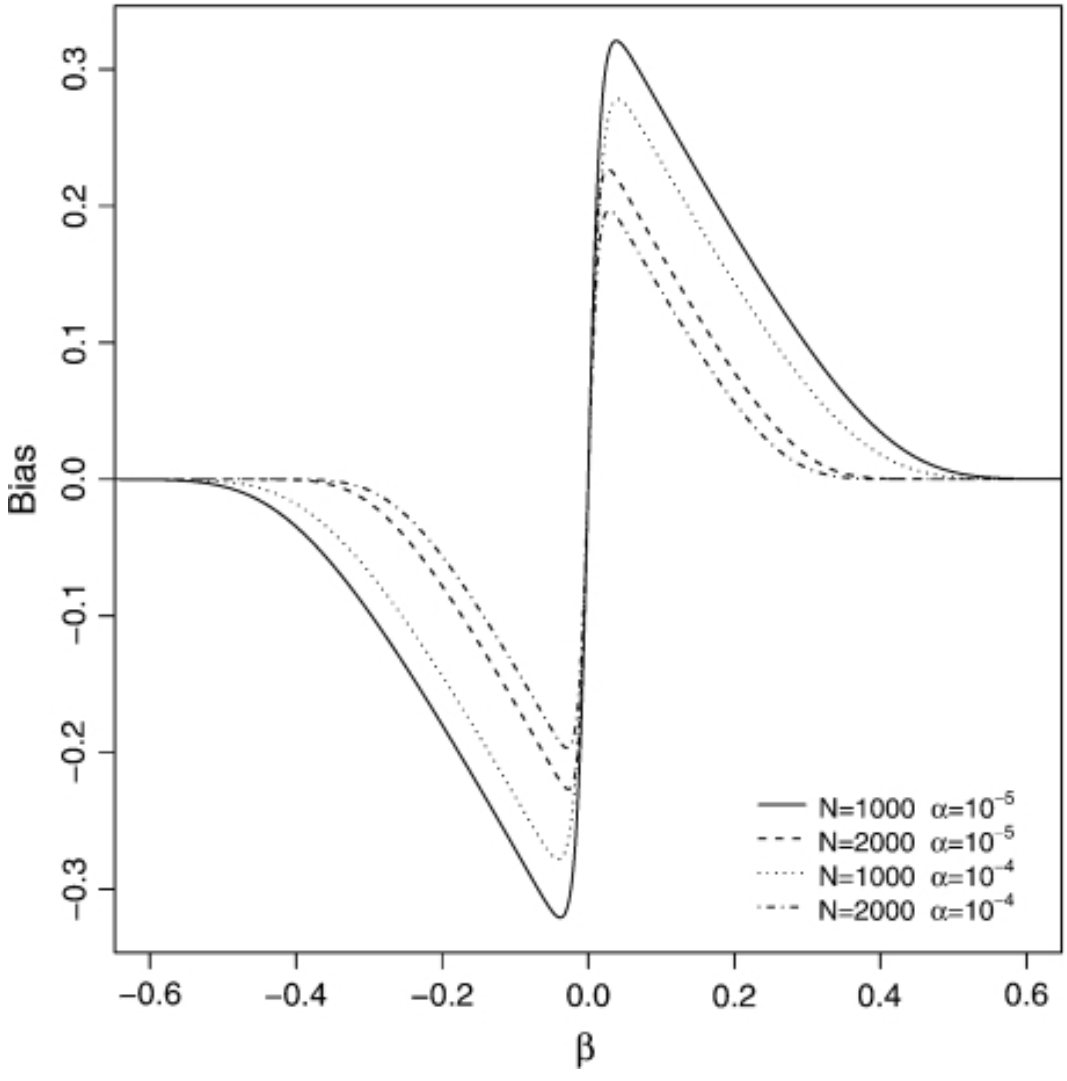
- Asymptotic distribution for $\widehat{\beta}$ after selection $|\widehat{\beta}/\widehat{\sigma}| > c$, where c is a cutpoint selected to control the family wise error rate

