

*Population Structure in  
Genetic Association Studies*

Biostatistics 666

# Sources of Association

---

- Causal association

- Genetic marker alleles influence susceptibility

best

- Linkage disequilibrium

- Genetic marker alleles associated with other nearby alleles that influence susceptibility

useful

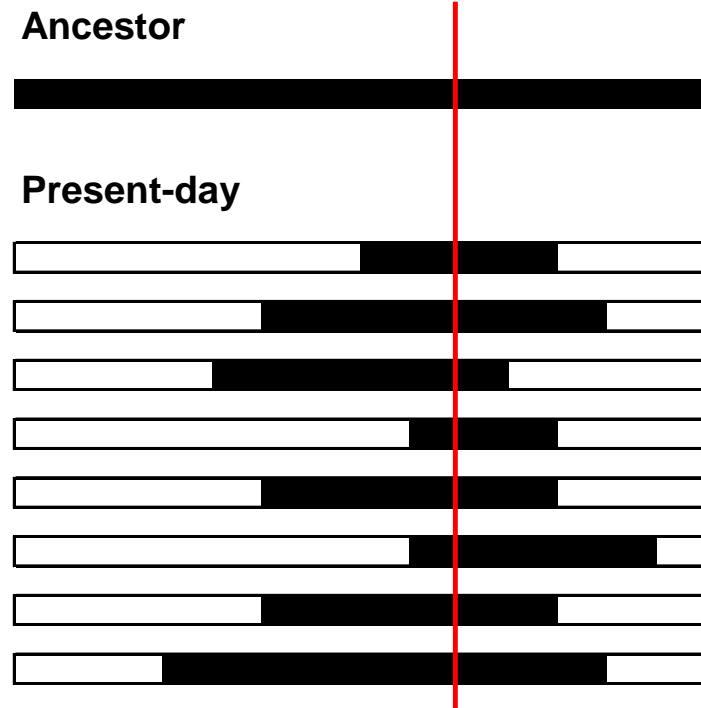
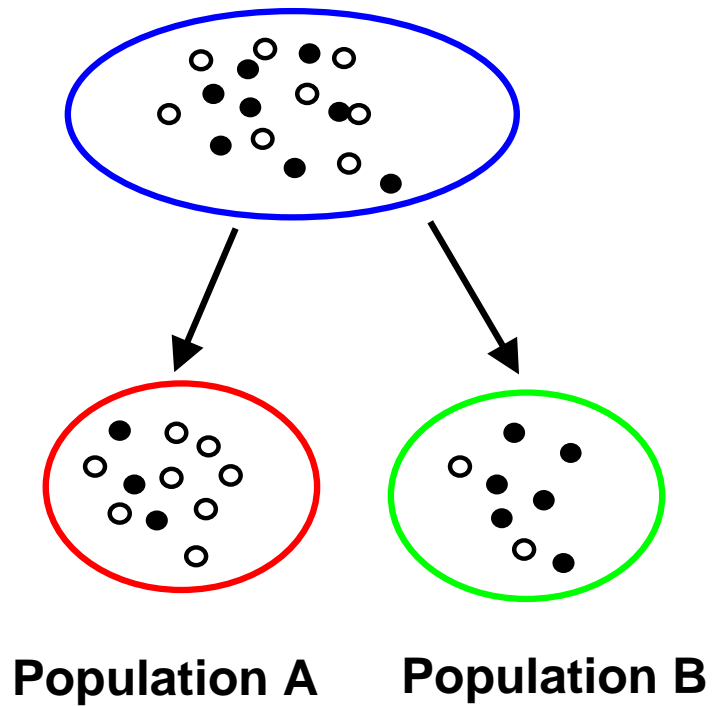
- Population stratification

- Genetic marker is unrelated to disease alleles

misleading

# Stratification vs Disequilibrium

---



# Impact of Stratification at One Locus – Numerical Example

	<u>Sample</u>		
	Population 1	Population 2	Combined
<b><u>Allele Frequencies</u></b>			
$p_1$	0.20	0.80	0.50
$p_2$	0.80	0.20	0.50
<b><u>Genotype Frequencies</u></b>			
$p_{11}$	0.04	0.64	0.34 (0.25 expected)
$p_{12}$	0.32	0.32	0.32 (0.50 expected)
$p_{22}$	0.64	0.04	0.34 (0.25 expected)

Notice the excess of homozygotes and deficit of heterozygotes.

