

# Genome-wide Association Studies

## Part 2: Association Testing

# GWAS

- Association analysis is relatively straight forward...
- At each SNP in the genome, a simple statistical test is performed to assess the association between the SNP and trait of interest.

# Quantitative Traits

- Test *correlation* between trait and SNPs
- Typically uses a simple “additive” model
- Each SNP is encoded 0, 1 or 2 representing the number of B alleles in the genotypes AA, AB and BB
- This is referred to as the additive model of association → each copy of the B allele is adding to the trait

# Quantitative Traits

- Additional covariates (age, sex, ...) can be included in a linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \beta_3 D_i + \dots$$

- It is important that the assumptions of linear regression are met
- Particularly normality of residuals
  
- Phenotypes may be transformed with a rank-based inverse normal transformation
- Could precorrect data for covariates with large effects and then ensure the normality

# Disease Traits

- Test whether the proportion of B alleles at a SNP differs between cases and controls
- This is a multiplicative model of association
- The risk of developing the disease by a factor  $r$  for each B allele carried
- i.e.
  - baseline risk of  $b$  for genotype AA
  - risk of  $br$  for genotype AB
  - risk of  $br^2$  for genotype BB

# Disease Traits

- Testing for association can be done using a simple chi-square contingency table test with a 2x2 matrix containing the counts of A and B alleles for cases and controls in each row

## Alleles

	1	2	Total
Case	$n_1$	$n_2$	2N
Ctrl	$m_1$	$m_2$	2M
Total	$T_1$	$T_2$	2(N+M)

