Lecture 8: Advanced Topics of GWAS

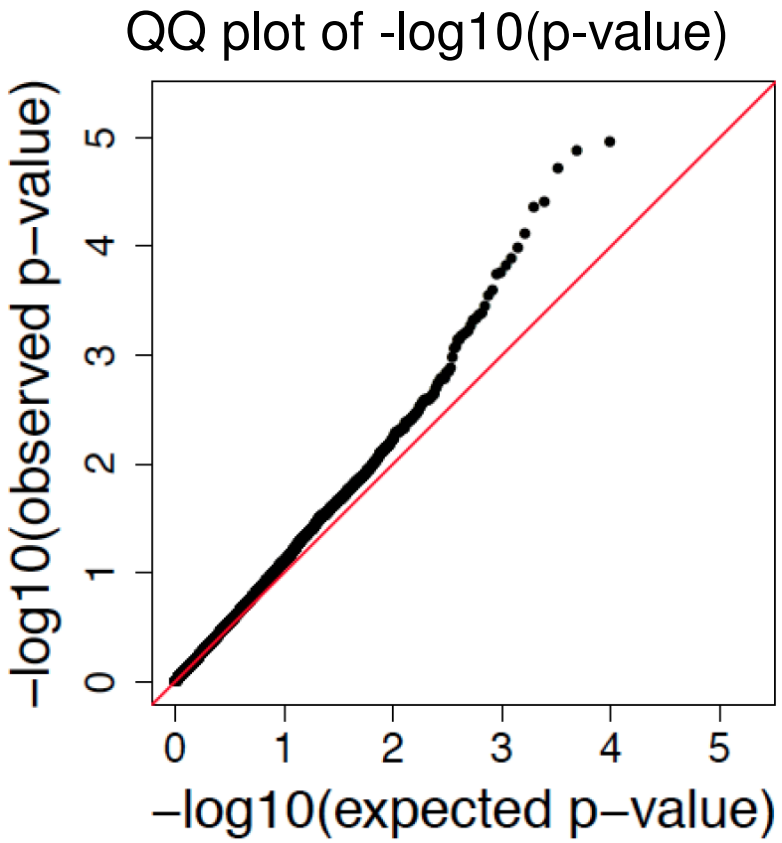
Advanced Topics of GWAS

- Account for population stratification
  - Genomic control factor
  - Structured association study
  - PCA
- Multiple testing
- Conditional analysis
- Meta-analysis
- Rare variant association study

Population Stratification

Population stratification is a confounding factor that must be accounted for in GWAS.

If not accounted for
$\Rightarrow$ the number of false-positive findings can be notably higher than would be expected (i.e., inflated type I error).



Adjust for Population Stratification

- Family-based association analysis: TDT, FBAT (previous lecture)
- Genomic control factor, *Devlin and Roeder, 1999, Biometrics*
- Principle component analysis (PCA), *Price et al. 2006, Nature Genetics*

Genomic Control Factor

Genomic Control Factor is used to control for systematic inflation of type I error.

The idea is that the statistic T is inflated by an inflation factor λ (i.e., genomic control factor) so that

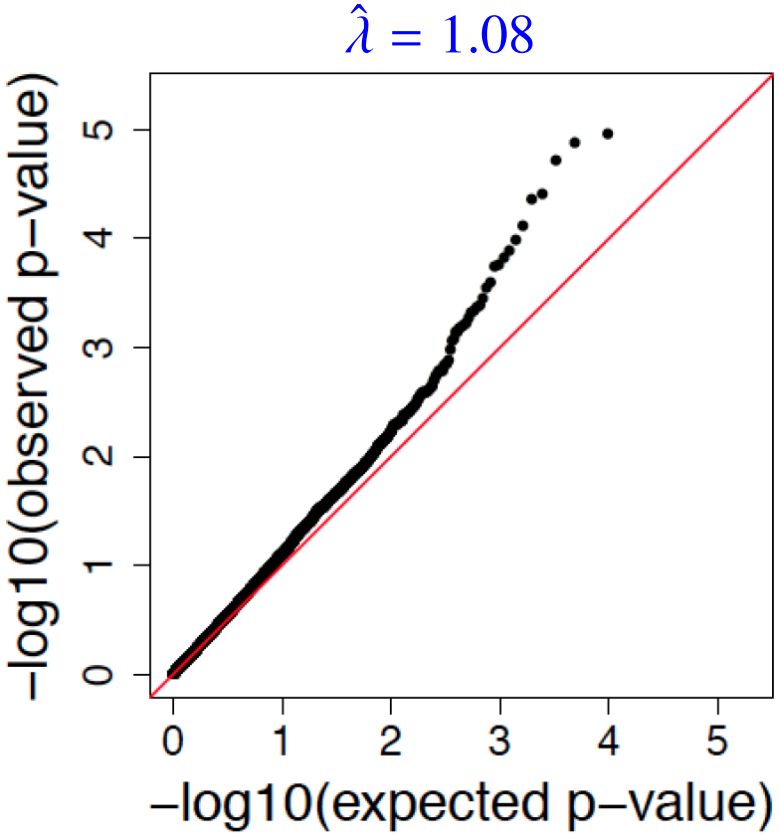
$$T \sim \lambda \chi_1^2$$

where λ can be estimated by

$$\hat{\lambda} = \text{median}(T_1, T_2, \ldots, T_M)/0.456$$

- M is the number of independent tests, though in practice all tests are included.
- The denominator is the median of χ_1^2 distribution.
- $\hat{\lambda}$ should be 1 under H_0.

Genomic Control Factor



Divide all chi-square test statistics T by the estimated inflation (GC) factor to get corrected test statistics

$$T/\hat{\lambda} \sim \chi^2_1$$

under H_0 of no association.