# Impact of Stratification at Two Loci – Numerical Example

---

**Population A**

	B <sub>1</sub>	B <sub>2</sub>
A <sub>1</sub>	160	160
A <sub>2</sub>	40	40

Chi<sup>2</sup> = 0.0

**Population B**

	B <sub>1</sub>	B <sub>2</sub>
A <sub>1</sub>	160	40
A <sub>2</sub>	160	40

Chi<sup>2</sup> = 0.0

**Combined Population**

	B <sub>1</sub>	B <sub>2</sub>
A <sub>1</sub>	320	200
A <sub>2</sub>	200	80

Chi<sup>2</sup> = 7.83

## The Stratification Problem

---

- If phenotypes differ between populations
- And allele frequencies have drifted apart
- Unlinked markers exhibit association
- Not very useful for gene mapping!

# Stratification

---

- Due to non-random mating
  - Eg. Mating based on proximity or culture
- Allele frequencies drift apart in each group
  - Eg. Allele frequency differences at many genes between African-Americans and Caucasians
- If disease prevalence also differs, association studies can produce misleading results
  - Eg. Glaucoma has prevalence of ~2% in elderly Caucasians, but ~8% in African-Americans

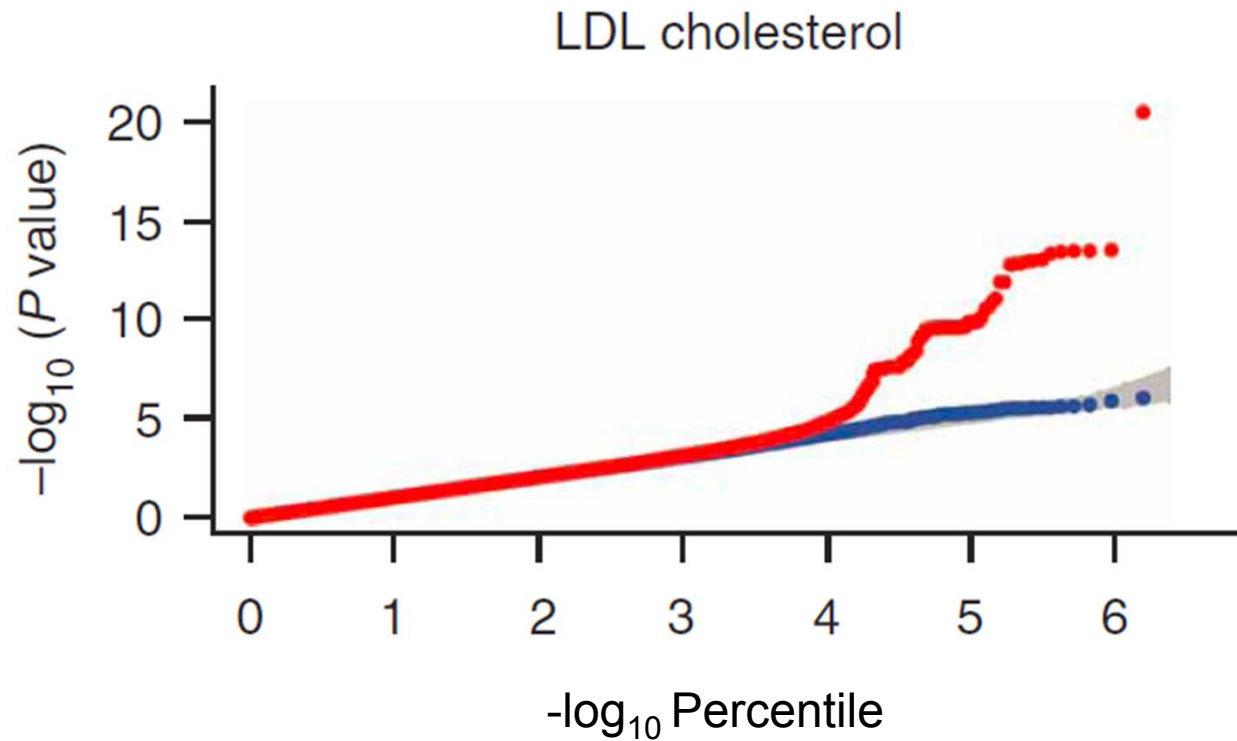
## Possible solutions

---

- Collect a better matched sample
- Identify population groupings
  - Using self reported ethnicity or genetic markers
  - Carry out association analysis within each group
- Account for inflated false-positive rate
- Use family based controls