2x2 contingency table

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

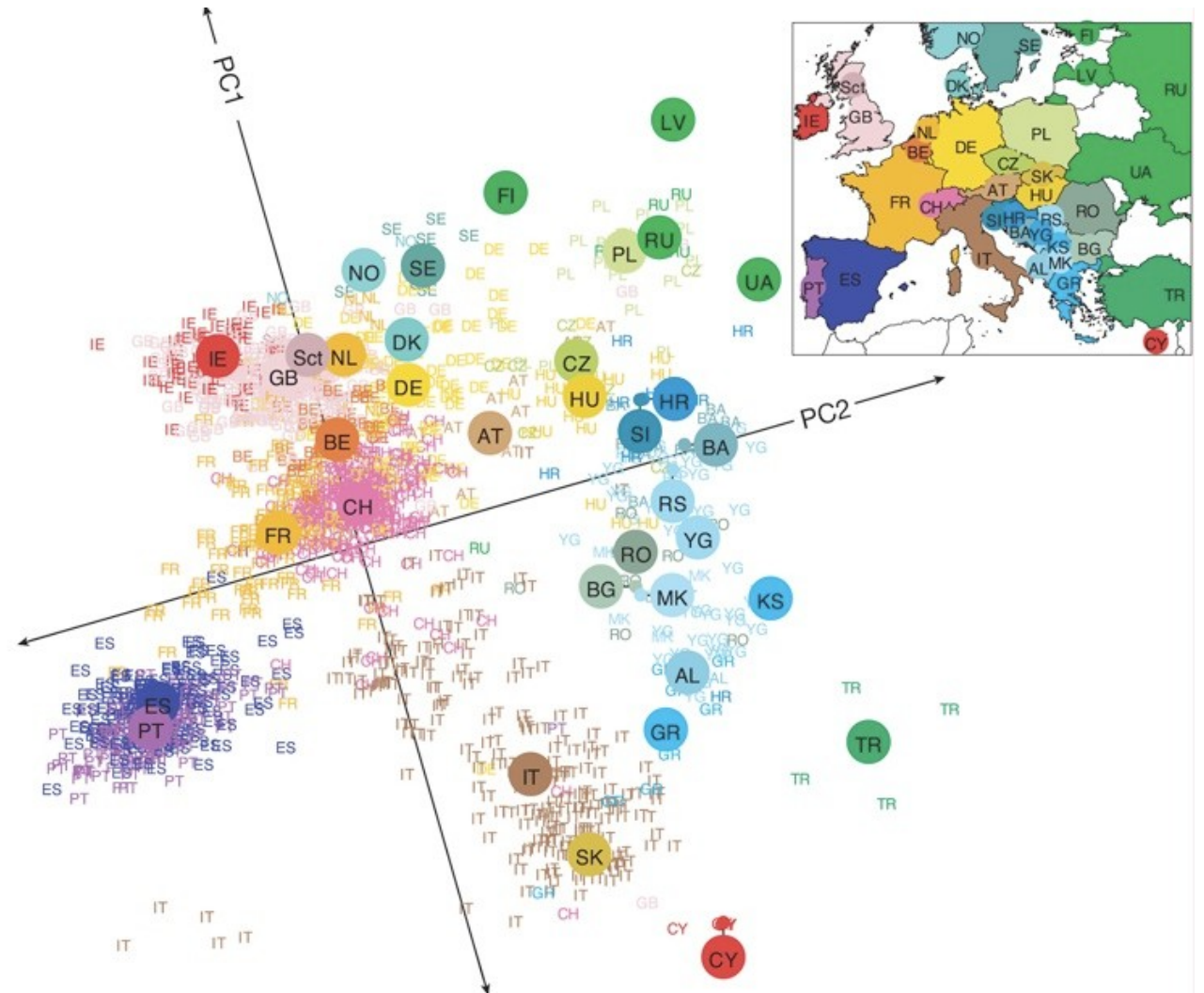
# Including Genotype PC Covariates

- Ancestral outliers were removed during the cleaning stage
- Smaller scale differences in ancestry will still be present in the data and can be corrected for by including PCs from genotypes
- PCs can also correct for possible biases induced by sample collection or non-genetic geographical effects on phenotype
- How many PCs to include? 10 or 20 are common guidelines

# Principle Components

- Include PCs as covariates in your GWAS to control for remaining population stratification
- Would have prevented association of height at the lactase gene

a





# Significance

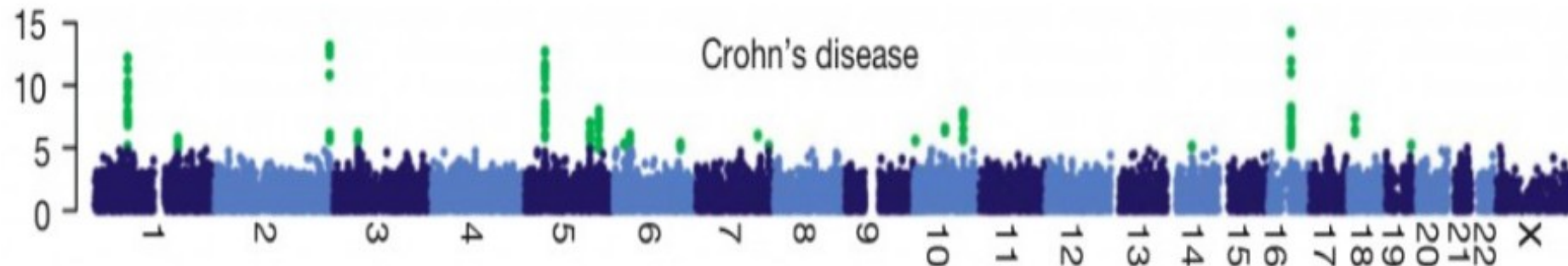
- It is important to correct for the large number of tests performed in a GWAS study when assessing the significance of a result
- Correcting for the number of SNPs tested using (e.g.) a Bonferroni correction is overly conservative due to the linkage disequilibrium between SNPs
- In humans significance threshold of  $5 \times 10^{-8}$  corrects for the effective number of independent tests genome-wide
- A less stringent threshold of  $1 \times 10^{-5}$  is widely used to indicate “suggestive” significance, but never do that...

# Manhattan Plots

- GWAS results are typically represented using a Manhattan plot
  - genomic locations along the X-axis
  - negative logarithm (base 10) of the p-value along the Y-axis
  - each point is the result from a single SNP
- The SNPs with the strongest associations will have the greatest negative logarithms, and will tower over the background of unassociated SNPs (like skyscrapers in the Manhattan skyline)