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

  - ϕ: standard normal density.
  - Φ: standard cumulant density function
- The expectation of $\widehat{\beta}$ for the selected SNP is

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

Bias


Bias
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
β
N=1000 α=10^-5
N=2000 α=10^-5
N=1000 α=10^-4
N=2000 α=10^-4

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, $\widehat{\beta}_D$
  - The 'validation' sample consists of subjects that are not selected in the bootstrap sample, estimate, $\widehat{\beta}_E$
  - $\widehat{Bias} = \overline{\widehat{\beta}_D - \widehat{\beta}_E}$
- 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)

Sun L, & Bull SB. (2005) Reduction of selection bias in genomewide studies by resampling. *Gen Epidem* 28(4):352–367.

Bias Correction Method

- The maximum likelihood estimator

$$\widehat{\beta}_{\text{MLE}} = \underset{\beta}{\text{argmax}} \; f_{\widehat{\beta}|\{|\widehat{\beta}|>c\widehat{\sigma}\}}(\widehat{\beta};\beta)$$

Adjusted Confidence Interval (CI)

- The likelihood ratio test

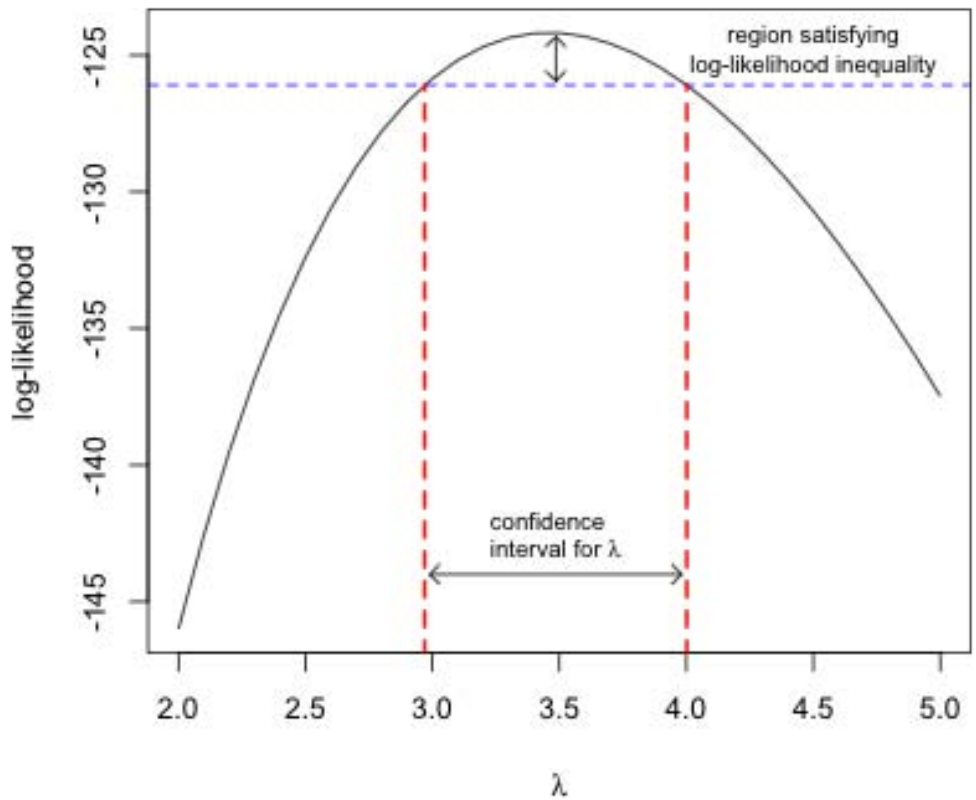
$$T = 2\{\log L(\widehat{\beta}_{\text{MLE}}) - \log L(\beta_0)\}$$

- A 95% CI for $\widehat{\beta}_{\text{MLE}}$ consists of those values of β for which the test is non-significant at significance level 0.05.