Genomic Control Factor

Limitation:

- Genomic control corrects for stratification by adjusting association statistics at each marker by a uniform overall inflation factor.
- However, some markers differ in their allele frequencies across ancestral populations more than others.
- Thus, the uniform adjustment applied by genomic control may be insufficient at markers having unusually strong differentiation across ancestral populations and may be superfluous at markers devoid of such differentiation, leading to a loss in power

Structured Association Study

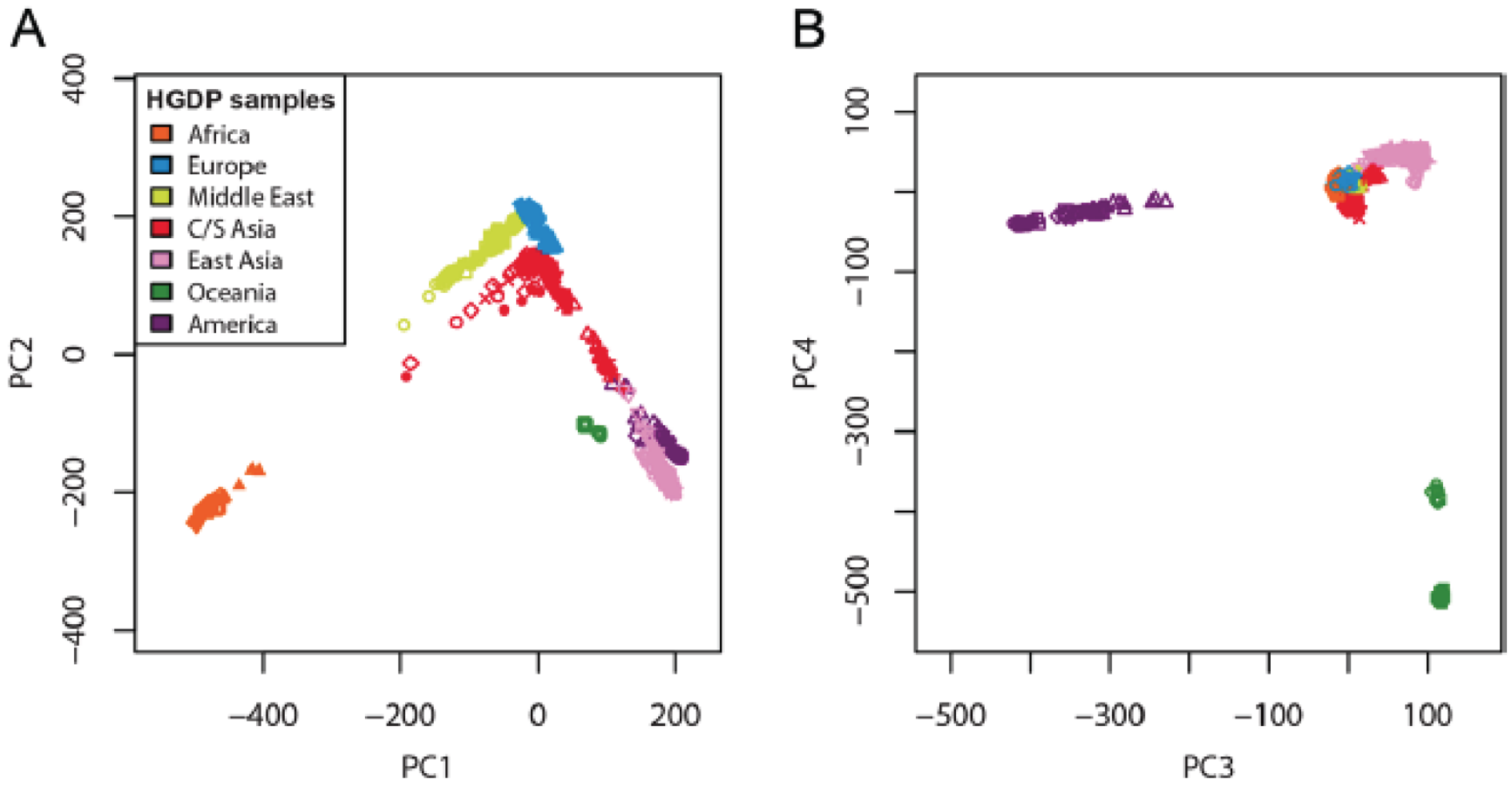
- **Software:** STRUCTURE
- **Idea**: stratified analysis of association by ethnicity to account for population stratification
- **Method**: Use genotype markers to infer the latent population structure, assign the samples to discrete subpopulation clusters, and then aggregates evidence of association within each cluster
- **Limitation**:
  - Discrete subpopulation clusters are not clear in admixed population
  - Not work well for fractional membership
  - Assignments of individuals to clusters are highly sensitive to the number of clusters, which is not well defined

Principal Component Analysis (PCA)

The principal component analysis (PCA) has become one of the standard ways to adjust for population stratification in population-based GWAS.