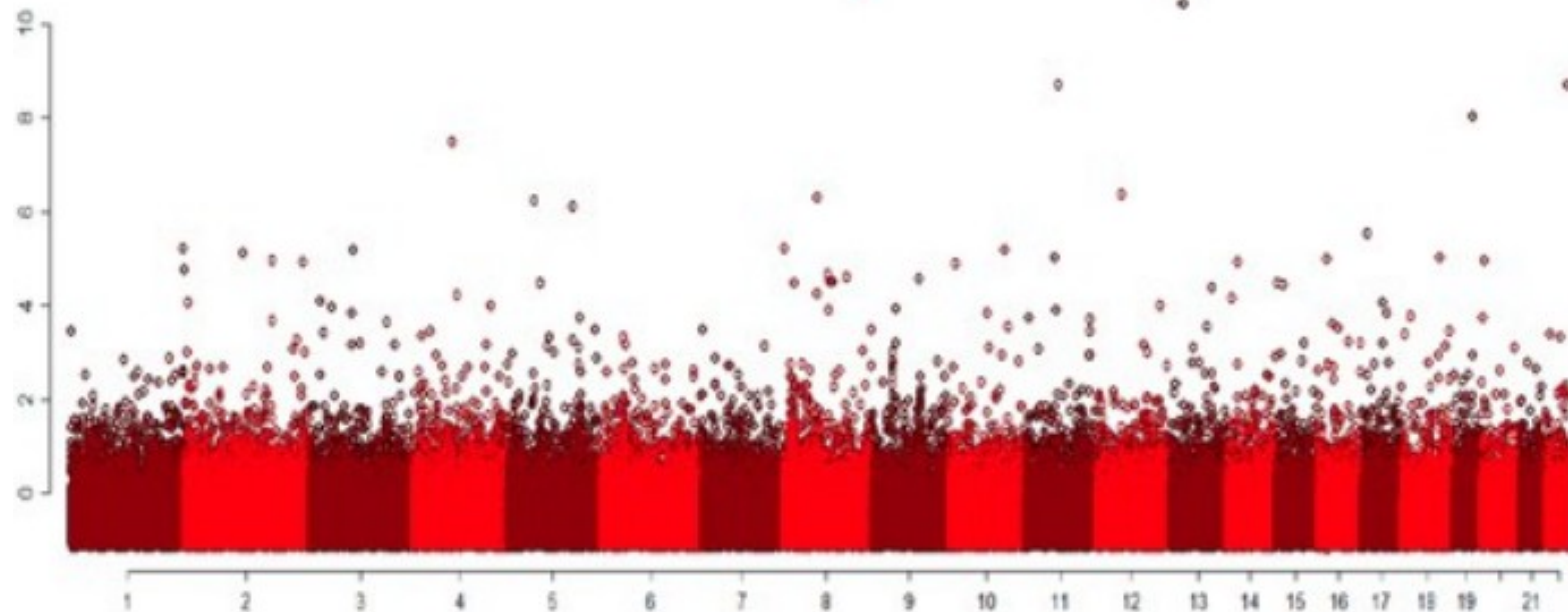
# Manhattan Plots

- A *good* Manhattan plot
- Wellcome Trust Case Control Consortium, Crohn's disease, Nature 2007
- Shows signals supported by many neighboring SNPs