Adjusted Confidence Interval (CI)

- $T \leq 3.84 = \chi^2_{1,0.95}$
- Henc, the CI consists of the β_0 values for which

$$\begin{aligned} \log L(\beta_0) &\geq \log L(\widehat{\beta}_{\text{MLE}}) - 3.84/2 \\ &= \log L(\widehat{\beta}_{\text{MLE}}) - 1.92 \end{aligned}$$



Practice

- $\widehat{\beta}$ has upward bias; however, $\widehat{\beta}_{\text{MLE}}$ tends to overcorrect and to underestimate β.
- Combine these two estimators with a weight

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

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

- The lower bound of CI

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

- The upper bound of CI

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

Example: Colorectal Cancer

- 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

					Trend p-value		Per Allele OR (95% CI)				
rsID	Gene	Allele[a]	Chr	Position[b]	Stages 1&2	Replication	Unadjusted	Adjusted	Replication	P^{c}_{het}	Combined
rs10411210	RHPN2	C/T	19	38224140	2.0×10^{-7}	6.9×10^{-4}	0.79 (0.72-0.86)	0.81 (0.72-0.95)	0.90 (0.85-0.96)	0.24	0.89 (0.84-0.94)
rs961253		C/A	20	6352281	7.8×10^{-7}	3.4×10^{-5}	1.13 (1.08-1.19)	1.10 (1.00-1.18)	1.11 (1.06-1.17)	0.87	1.11 (1.06-1.15)
rs355527		G/A	20	6336068	7.8×10^{-7}	3.4×10^{-5}	1.13 (1.08-1.19)	1.10 (1.00-1.18)	1.11 (1.06-1.17)	0.87	1.11 (1.06-1.15)
rs9929218	CDH1	G/A	16	67378447	1.1×10^{-6}	1.5×10^{-4}	0.88 (0.84-0.93)	0.91 (0.84-1.00)	0.93 (0.90-0.97)	0.71	0.93 (0.90-0.96)
rs4444235	BMP4	T/C	14	53480669	5.6×10^{-6}	1.8×10^{-4}	1.12 (1.07-1.18)	1.03 (0.99-1.17)	1.10 (1.05-1.16)	0.42	1.09 (1.04-1.14)
rs1862748	CDH1	C/T	16	67390444	8.5×10^{-7}	1.5×10^{-4}	0.88 (0.84-0.93)	0.91 (0.84-1.00)	0.93 (0.90-0.97)	0.64	0.93 (0.90-0.96)
rs4951291		G/A	1	202273161	6.6×10^{-6}	5.7×10^{-1}	0.85 (0.79-0.91)	0.97 (0.80-1.01)	1.02 (0.95-1.09)	0.35	0.99 (0.95-1.01)
rs7259371	RHPN2	G/A	19	38226481	3.4×10^{-6}	2.1×10^{-3}	0.86 (0.81-0.92)	0.93 (0.81-1.01)	0.91 (0.86-0.97)	0.84	0.91 (0.86-0.97)
rs4951039		A/G	1	202273220	6.6×10^{-6}	5.2×10^{-2}	0.85 (0.79-0.91)	0.97 (0.80-1.01)	1.09 (1.00-1.19)	0.03	0.99 (0.96-1.01)

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

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

- $\widetilde{\beta}$ is a posterior mean with a flat prior on β and has favorable MSE properties
- Averaging $\widetilde{\beta}$ and $\widehat{\beta}_{\text{MLE}}$ to balance out the strengths of the two estimators

---

Ghosh et al. (2008) Estimating Odds Ratios in Genome Scans: An Approximate Conditional Likelihood Approach. AJHG 82: 1064–1074

Summary

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

Recommended Reading

- Devlin B & Roeder K (1999) Genomic control for association studies. *Biometrics* 55(4):997–1004.
- Lin DY & Zeng D (2011) Correcting for Population Stratification in Genomewide Association Studies, *J Am Statist Assoc* 106:997–1008.
- Pritchard JK, Stephens M, Rosenberg NA, & Donnelly P. (2000) Association mapping in structured populations. *Am J Hum Genet* 67(1):170–181.
- Zhong H & Prentice RL (2008) Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* 9(4):621–634.