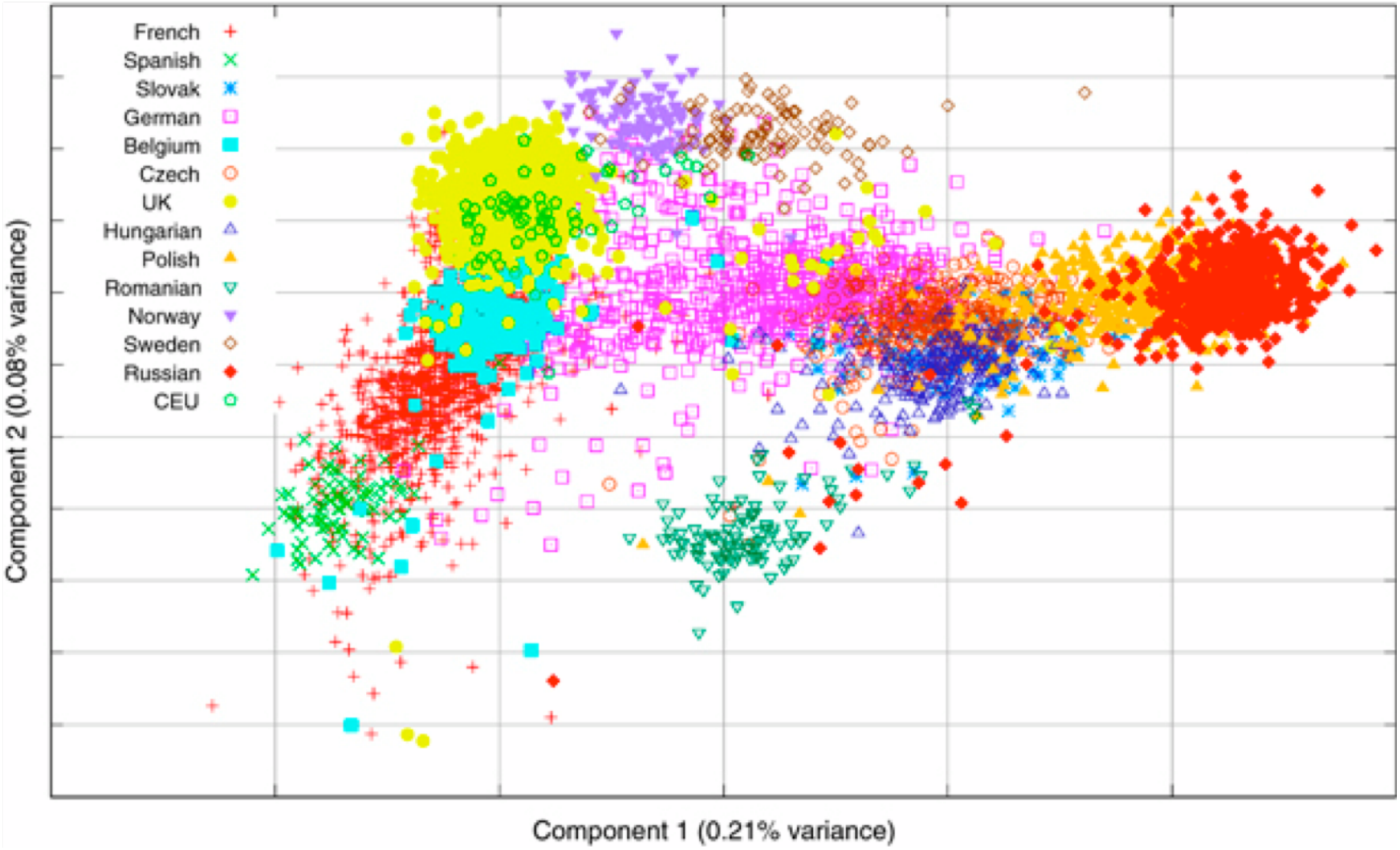
- Apply PCA to genotype data to obtain top principal components (PCs) that explain most genotype data variation
- PLINK can be used to generate top PCs
- Top PCs will reflect sample ancestry
- Include top PCs as covariates in GWAS

Top PCs often reflect geographic distribution (e.g, PC1 - PC4 as follows)



Li et al. Science. 2008; Jakobsson et al. Nature. 2008.

PC1 vs. PC2 among European samples



Heath et al. 2008.

PCA Algorithm

Notation:

- M : Number of SNPs
- N : Number of subjects
- $Z = (z_{ij})$: an $M \times N$ matrix of standardized genotyped coded for the additive model for the ith SNP in the $j$the subject, i.e.,

$$z_{ij} = (X_{ij} - \bar{X}_{i.}) / \sqrt{2\hat{p}_i(1 - \hat{p}_i)}$$

where $\hat{p}_i$ denotes the MAF of the ith SNP.

Algorithm:

- Compute the $N \times N$ variance-covariance matrix as $\Sigma = Z^{\mathrm{T}} Z/(N-1)$.
- Compute the eigenvalue decomposition of Σ: e.g., using R function `eigen`
- Select the top K eigenvalues that are significantly large ($K = 5$ or 10) by a scree plot.
- Include the K eigenvectors (PCs) as additional covariates in the generalized linear regression models that are used for GWAS.

Multiple Testing

For GWAS with millions of single variant tests, the problem of multiple testing must be accounted for.

How does the problem of multiple testing arise?

– M: the number of markers for testing
– $H_0^{(m)}$: no association between the mth SNP and disease, $m = 1, \ldots, M$

In testing single marker, we set our significance level, α':

$$\alpha' = P(\text{reject null hypothesis } H_0^{(m)} \mid H_0^{(m)} \text{ is true})$$

But in testing multiple SNPs, our interest is in the family-wise error rate (FWER):

$$\alpha = P(\text{reject at least one } H_0^{(m)} \mid H_0^{(m)} \text{ is true for all } m)$$

Assume 1) $M = 500,000$, 2) all markers independent, 3) $\alpha' = 0.05$. Consequenctly,

$$\begin{aligned}
\alpha &= 1 - P(\text{not reject any } H_0^{(m)} \mid H_0^{(m)} \text{ is true for all } m) \\
&= 1 - \prod_{m=1}^{M} P(\text{not reject } H_0^{(m)} \mid H_0^{(m)} \text{ is true for all } m) \\
&= 1 - (1 - \alpha')^M \\
&= 1
\end{aligned}$$

Bonferroni correction: An easy and popular approach to adjust the significance-level of each test so as to preserve the overall error rate

