Statistical Genetics

Yijuan Hu

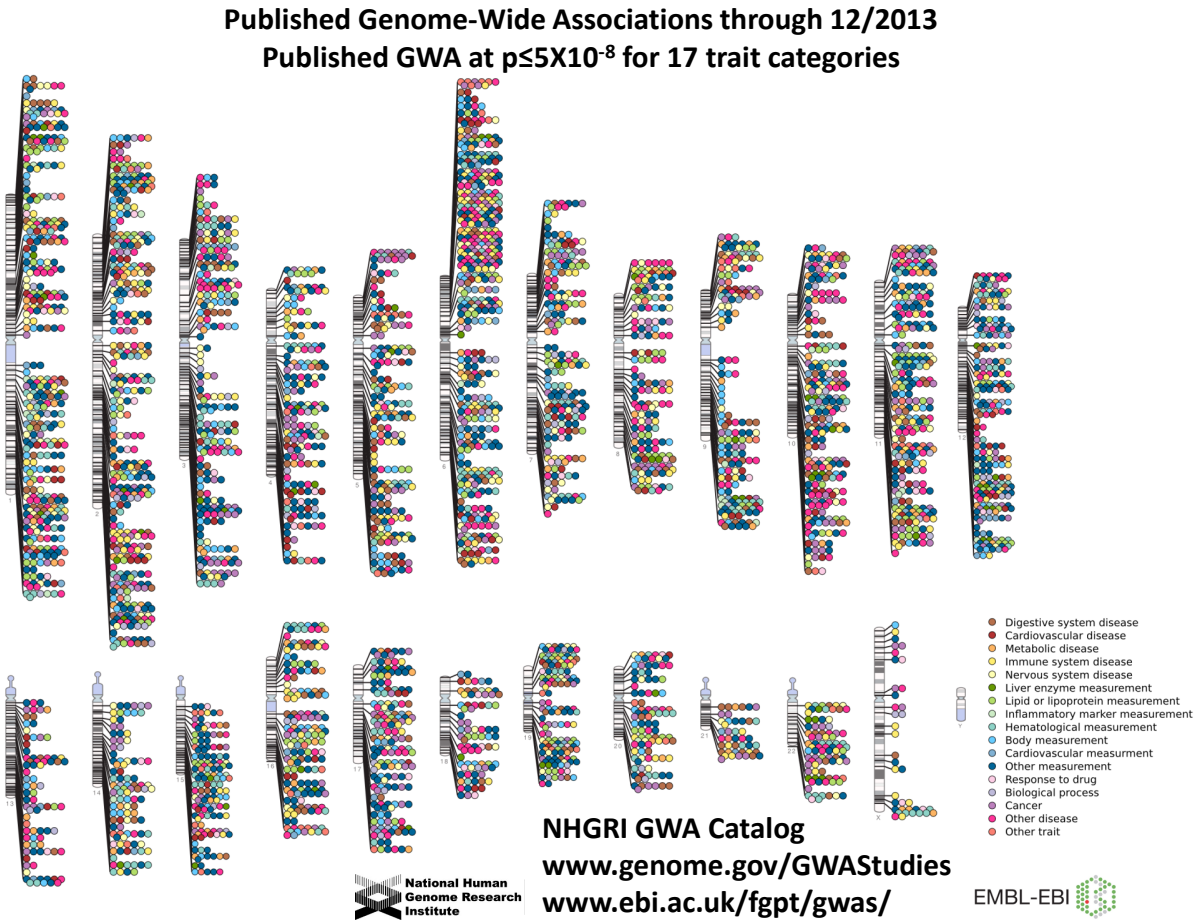
2017-10-5

Outline

- What is statistical genetics?
- Genetics background: SNP, linkage disequilibrium (LD)
- Statistical methods for association analysis
- Statistical problems in Genome-Wide Association Studies (GWAS)

What is Statistical Genetics?

- To develop statistical methods for drawing inferences from genetic data
- To discover genetic variants that influence human traits/diseases (height, BMI, diabetes, cancers, ...)

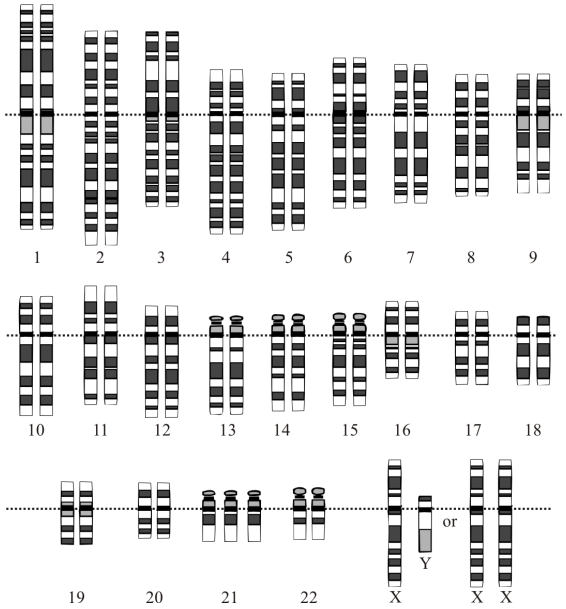


Success “stories” of statistical genetics

Genetics Background

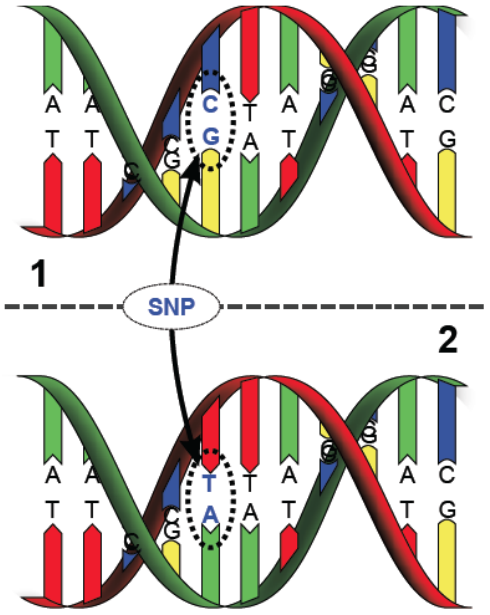
Human Genome

- Autosome: 22 pairs, numbered 1 to 22
- Sex Chromosome: X and Y chromosomes
- Homologous Chromosomes: two non-identical chromosomes in an autosome pair