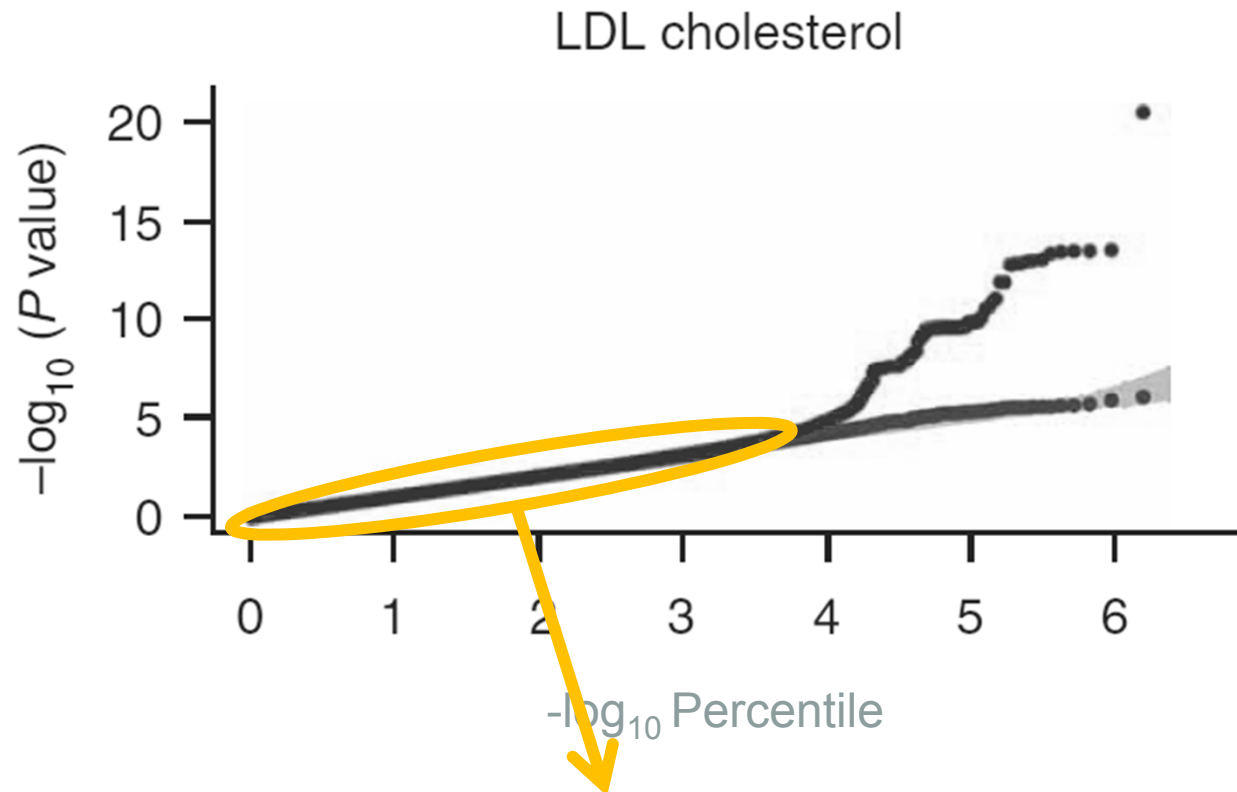


# Q-Q Plots: A Useful Diagnostic



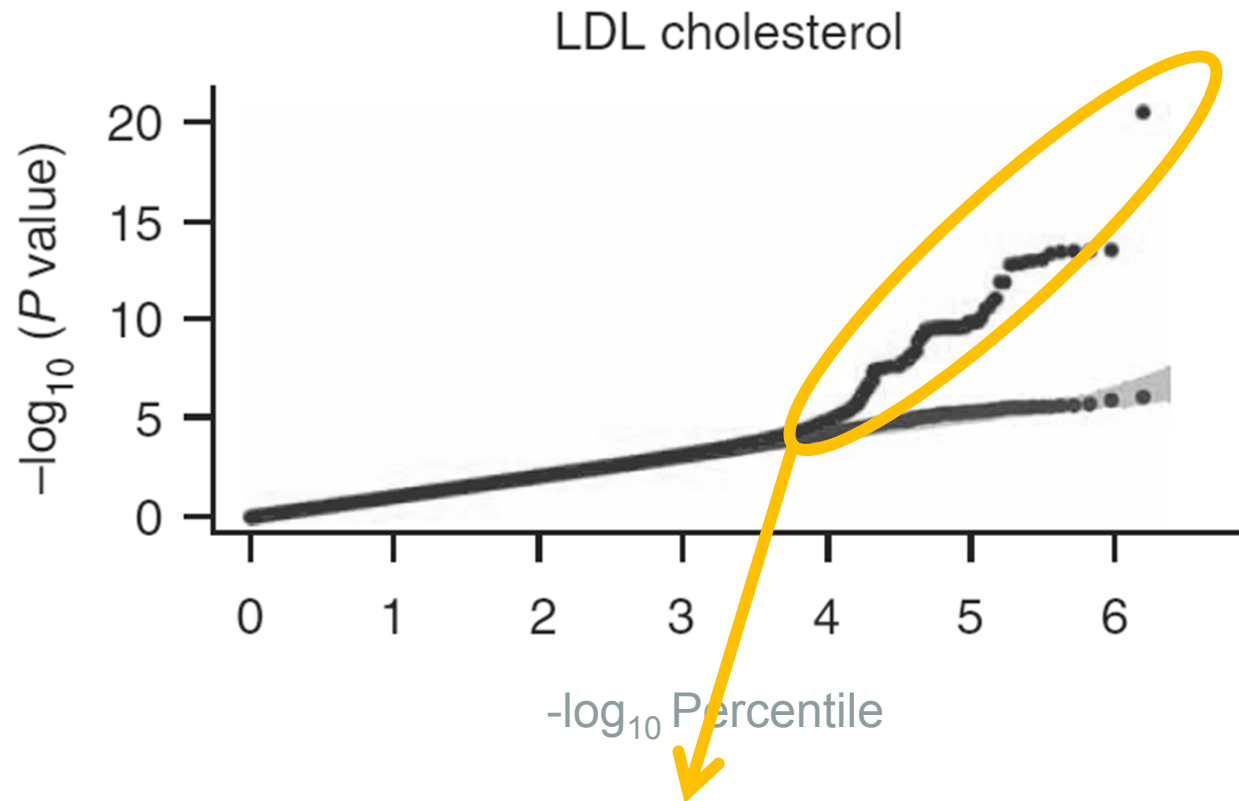
Comparison of expected and observed p-values in a study of LDL cholesterol for all markers (**red**) and for markers in regions not known to impact LDL levels (**blue**).

# Q-Q Plots: A Useful Diagnostic



In genomewide studies, most markers show no association with the trait and, therefore, very similar observed and expected p-values.

# Q-Q Plots: A Useful Diagnostic



In genomewide studies, only a small subset of markers is expected to show association with any particular trait.

# Family-Based Tests

---

- Use family information to define well-matched controls
- Distinguish “true” association from population stratification

## Trio Families

---

- Families with two genotyped parents
- One affected child
- Calculate distribution of child genotypes conditional on parental data
  - Focus on children with heterozygous parents