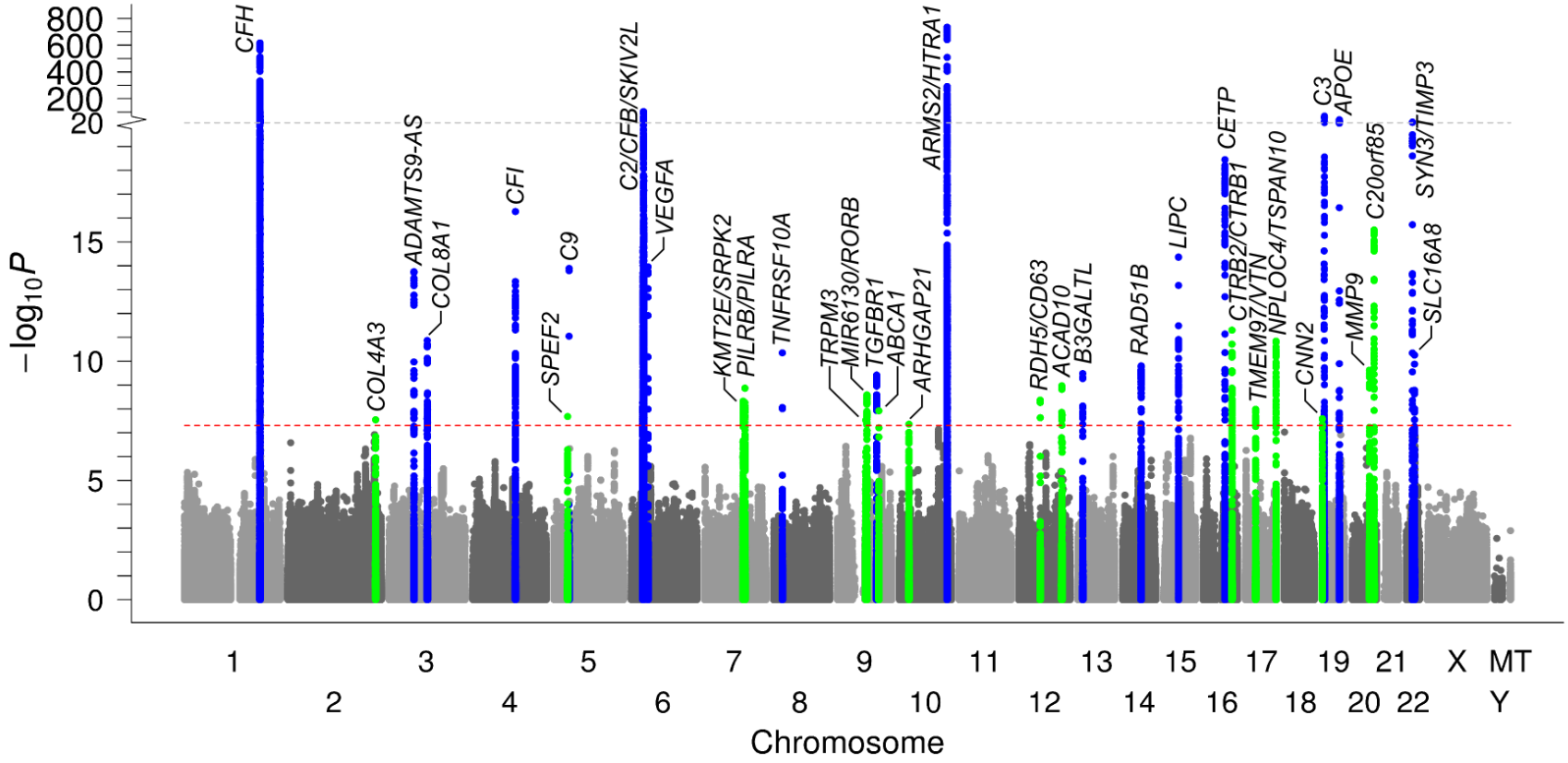
It follows from Boole's inequality:

$$\begin{aligned}\alpha &= P(\text{reject at least one } H_0^{(m)} \mid H_0^{(m)} \text{ is true for all } m)\\ &= P\left(\bigcup_m \{\text{reject } H_0^{(m)} \mid H_0^{(m)} \text{ is true}\}\right)\\ &\leq \sum_m P(\text{reject } H_0^{(m)} \mid H_0^{(m)} \text{ is true}) = M\alpha'\end{aligned}$$

FWER can be kept less than α if each individual test has significance level α/M.

- If $\alpha = 0.01$ and $M = 500,000$, then $\alpha' = 2 \times 10^{-8}$.
- Bonferroni correction is conservative when tests are not independent.
- Use $\alpha = 5 \times 10^{-8}$ to account for independent tests in GWAS.

Recall the Example Manhattan Plot of GWAS Result



18 known AMD loci and **16 novel AMD loci**

Sequential Forward Selection

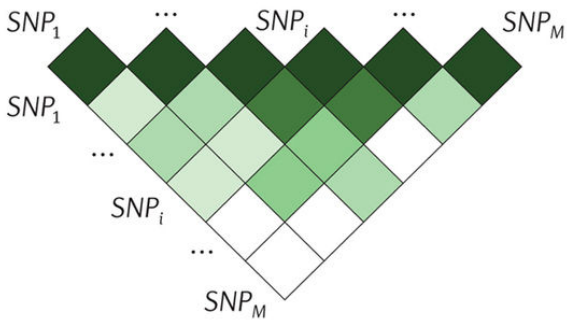
Aim: Within each region of interest, identify all statistically independent variants

1. Select variant with smallest P value ($P < 5x10^{-8}$), write into results file

2. Conduct region-wide association analysis conditioning on variants in results file

3. From the results of 2., if smallest $P < 5x10^{-8}$, select variant write into results file; otherwise stop

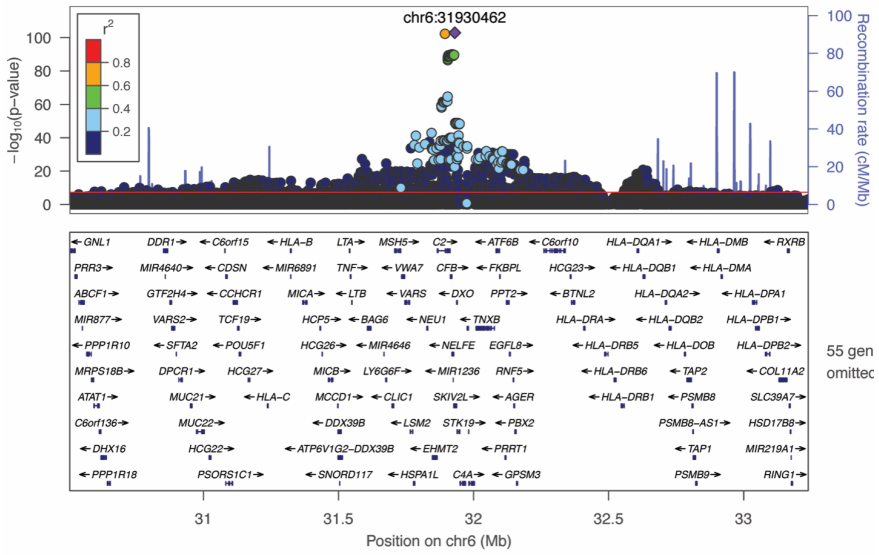
4. Repeat 2. and 3.