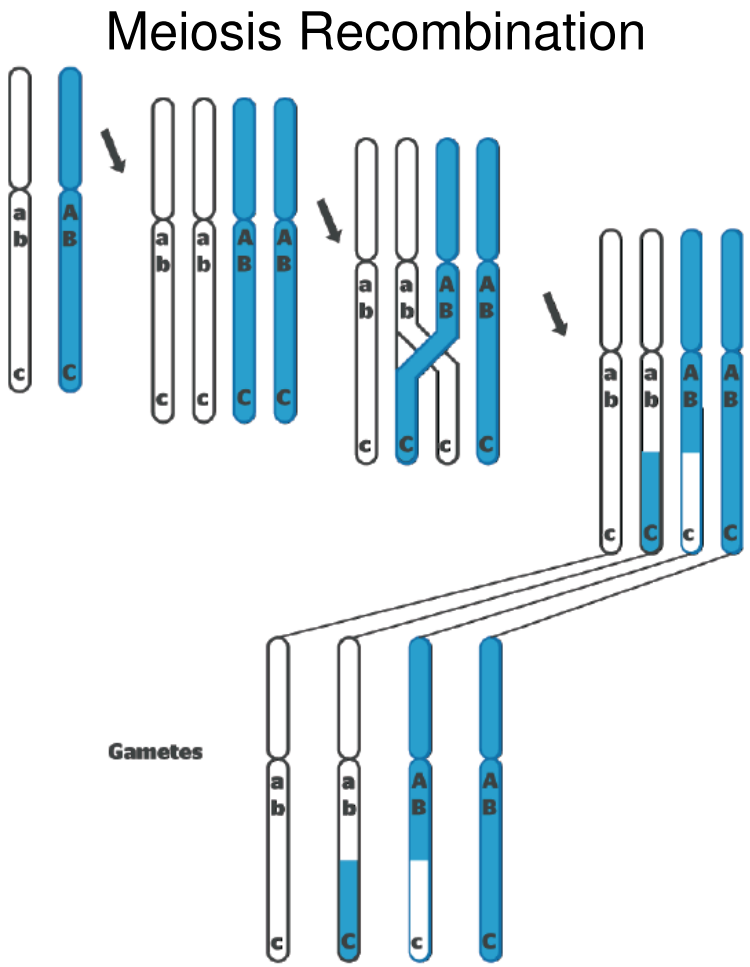
Single Nucleotide Polymorphisms (SNP)

- A DNA sequence variation occurring at a single base pair
- >10 million catalogued in human genome with rs numbers
- Most important in genetic association studies



Linkage Disequilibrium (LD)

Alleles of genes that are close together on the same chromosome are not passed independently. The closer that genes are, the more likely their alleles are passed together.



Measuring LD

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b}$$

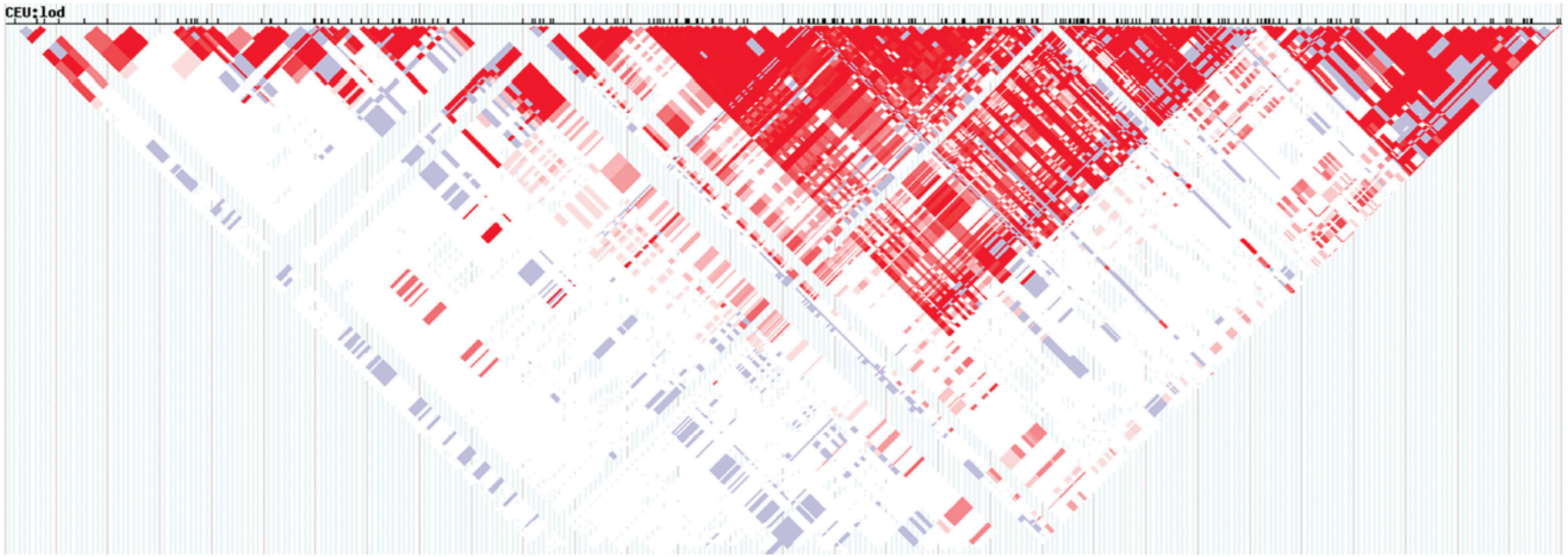
- r^2 ranges between 0 and 1
- $r^2 = 1$ when the two markers provide identical information
  - knowing the allelic status at one marker perfectly predicts that at the other
  - only two out of the four two-locus haplotypes are observed in the sample
  - perfect LD
- $r^2 = 0$ when they are in perfect equilibrium
- r^2 is the squared Pearson correlation

LD and Physical Distance

The relationship is not monotonic.

In some regions, LD is found to be very weak, even over very short distances; in others, LD can stretch over more than 100Kb or even 1Mb.

Also, LD strongly depends on the chromosomal region.



Statistical Methods for Population-Based Association Analysis

Population-Based Association Analysis

Association analysis based on unrelated study subjects is not fundamentally different from any other statistical association analysis.

The objective is to establish an association between two variables: a disease trait and a genetic marker.

- The disease trait can be dichotomous, quantitative, or (potentially censored) time-to-onset of a disease.
- The genetic marker can be a known or suspected disease-causing mutation, or a marker without any known effect on DNA coding.

Study Designs

Use standard epidemiological designs for studying the relationship between general risk factors and diseases.