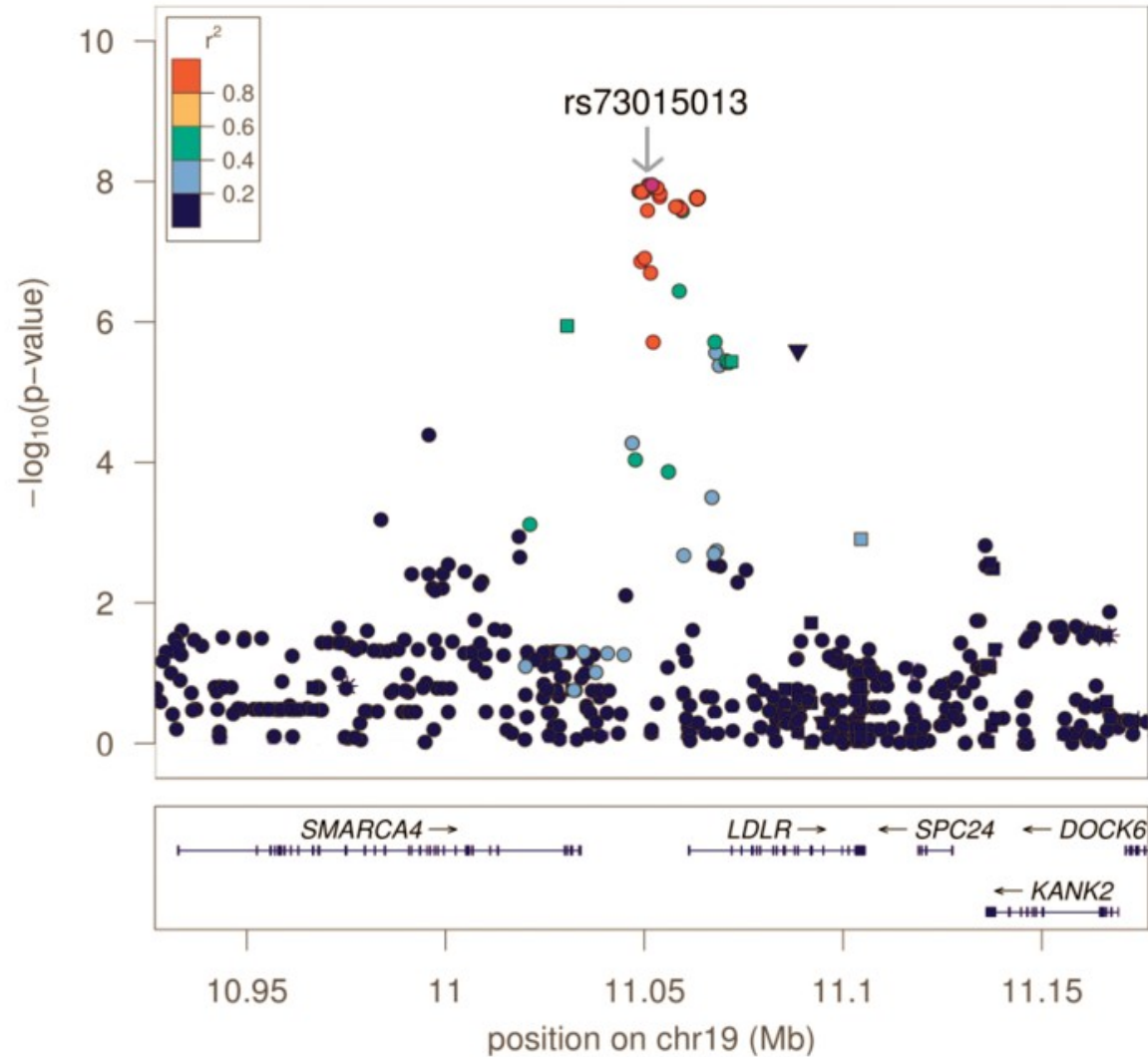


# Manhattan Plots

- A *bad* Manhattan plot
- Sebastiani et al. “Genetic signatures of exceptional longevity in humans” Science July 2010
- Retracted July 2011 because of poor QC