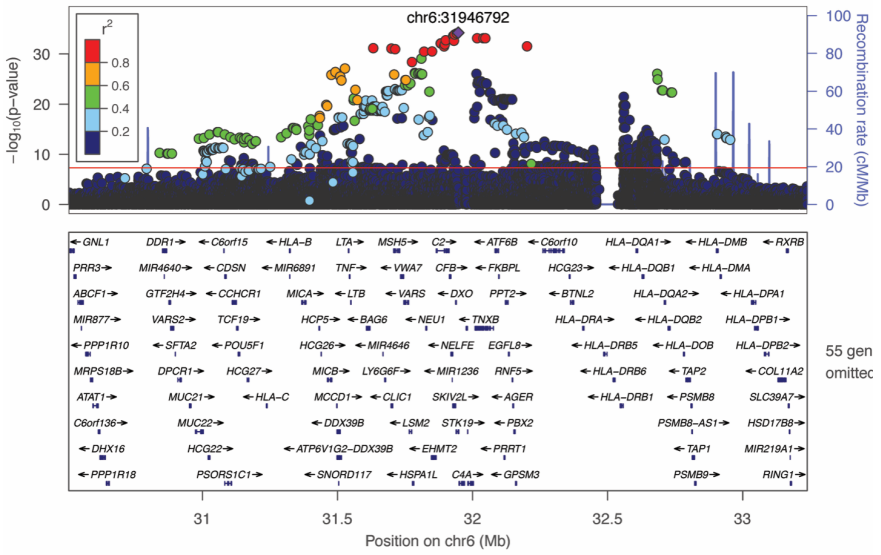
From Fritsche L and Pasaniuc B & Price AL, Nat. Rev. 2017