Case-Control Design

Ascertain subjects on the basis of dichotomous disease outcome

- informative, efficient, low cost
- selection bias, recall bias
- cannot estimate disease prevalence

Cohort Design

Follow subjects over time for development of disease and/or risk factors

- no selection and recall bias
- reliable pre-disease exposure information
- a full range of diseases and traits
- many years of follow-up

Single-SNP Test of Association in Case-Control Studies

Often rely on standard contingency table methods:

1. Pearson Chi-square test
2. (Large-sample) Z-test comparing two proportions
3. (Small-sample) Fisher exact test
4. Logistic regression

Frequently-used tests:

1. Genotype test (2-*df* test)
2. Dominant/recessive test (1-*df* test)

Genotype Test

- Compare genotype frequencies in cases and controls in a 2×3 table

	AA	Aa	aa	Total
Case	n_{10}	n_{11}	n_{12}	$n_{1.}$
Control	n_{00}	n_{01}	n_{02}	$n_{0.}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	n

The genotype/codominant test: D – disease status; G – genotype

$$H_0 : \Pr(D = 1|G = \text{AA}) = \Pr(D = 1|G = \text{Aa}) = \Pr(D = 1|G = \text{aa})$$

$$H_1 : \text{At least one inequality holds}$$

The standard $2\ df$ Pearson Chi-square test of independence for a 2×3 table is:

$$X^2_G = \sum_{i=0,1} \sum_{j=0,1,2} (O_{ij} - E_{ij})^2 / E_{ij} \quad \sim \quad \chi^2,\ df = 2$$

– $O_{ij} = n_{ij}$: observed count in the cell

– $E_{ij} = n_{i.}n_{.j}/n$: expected count under independence: $np_{D=i}p_{G=j} = n(n_{i.}/n)(n_{.j}/n)$

Genotype Test: Example

- TCF7L2 for Type 2 Diabetes in Finns
- SNP rs12255372 has alleles T and G

	GG	GT	TT	Total
Case	661	255	20	936
Control	724	354	50	1128
Total	1385	609	70	2064

$$X_G^2 = (661 - 628.08)^2/628.08 + \ldots \approx 14.08 \quad \sim \quad \chi^2,\ df = 2$$

$$p = .0009$$

$\Pr(G|D)$

	GG	GT	TT	Total
Case	0.71	0.27	0.02	1.0
Control	0.64	0.31	0.05	1.0

$\Pr(D|G)$

	GG	GT	TT	Total
Case	0.48	0.42	0.29	0.45
Control	0.52	0.58	0.71	0.55

Dominant or Recessive Test

- Compare frequencies of AA or Aa with aa in cases and controls in a 2×2 table
- Assume dominant or recessive Mendelian genetic disease model
- More powerful than genotype test if the genetic model is true

	AA or Aa	aa	Total
Case	$n_{10} + n_{11}$	n_{12}	$n_{1.}$
Control	$n_{00} + n_{01}$	n_{02}	$n_{0.}$
Total	$n_{.0} + n_{.1}$	$n_{.2}$	n

The dominant/recessive test:

$$H_0 : \Pr(D = 1|\text{AA}) = \Pr(D = 1|\text{Aa}) = \Pr(D = 1|\text{aa})$$
$$H_1 : \Pr(D = 1|\text{AA or Aa}) \neq \Pr(D = 1|\text{aa})$$

The standard $1\ df$ Pearson χ^2 test of independence for a 2×2 table is:

$$X_D^2 = \sum_{i=0,1} \sum_{j=0,1} (O_{ij} - E_{ij})^2 / E_{ij} \quad \sim \quad \chi^2,\ df = 1$$

How to obtain E_{ij}?

Dominant or Recessive Test: Example

- TCF7L2 for Type 2 Diabetes in Finns
- SNP rs12255372 has alleles T and G
- Allele T is dominant to G

	GG	GT+TT	Total
Case	661	255+20=275	936
Control	724	354+50=404	1128
Total	1385	609+70=679	2064

$$X_D^2 \approx 9.60 \quad \sim \quad \chi^2,\ df = 1$$

$$p = .0019$$

Measure of Association Strength: Odds Ratio

	Exposed (E)	Not Exposed ($\bar{E}$)
Case (D)	a	b
Control ($\bar{D}$)	c	d

Odds ratio:

$$\begin{aligned} OR &= \frac{P(D|E)/P(\bar{D}|E)}{P(D|\bar{E})/P(\bar{D}|\bar{E})} \\ &= \frac{P(E|D)/P(\bar{E}|D)}{P(E|\bar{D})/P(\bar{E}|\bar{D})} \\ &= ad/bc \end{aligned}$$

– Y = dichotomous phenotype
– X = a coding for the genotype

Genotype	Codominant	Dominant	Recessive	Additive
AA	$X = (0,1)^{\mathrm{T}}$	$X = 1$	$X = 1$	$X = 2$
Aa	$X = (1,0)^{\mathrm{T}}$	$X = 1$	$X = 0$	$X = 1$
aa	$X = (0,0)^{\mathrm{T}}$	$X = 0$	$X = 0$	$X = 0$

The association can be specified in a logistic regression model:

$$\log\left[\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)}\right] = \beta_0 + \beta_1 X$$

where β_0 is the intercept, and β_1 is the association parameter.

The Wald, score, or likelihood ratio test for logistic regression can be used to test

$$H_0 : \beta_1 = 0$$

General Regression Approach

Extension to other phenotypes:

- The phenotype Y can be a count or a continuous outcome.
- The generalized linear model is given by

$$g[\mathrm{E}(Y|X)] = \beta_0 + \beta_1 X$$

where $g(.)$ is a link function.

Extension to covariate adjustment:

- An added advantage of the regression approach
- Letting Z_i denote a vector of covariates, such as race, age, and sex, we can incorporate covariates into the regression by

$$g[\mathrm{E}(Y|X,Z)] = \beta_0 + \beta_1^{\mathrm{T}} X + \beta_2^{\mathrm{T}} Z$$

- Allows gene-environment interactions.