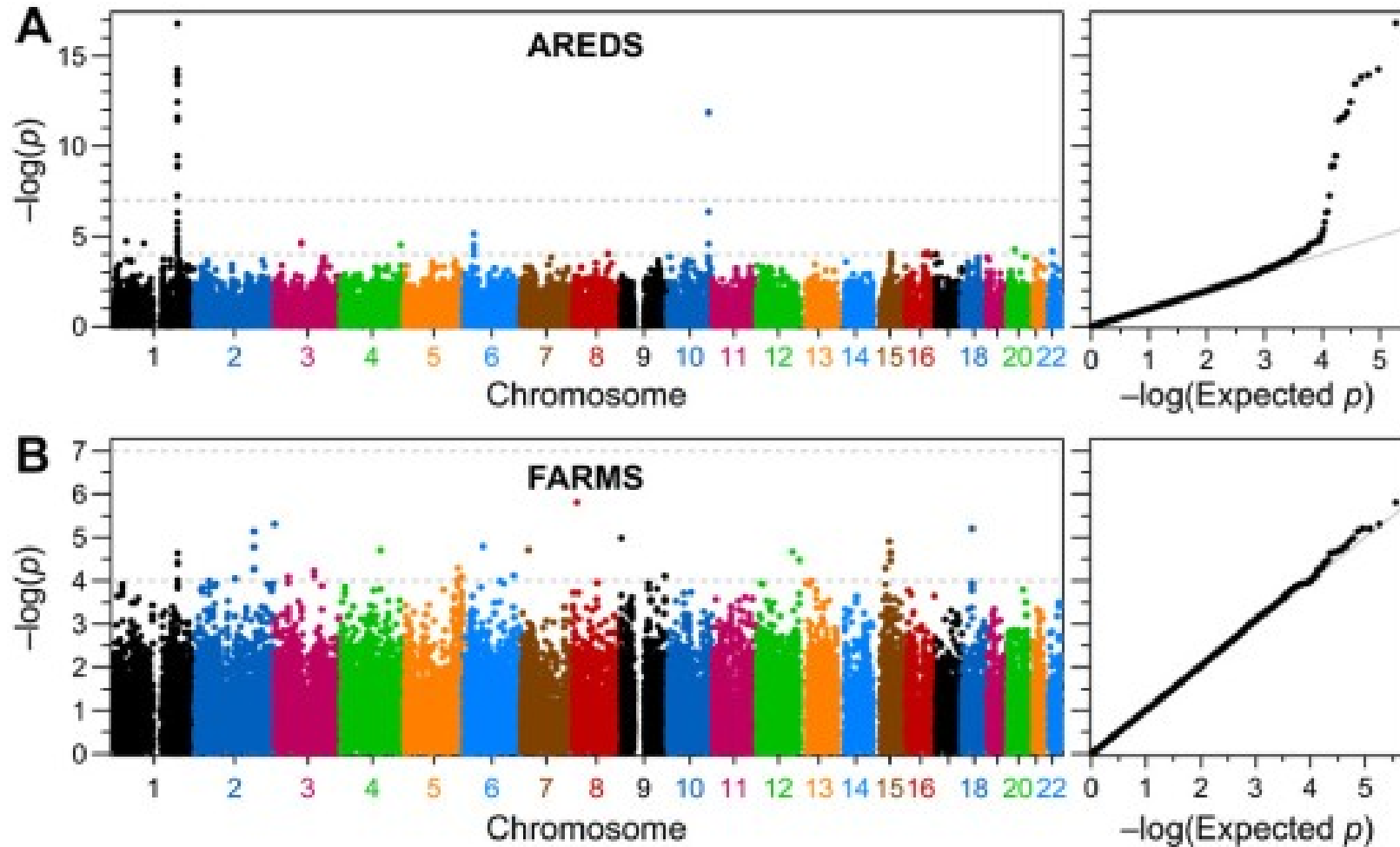
# Regional Association Plots



# QQ Plot

- A QQ plot is a common way to demonstrate the lack of confounding effects in a GWAS
- 
- The ordered observed negative logarithm of the p-values are plotted against the expected distribution