Conditional Analysis

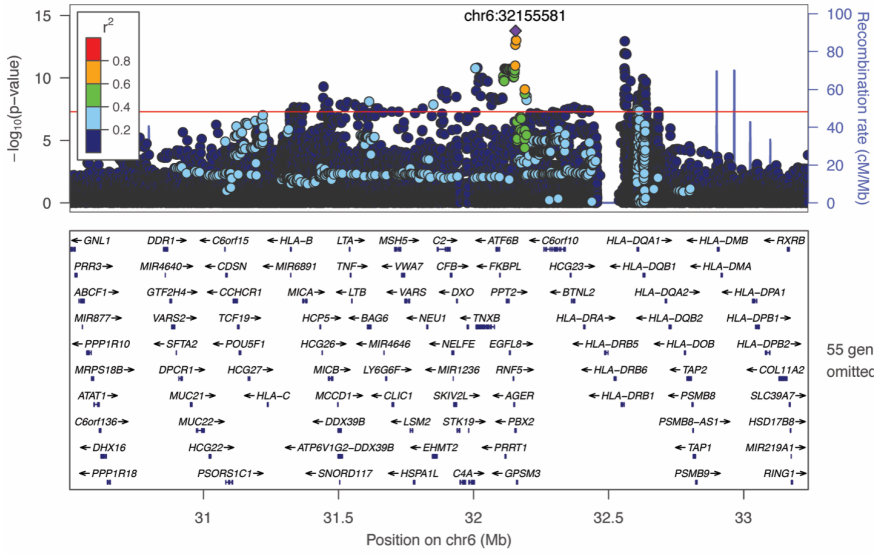
Locus #8.1: rs116503776



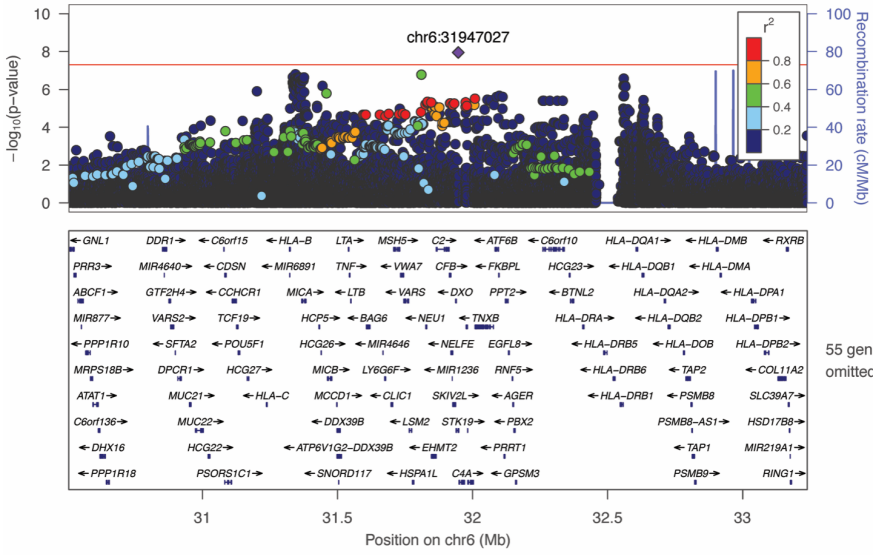
Locus #8.2: rs144629244



Locus #8.3: rs114254831



Locus #8.4: rs181705462

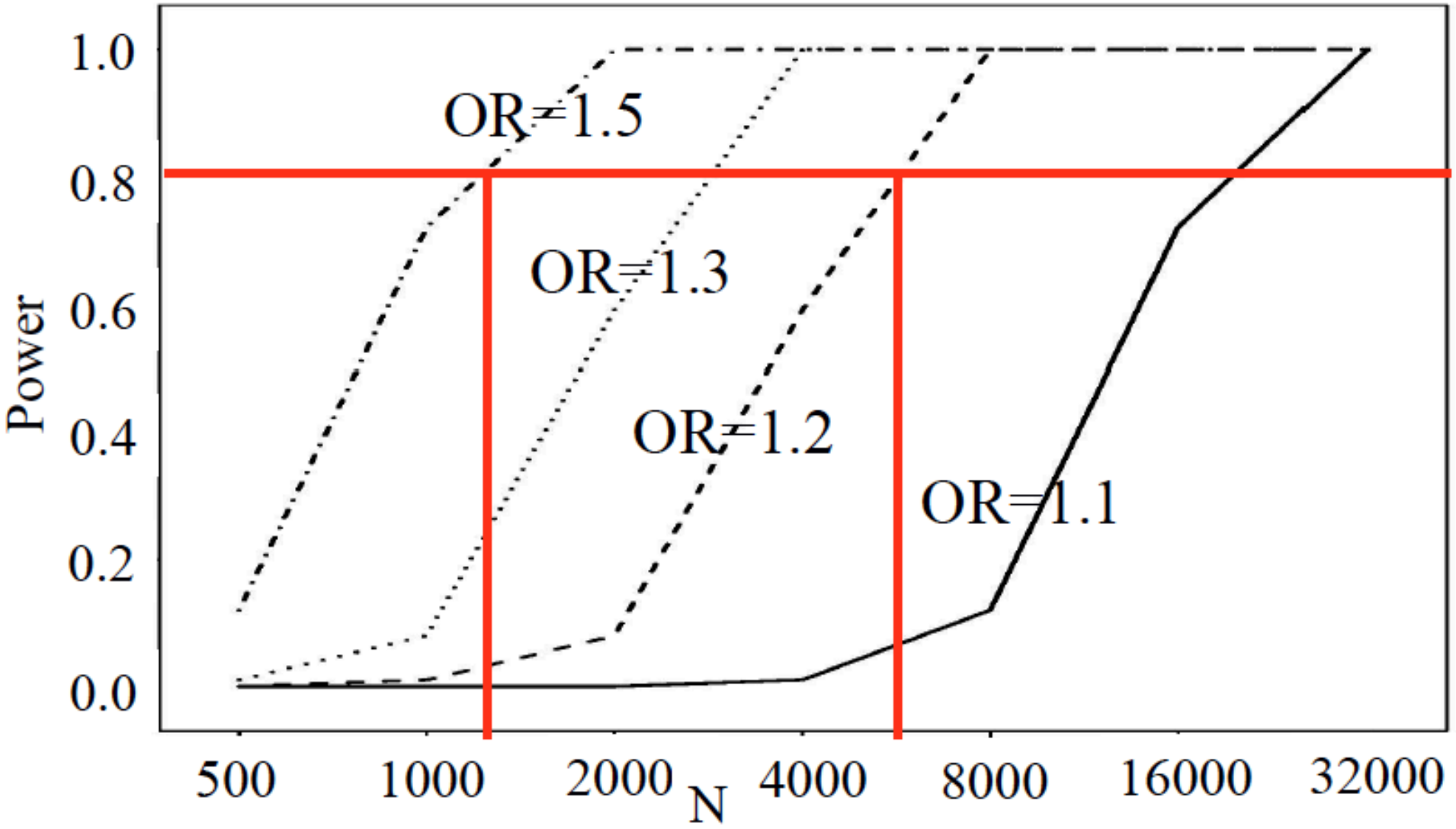


Meta-Analysis

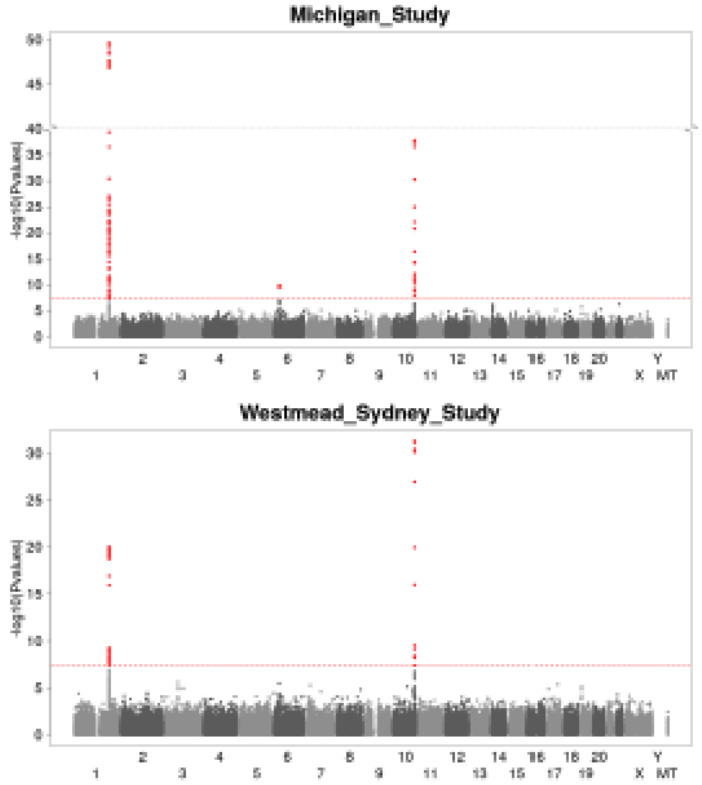
- Combine summary statistics (e.g., p-values, odds ratios, effect-sizes) across multiple studies for the same phenotype
- Improve power for increasing total sample size
- Address between study variances (due to population stratification, study design)
- Avoid the hassle of sharing individual-level genotype/phenotype/covariate data (e.g., privacy protocols)
- Lin and Zeng (2010) showed that meta-analysis with summary results is statistically as efficient as using individual-level data, under an idea situation:
  - Phenotype distributions are the same across all studies
  - Same covariates were included across all studies
  - Equal case/control ratios across all studies

Why Meta-Analysis?