Statistical problems in Genome-Wide Association Studies (GWAS)

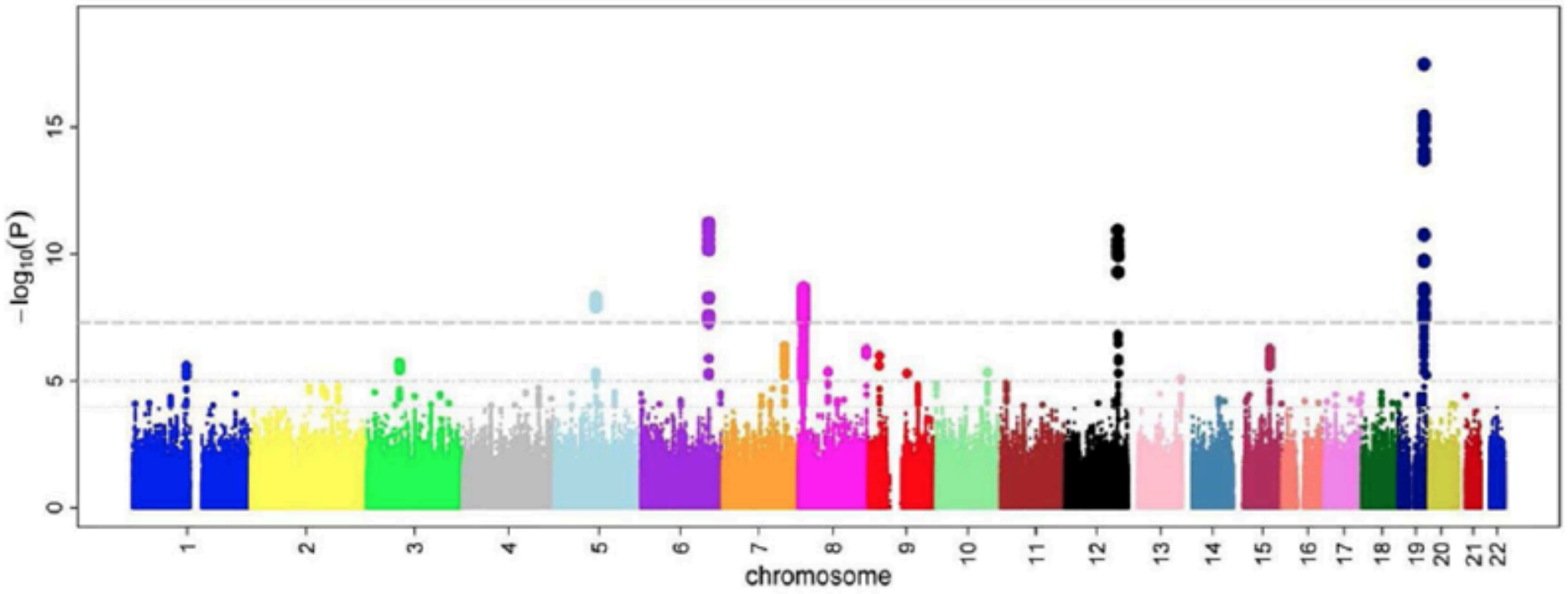
Single-SNP Association Testing

- Genotype test (2-*df* test)
- Dominant/recessive test
- Allele test
- Cochran-Armitage tend test
- Logistic regression, linear regression, proportional hazards model (cox model)

Results of GWAS: Manhattan Plot

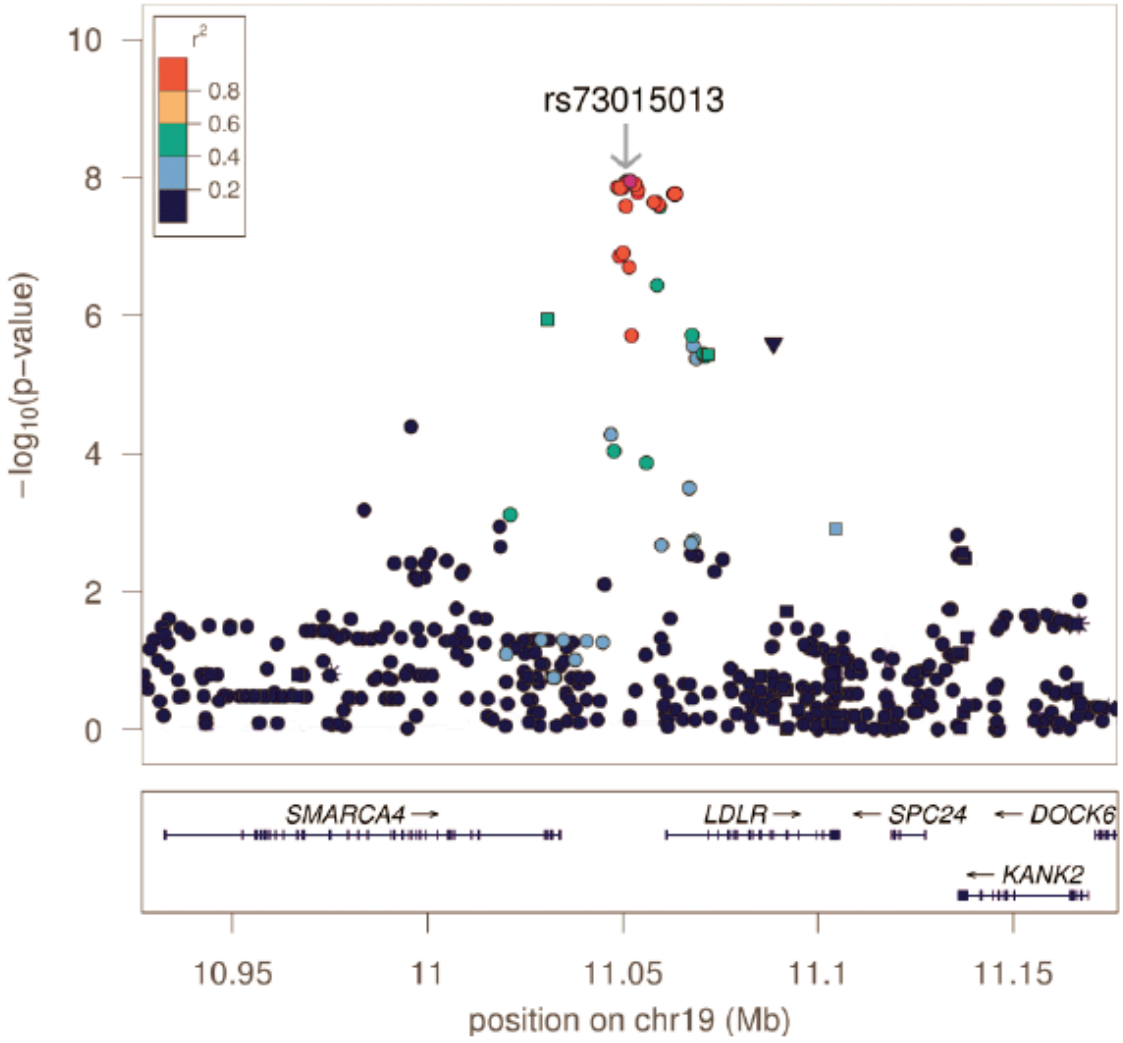
– Results plotted across genomic position on X axis and $-\log 10(\text{p-value})$ on Y axis

– To identify signal peaks



Results of GWAS: Locus Zoom Plot

– Focus on a narrower region
– Describe the linkage disequilibrium between analyzed markers