under the null hypothesis of no association
- Ideally, the points in the plot should align along the  $X = Y$  line, with deviation at the end for the significant associations

# QQ Plot

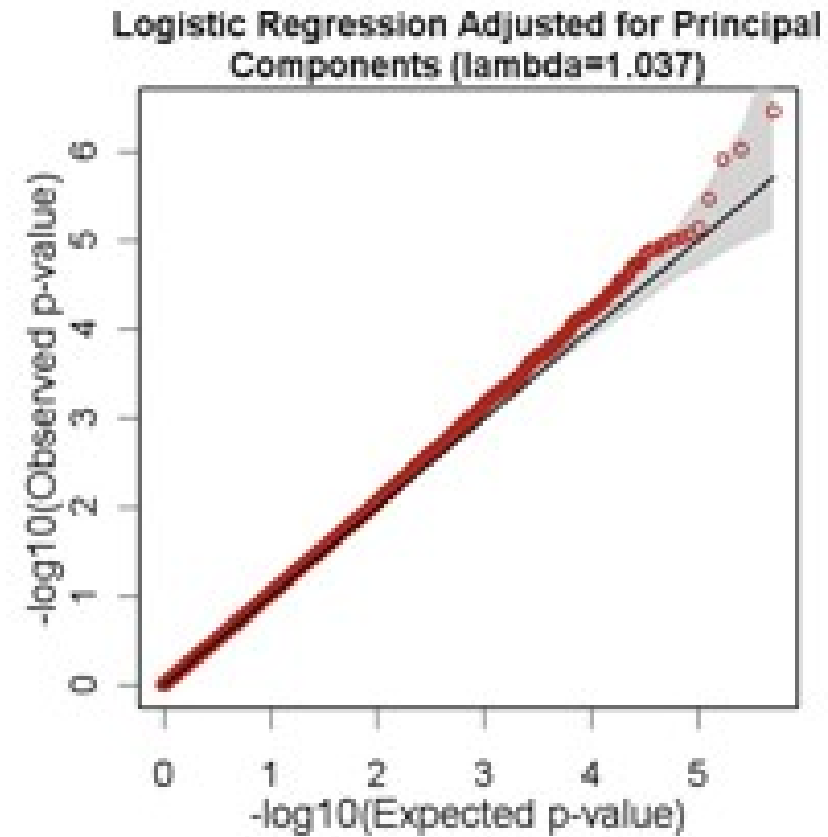
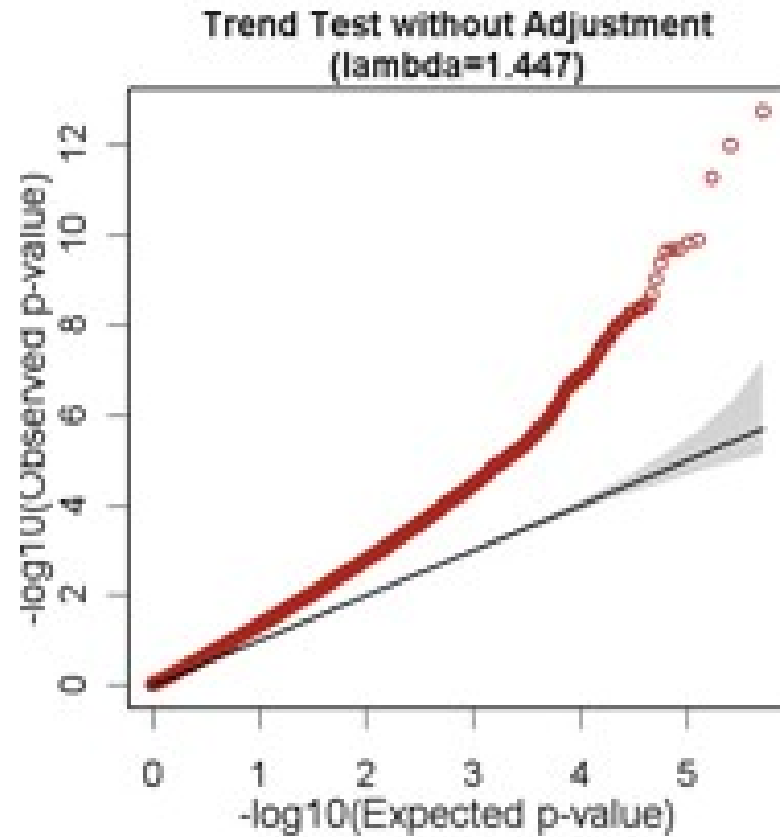