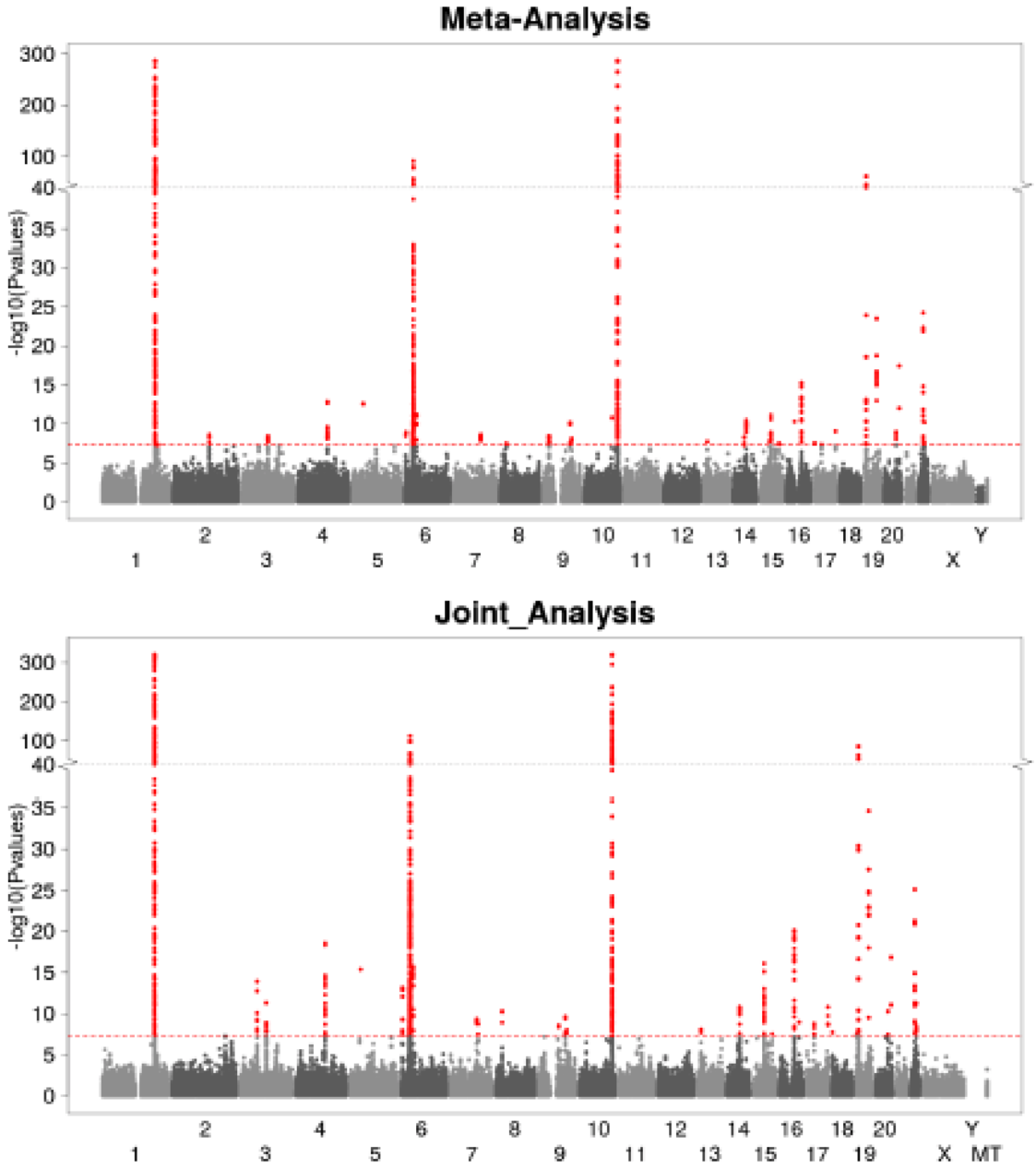
Additive model, N cases, N controls, MAF = .3, $\alpha = 5 \times 10^{-8}$



Why Meta-Analysis?



Example two individual studies of AMD.



Meta-Analysis Methods

- Fisher's Method: combining p-values
- Stouffer's Z-score method
- Fixed Effect Model: combining standardized effect-sizes
- Software: METAL
  https://genome.sph.umich.edu/wiki/METAL_Documentation

Fisher's method

Given summary statistics from individual studies of the same genetic variant

- p_k: p-value from the kth study, $k = 1, \ldots, K$

The test statistic

$$-2 \sum_k \log(p_k) \sim \chi^2_{(2K)}$$

Derivation:

- Under the null, each p_k follows $U[0, 1]$
- The $-\log$ of a uniformly distributed value follows an exponential distribution
- Scaling a value that follows an exponential distribution by a factor of two yields a quantity that follows a χ^2 distribution with 2 df
- The sum of K independent χ^2 values follows a χ^2 distribution with $2K$ df

Stouffer's Z-score method

Given summary statistics from individual studies

- n_k: sample size of the kth study
- p_k: p-value from the kth study
- β_k: effect-size for the kth study

Then, we obtain

- $Z_k = \text{sign}(\beta_k)\Phi^{-1}(1 - p_k/2)$, where Φ is standard normal CDF.
- $w_k = \sqrt{n_k}$: weight

Stouffer's Z statistic is given by

$$\frac{\sum_k w_k Z_k}{\sqrt{\sum_k w_k^2}} \sim N(0, 1)$$

Fixed Effect Model

Inverse-variance estimator

Given summary statistics from individual studies

- $\hat{\beta}_k$: genetic effect-size from the kth study
- v_k: variance of $\hat{\beta}_k$ from the kth study

Then, consider

- $\beta_{meta} = \frac{\sum_k w_k \beta_k}{\sum_k w_k}, \; w_k = 1/v_k$
- $V_{beta} = \frac{1}{\sum_k w_k}$
- Inverse-variance weighting

The Wald test statistic is given by

$$\frac{\beta_{meta}}{\sqrt{V_{meta}}} \sim N(0, 1)$$

Rare Variants

Genetic variants identified in the TOPMed project (2015 - Present)

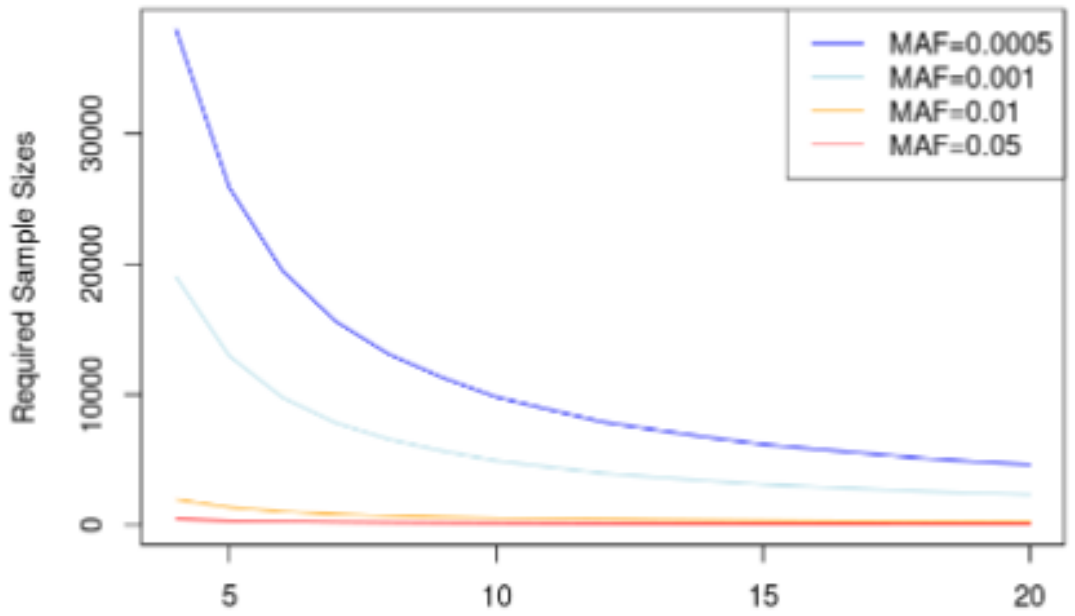
Variant Type	Category	# PASS	# FAIL	% dbSNP (PASS)	Known/Novel Ts/Tv (PASS)
SNP	**All**	**438M**	**85M**	**22.9%**	**1.93 / 1.69**
	Singleton	202M	24M	8.5%	1.23 / 1.54
	Doubleton	69M	8.8M	12.6%	1.61 / 1.74
	Tripleton ~ 0.1%	142M	24M	34.9%	2.23 / 1.99
	0.1% ~ 1%	13M	4.5M	98.2%	2.17 / 1.79
	1 ~ 10%	6.5M	2.9M	99.6%	1.82 / 1.75
	>10%	5.3M	2.0M	99.8%	2.11 / 1.88
Indels	**All**	**33.4M**	**26.2M**	**20.1%**	
	Singleton	15.7M	4.7M	10.1%	
	Doubleton	5.3M	1.8M	12.6%	
	Tripleton ~ 0.1%	10.7M	8.0M	26.7%	
	0.1% ~ 1%	2.8M	968K	88.9%	
	1 ~ 10%	432K	2.3M	98.5%	
	>10%	298K	1.4M	99.6%	