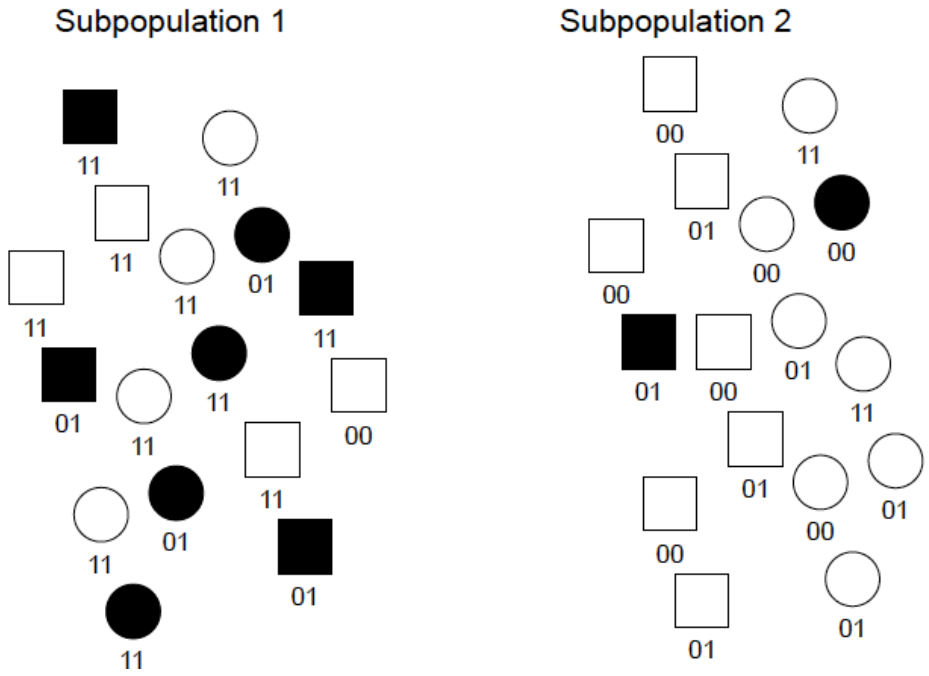
Population Stratification

Aka, population substructure, population admixture

- Allele frequencies can differ substantially between different subpopulations
- Risk of disease can also differ substantially
- Study includes a mixture of different subpopulations

Population Stratification: An Intuitive Example

Consider genotypes (coded as 00, 01 and 11) at a marker locus (e.g., SNP)



Subpopulation 1

	1	0
Case	12	4
Control	14	2

Subpopulation 2

	1	0
Case	1	3
Control	10	18

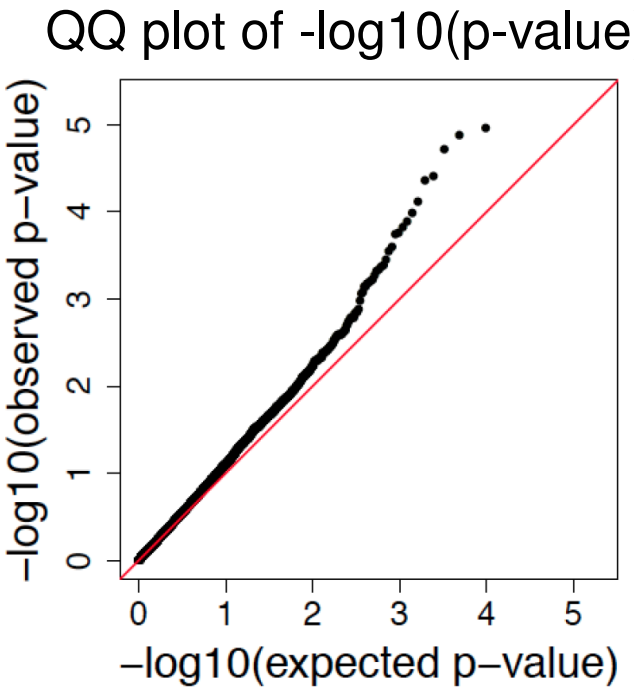
Combined

	1	0
Case	13	7
Control	24	20

A combined sample tends to show association (p = .035), even though there is not association within each subpopulation.

If Not Accounting for Population Stratification...

The number of false-positive findings can be notably higher than would be expected based on the specific significance level (i.e., inflated type I error).



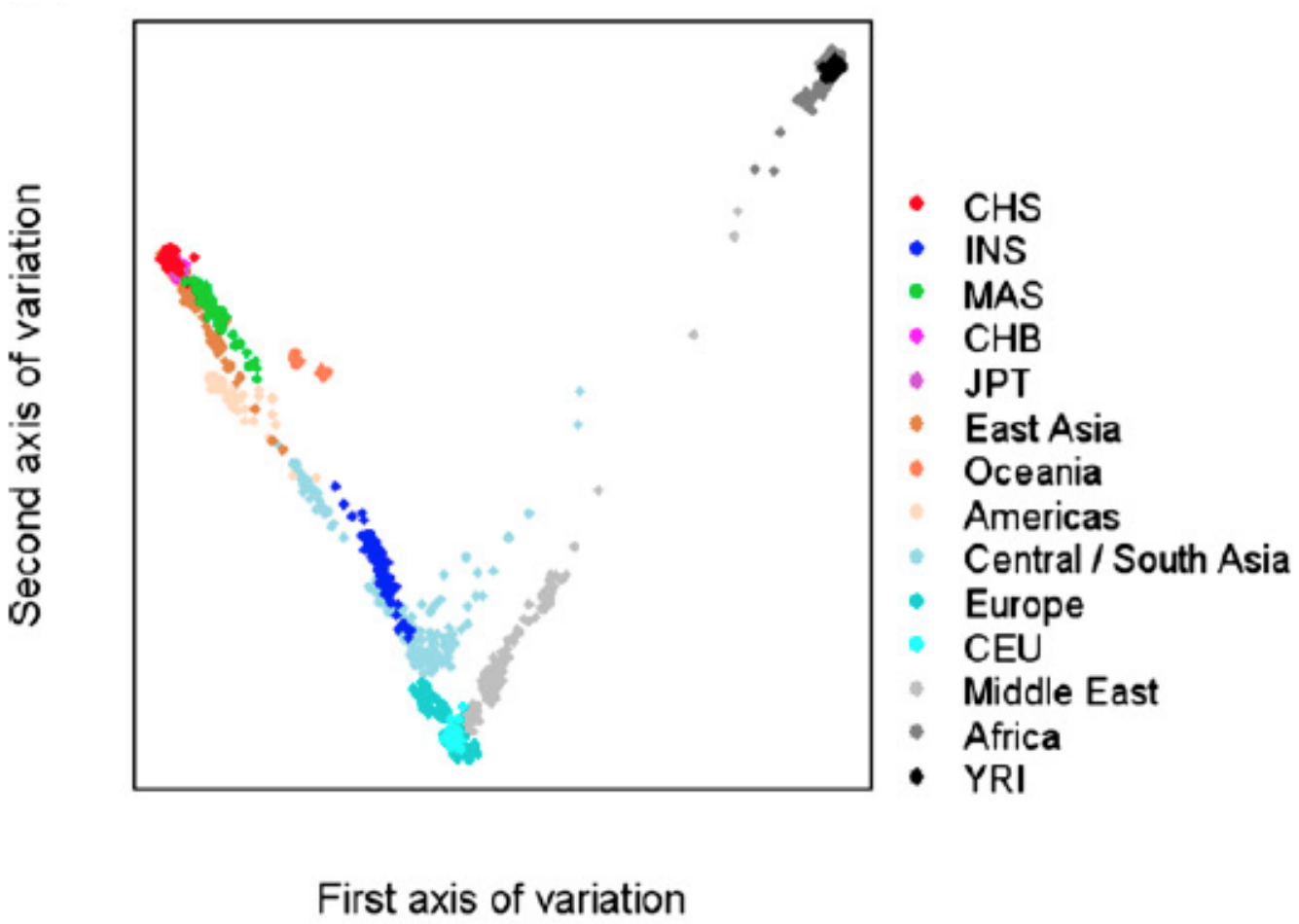
- Late departures are to be expected if there are true causal variants.
- Early departures (upwards) indicate systematic inflation of test statistics across the genome, which might reflect unaccounted population stratification, batch effects or subject relatedness.