# The Spielman TDT

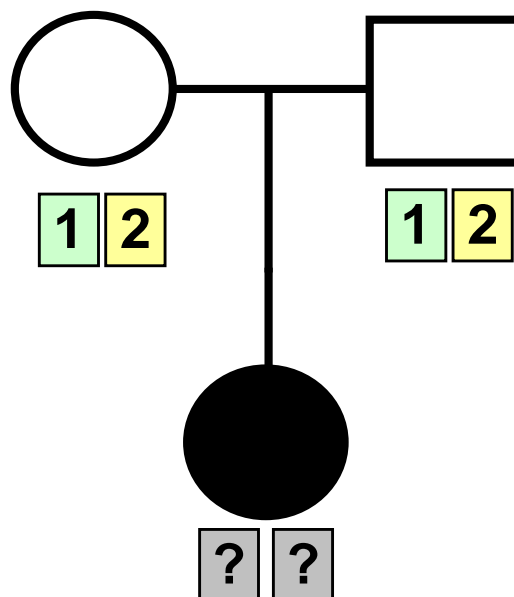
---

- Traditional case-control
  - Compare allele frequencies in two samples
    - Cases and controls must be one population
- Heterozygous parents
  - Parental alleles are the study population
  - Population allele frequencies fixed
    - 50:50, independent of original
  - Check proportion of each allele transmitted to affected offspring

## Basic TDT

---

- Is allele consistently transmitted from heterozygotes?
- Affirmative answer requires
  - Allele is associated
  - Allele is linked
  - Or we have a false-positive



## The TDT statistic

---

	Transmitted 1	Transmitted 2
Not-Transmitted 1	a	b
Not-Transmitted 2	c	d

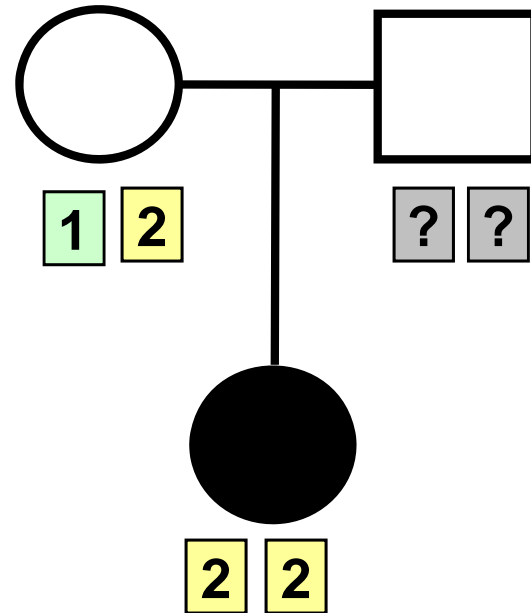
$$\frac{(b - c)^2}{b + c} \sim \chi_1^2$$



# Caution: Parental genotypes crucial!

---

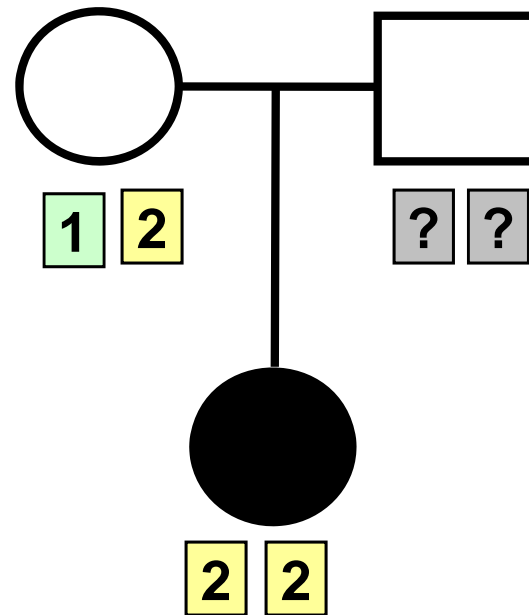
- It seems we can deduce transmitted allele...
- However, this leads to bias...
  - Why?