# Genomic Inflation

- One way to quantify the lack of global inflation in the QQ plot is the genomic inflation factor ( $\lambda_{GC}$ )
- This is calculated by:
  - determining the median p-value of GWAS test statistics
  - calculating the quantile in a chi-squared distribution with one degree of freedom that would give this p-value
  - divide this by the median of a chi-squared distribution with one degree of freedom (0.4549)
- Deviations of this value away from 1.0 indicate genome-wide confounding in the data.



# Genomic Inflation



# How Many Signals in a GWAS Peak

- The SNP showing the strongest statistical evidence for association in a genomic region (for example, a 2-Mb window centered on the locus) is often reported to represent the association in this region
- This assumes that the detected association at the top SNP captures the maximum amount of variation in the region by its LD with an unknown causal variant and that other SNPs in the vicinity show association because they are correlated with the top SNP
- However, there may be multiple causal variants at the locus

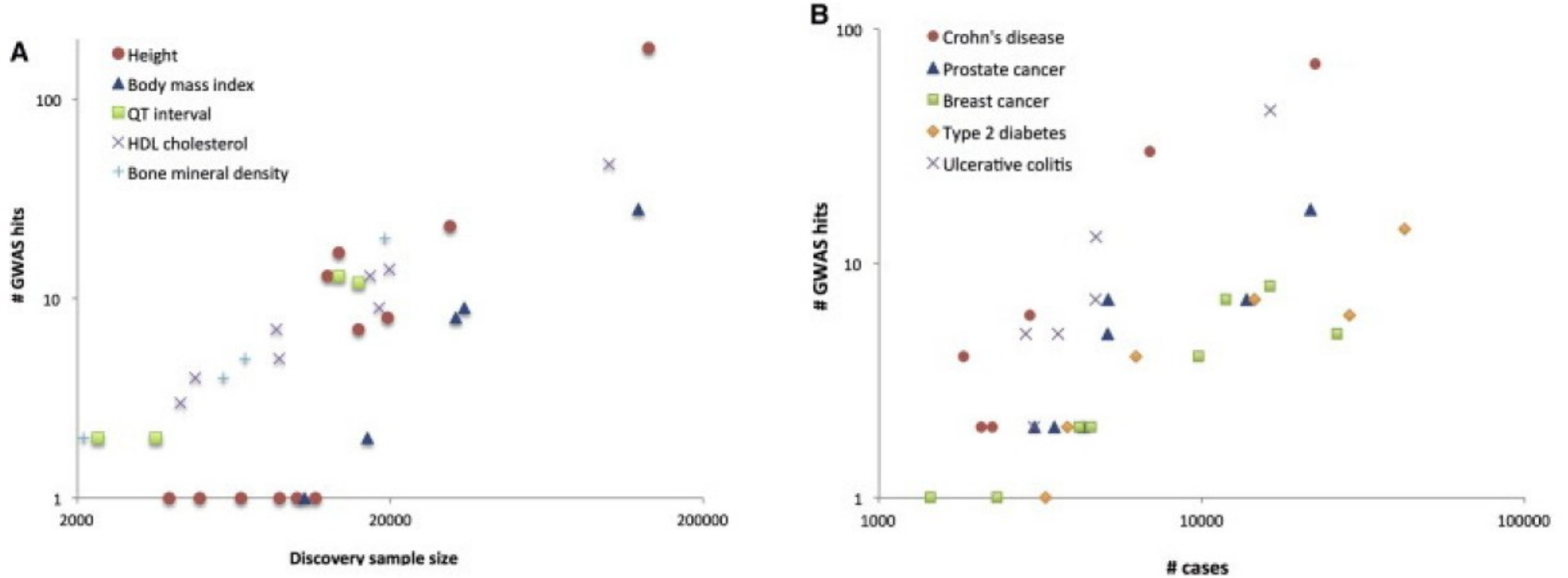
# How Many Signals in a GWAS Peak

- PLINK provides a LD-based result clumping procedure
- SNPs are “clumped” into groups with high-linkage disequilibrium
- Procedure:
  - take most significant “unclumped” SNP (lowest p-value)
  - look at all SNPs with  $R^2 > x$  and within  $y$  distance
  - clump all significant SNPs in that group to one set
  - repeat until significant SNPs are all clumped
- Somewhat arbitrary choice of thresholds...

# How Many Signals in a GWAS Peak

- GCTA-COJO: Conditional and Joint analysis of significant SNP
  - Perform GWAS
  - Take most significant SNP and either:
    - (1) add it as covariate to GWAS, or
    - (2) regress its effect out of the phenotype
  - Repeat until no significant SNPs left
- Can focus on just region of interest or whole genome
- Can identify new signals in regions with no previous association signal

# Sample Size Requirements



# Meta-Analysis

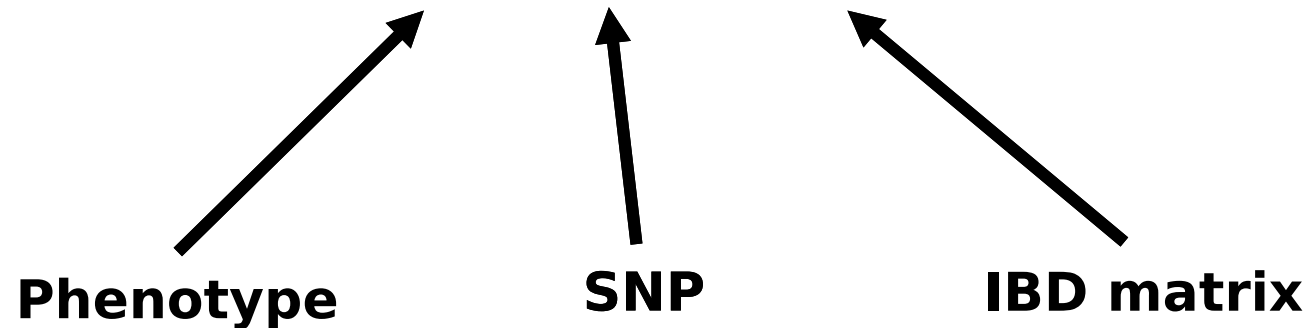
- We need very large sample sizes to detect associations with variants of small effect in GWAS
- 
- It is rare to have a large enough cohort to detect a large number of variants, particularly for disease case control analysis
- We can use a meta-analysis to combine results from a number of studies to effectively increase our sample size
- Common approach for international consortia

# Meta Analysis

- Why not combine datasets?
  - Privacy
  - Ethics
  - Population Stratification
  -

# Including Relatives in GWAS

- Including relatives in a GWAS can create false positive results *unless accounted for properly*
- Used a mixed model to account for the relatedness (covariance) between relatives