Accounting for Population Stratification: PCA

The principal component analysis (PCA) has became one of the standard ways to adjust for population substructure in population-based designs when thousands of markers are available.

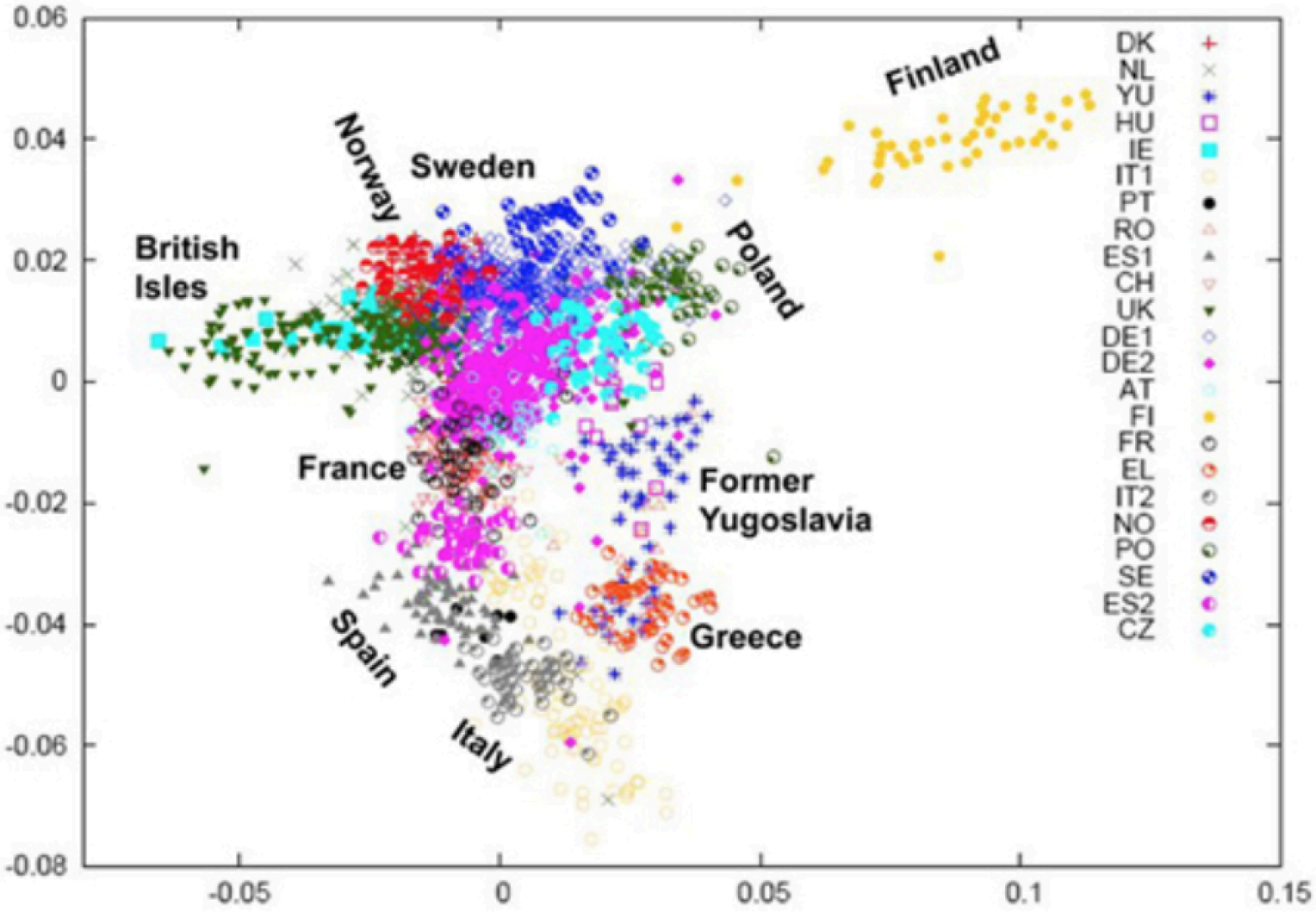
- Software: EIGENSTRAT/EIGENSOFT
- Apply principal components analysis to genotype data to infer continuous axes of genetic variation.
- Intuitively, the axes of variation reduce the data to a small number of dimensions, describing as much variability as possible; they are defined as the **top eigenvectors of a covariance matrix between samples**
- Adjust genotypes and phenotypes by amounts attributable to ancestry along each axis, via computing residuals of linear regressions. Intuitively, this creates a virtual set of matched cases and controls.
- Compute association statistics using ancestry-adjusted genotypes and phenotypes.