Why Study Rare Variants?

- Most genetic variants are rare
- Functional variants tend to be rare
- Number of samples N needed to observe a rare variant with at least 99.9% probability :

MAF	0.1	0.01	0.001	0.001
N	33	344	3453	34537

- Number of samples needed to achieve 80% power by single variant test (underpowered for rare variants):



Region (or Gene) Based Test

Test the joint effect of rare/common variants within a defined genome region (e.g., gene, regulatory region)

- Consider a total of p variants within a test genome region
- Genotype data: $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})', \ x_{ij} = 0, 1, 2$, for individual i
- Genetic effect-size vector $\beta_{1\times p}$
- Consider covariate vector Z_i for individual i, with extended intercept term
- General linear regression model

$$E[f(\mu_i)] = Z_i'\alpha + X_i'\beta \ ,$$

  where $f(\cdot)$ is a link function, e.g., logistic function for dichotomous traits, identity function for quantitative traits.

- Test region-based association

$$H_0 : \beta = (\beta_1, \cdots, \beta_p) = 0$$

Burden (collapsing test)

- Assume there is a shared genetic effect across all variants, i.e., $\beta_j = w_j \beta_{burden}$
- One commonly used variant weight is based on MAF, $w_j = 1/\sqrt{MAF_j(1 - MAF_j)}$
- Equivalent to consider a Burden genotype score

$$C_i = \sum_j w_j x_{ij}$$

- Model is equivalent to

$$E[f(\mu_i)] = Z_i' \alpha + C_i \beta_{burden}$$

- Region-based test is equivalent to test

$$H_0 : \beta_{burden} = 0$$

Burden (collapsing test)

- Score test statistic can be used

$$T_{score} = \sum_{i=1}^{n} C_i(Y_i - \widehat{\mu_{0i}}),\ var(T_{score}) = C'\left(\widehat{P} - \widehat{P}Z(Z'\widehat{P}Z)^{-1}Z'\widehat{P}\right)C;$$

$$\frac{T_{score}}{\sqrt{var(T_{score})}} \sim N(0,1);$$

  - Let $\widehat{\alpha}$ denote the covariate coefficients estimated under the NULL model $E[f(\mu_i)] = Z_i\alpha$
  - $\widehat{\mu_{0i}} = Z_i'\widehat{\alpha}$ and $\widehat{P} = \widehat{\sigma_\epsilon^2}I$, with error variance estimate $\widehat{\sigma_\epsilon^2}$ under NULL model, for standard linear regression model
  - $\widehat{\mu_{0i}} = logit^{-1}(Z_i\widehat{\alpha})$ and $\widehat{P} = diag\left(\widehat{\mu_{01}}(1-\widehat{\mu_{01}}), \cdots, \widehat{\mu_{0n}}(1-\widehat{\mu_{0n}})\right)$ for logistic regression model

- Underpowered when genetic variants have opposite effect-sizes within a tested region

Variance component test

- Consider the general linear regression model

$$E[f(\mu_i)] = Z_i'\alpha + X_i'\beta \ ,$$

- Assume $\beta_j \sim N(0, w_j^2\tau)$, sharing a common variance component τ
- Then region-based test is equivalent to test

$$H_0 : \tau = 0$$

Region (or Gene) Based Test

SKAT: Sequence Kernel Association Test, *Wu et al. (2011) AJHG*

- Test statistic (i.e., weighted sum of single variant score test statistics) :

$$Q_{SKAT} = (Y - \widehat{\mu_0})XWWX'(Y - \widehat{\mu_0}) = \sum_{j=1}^{p} w_j^2[X_j'(Y - \widehat{\mu_0})]$$

- Let $K = XWWX'$ denote the kernel matrix, $W = diag(w_1, \cdots, w_p)$
- $\widehat{\mu_0}$ is estimated under the NULL hypothesis
- Q_{SKAT} asymptotically follows a mixture of $\chi^2_{(1)}$ distribution under the NULL hypothesis

$$Q_{SKAT} \approx \sum_{j=1}^{p} \lambda_j \chi^2_{(1)}$$

- λ_j are eigenvalues of $P_0^{1/2} K P_0^{1/2}$, with projection matrix $P_0 = \widehat{P} - \widehat{P}Z(Z'\widehat{P}Z)^{-1}Z'\widehat{P}$.
- The mixture of $\chi^2_{(1)}$ distribution can be approximated with the computationally efficient Davies method, which will be used to calculate p-value.

Post-GWAS Analysis

- Replication study with independent datasets
- Fine-mapping GWAS loci while accounting for functional annotation
- Biological interpretation
- Biological replication (e.g., CRISPER-CAS9)

References

- Price A.L. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics volume 38, pages 904-909 (2006).
- Purcell, S. et. al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet.81, 559-575. 2007.
- Cristen J. Willer, Yun Li, Gonçalo R. Abecasis; METAL: fast and efficient meta-analysis of genome-wide association scans, Bioinformatics, Volume 26, Issue 17, 1 September 2010, Pages 2190-2191.
- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83(3):311-21.
- Wu M.C. et. al. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82-93.