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$


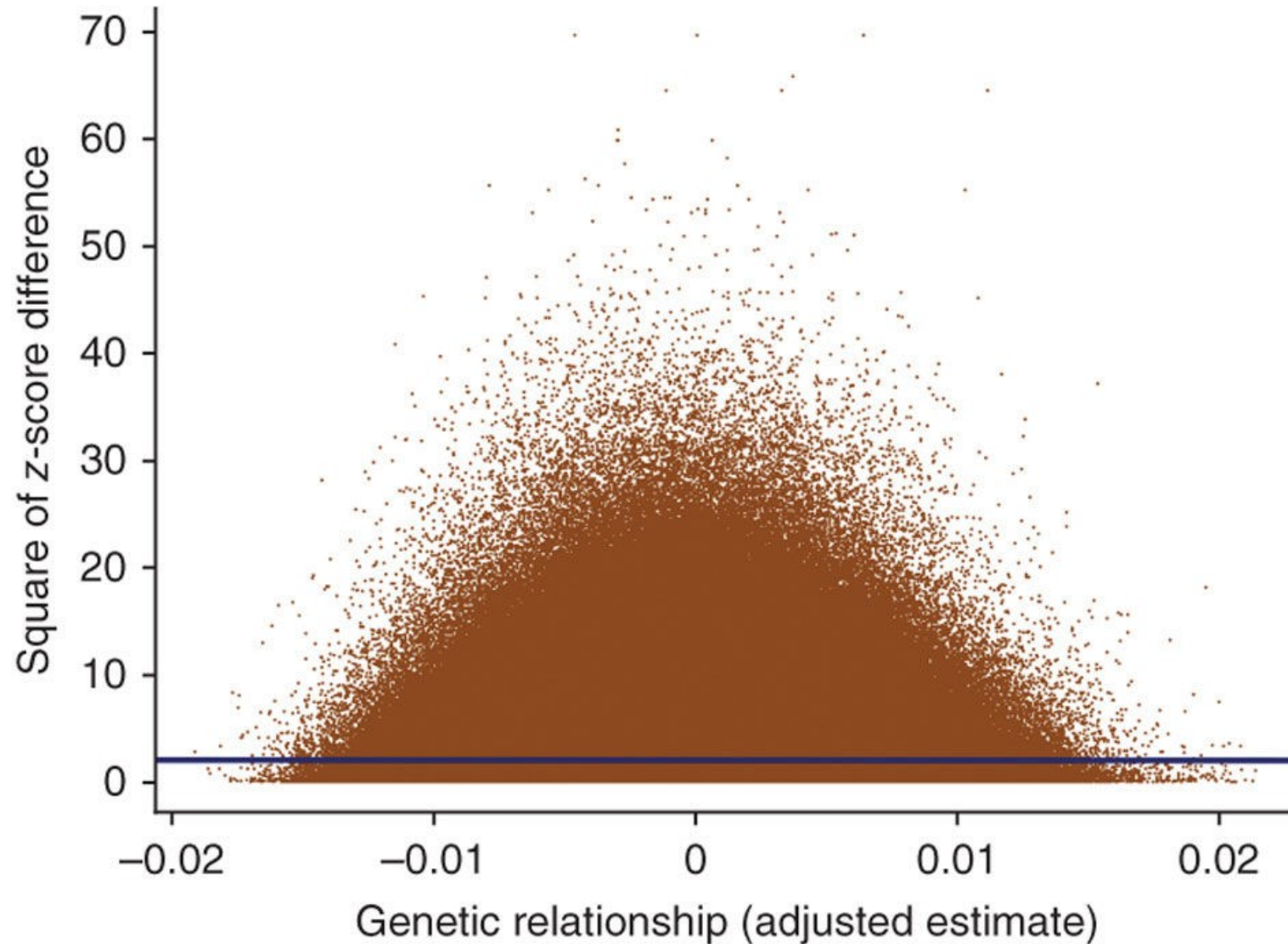
Phenotype

SNP

IBD matrix



# Everyone is Related to Some Extent



# Tools for GWAS with Relatives

- Many available tools to account for relatives. Some common tools are:
- GCTA
  - MLMA – mixed linear model analysis
  - LOCO – leave-one chromosome out
  - fastGWA – super efficient accounting for close relatives in Biobank sized cohorts
- BOLT-LMM – efficient Bayesian linear regressions
- ....