Axes of variation often have a geographic interpretation



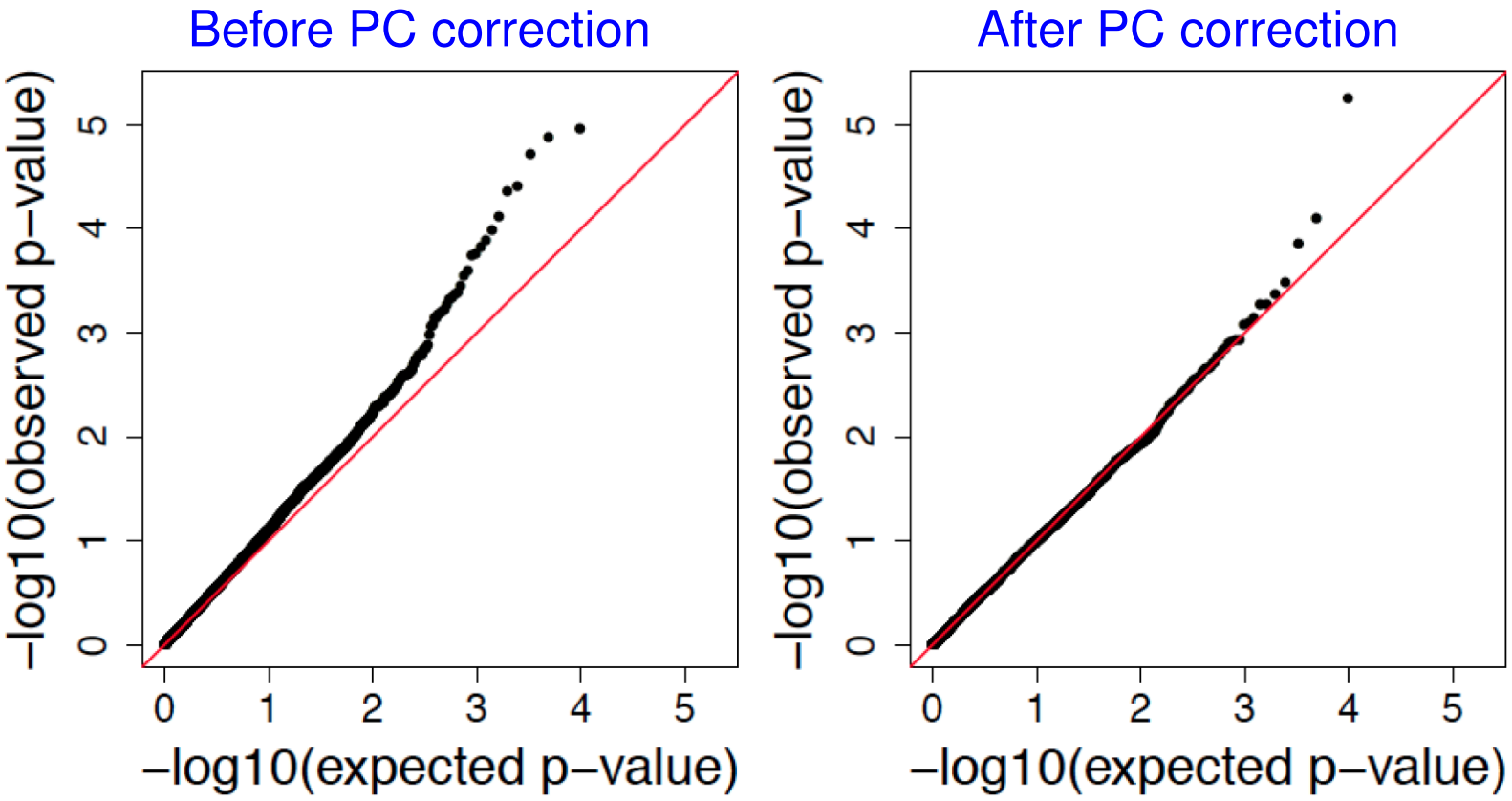
PCA: A Genetic Map of Europe

Geography determines genetics; Genetics predicts geography.



PC Correction

– Conduct association tests without and with principal component (PC)
– Rank observed $-\log 10(\text{p-value})$ from most significant to least
– Pair these with expected values from order statistic of a $\text{Uniform}(0, 1)$ distribution



For GWAS studies with $> 500,000$ markers that are tested for genetic association separately, the multiple testing problem is one of the major statistical hurdles.

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:

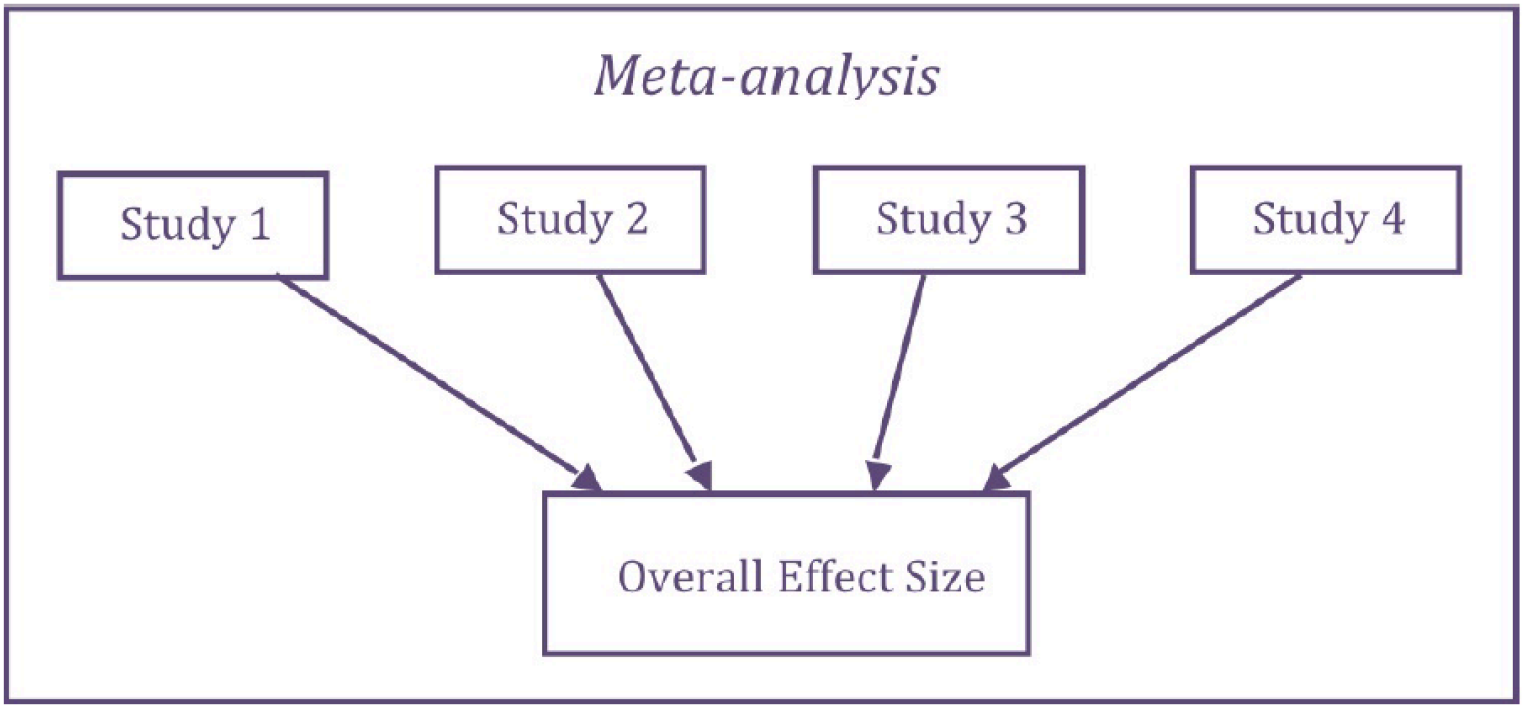
$$\begin{aligned}\alpha &= P(\text{reject at least one } H_0^{(m)} \mid H_0^{(m)} \text{ is true for all } m)\\ &= P(\bigcup_m \{\text{reject } H_0^{(m)} \mid H_0^{(m)} \text{ is true}\})\\ &\leq \sum_m P(\text{reject } H_0^{(m)} \mid H_0^{(m)} \text{ is true}) = M\alpha'\end{aligned}$$

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 $\alpha' = 2 \times 10^{-8}$.
- If the association tests are correlated, as we might expect them to be if there is much LD between the markers, then the Bonferroni method is very conservative.
- In the extreme, where rejection of one test implies rejection of all the rest, then Bonferroni method will yield the true FWER to be α/M.
- Current practice prefers a threshold of 5×10^{-8} (ad hoc).
- Obtaining a more accurate genome-wide threshold remains a difficult problem.

What is Meta-Analysis?

- Combine data from multiple studies (on the same disease) to improve power
- Based on summary results (e.g., $\hat{\beta}$, se, p-values), rather than raw individual-level data (mega-analysis)

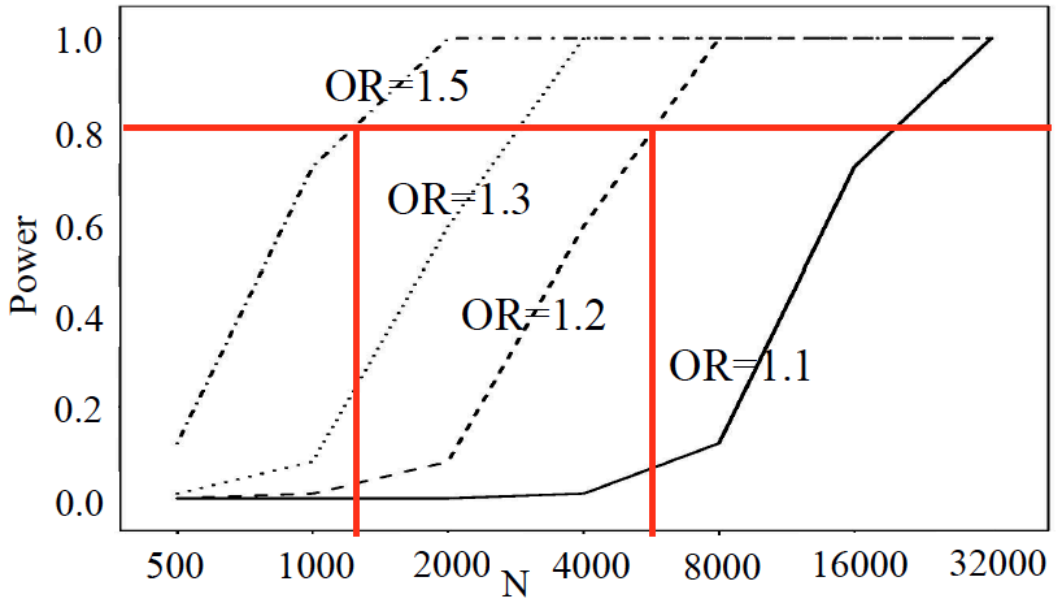


An Example: GIANT Data

- **G**enetic **I**nvestigation of **AN**thropometric **T**raits consortium
- The consortium was recently interested in identifying genetic loci associated with the extremes of height, body mass index (BMI), and waist-hip-ratio adjusted by BMI
- 51 GWAS submitted single-variant analysis results ($\hat{\beta}$, se and p-value) for case-control comparisons for ~ 2.8 million SNPs
- The 51 studies involved $\sim 160,000$ cohort members of the European ancestry

Why Meta-Analysis?

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



- Improve power
- Using summary results
  - Can bypass the privacy issues of study participants
  - Avoid cumbersome data integration from multiple studies
- Lin and Zeng (2010) showed that using summary results is statistically as efficient as using individual participant data.

How to Do Meta-Analysis?

How to combine information across studies?

- Studies generally vary in size.
- An estimate based on a sample of 100 individuals is generally assumed to be less precise or informative than that based on a sample of 1000 individuals.
- One possible weighting scheme?

Inverse-Variance Estimator for Meta-Analysis

Given summary statistics from individual studies

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

Then, we obtain

- $w_k = V_k^{-1} / \sum_{k=1}^{K} V_k^{-1}$: weight
- $\hat{\beta} = \sum_k w_k \hat{\beta}_k$: inverse-variance estimator for the common genetic effect
- $V = \sum_k w_k^2 V_k$: variance estimator for $\hat{\beta}$

Finally, the Wald test statistic

$$\frac{\hat{\beta}}{\sqrt{V}} \sim N(0, 1)$$

Fisher's Method for Meta-Analysis

Given summary statistics from individual studies

- 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

Getting Complicated ...

- Heterogeneous effects across studies

$$\beta_1 \neq \beta_2 \neq \cdots \neq \beta_K$$

- Meta-analysis of burden test based on single-variant statistics
  - $S_i = \sum_{j=1}^{m} \xi_j G_{ij}$, ξ : weights that may depend on the MAFs
  - To test $H_0 : \tau = 0$ in $g(\mathrm{E}(Y|S, X)) = \tau S + \gamma^{\mathrm{T}} X$
  - Test requires multivariate statistics
  - However, only single-variant statistics are available
  - How to recover the multivariate statistics from the single-variant statistics?

More Topics

- Missing genotypes/Untyped SNPs
- Gene-environment interaction
- Haplotype-based analysis
- Copy Number Variants (CNV) association analysis
- Trait-dependent sampling (case-control, extreme-trait sampling)
- Secondary phenotype
- . . .