# Caution: Parental genotypes crucial!

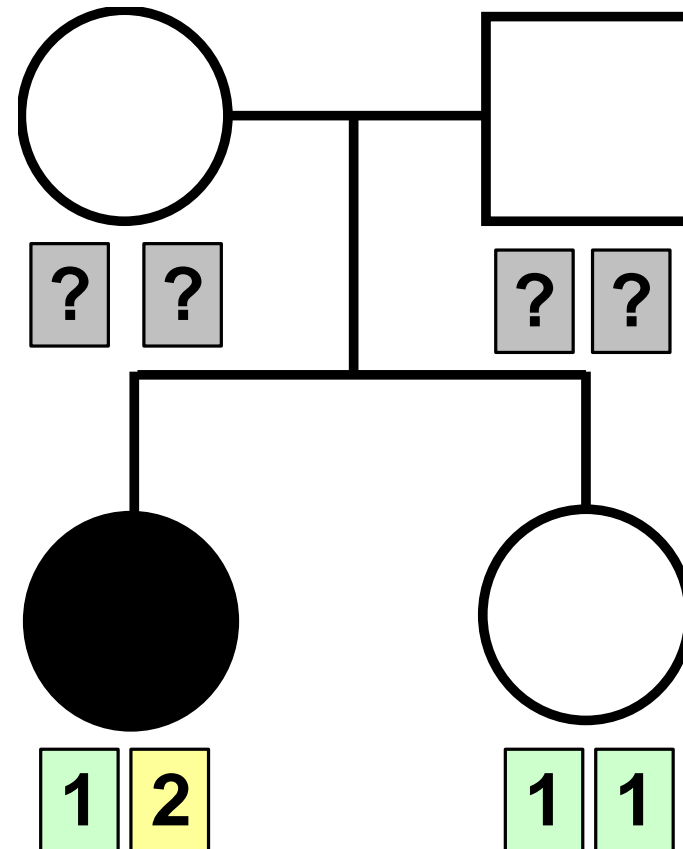
---

- Probability of inferring transmitted genotype depends on population allele frequencies
- Expected ratio of observed transmissions no longer 50/50.



# The Sib-TDT

- Parents may be missing
  - eg. late onset conditions
- Compare alleles that differ between siblings
  - When sib genotypes differ, which allele is carried by affected sib?



## The Sib-TDT statistic

---

	Affected has 1	Affected has 2
Unaffected has 1	a	b
Unaffected has 2	c	d

$$\frac{(b - c)^2}{b + c} \sim \chi_1^2$$

## Further Extensions

---

- The TDT can be further extended to model genotype, rather than allele, distributions
  - Schaid (1999) *Genet Epidemiol* **16**:250-260
- The TDT can be extended to accommodate different family structures

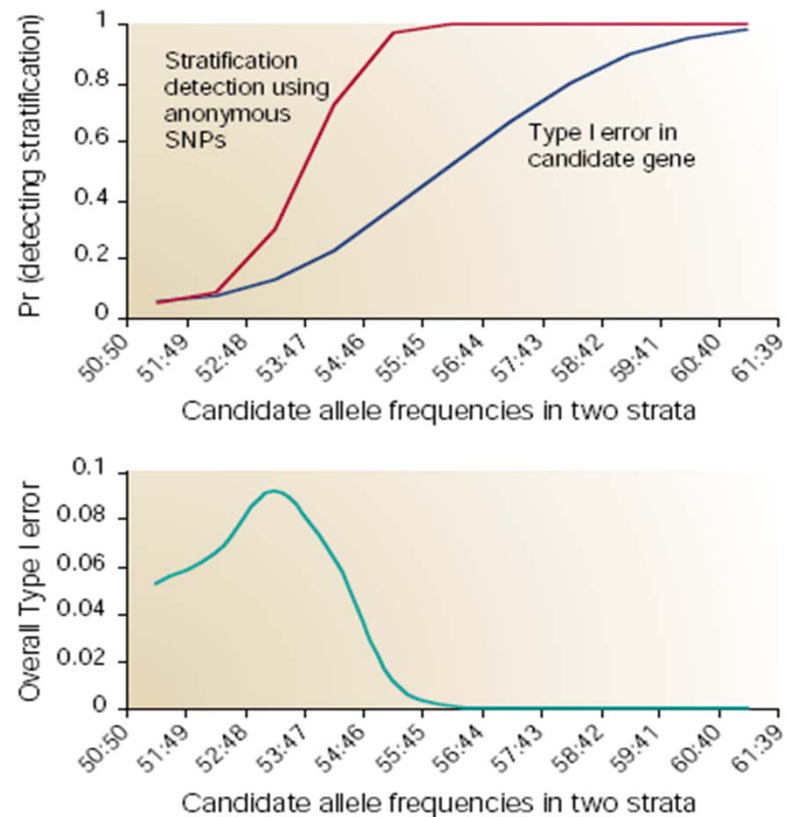
## What if families are not available?

---

- Test null markers across genome
  - Markers that are unlikely to be associated
    - Markers that are outside genes
    - Markers in genes that are unlikely to be involved
- Initially, 50 markers suggested as minimum
- Now, typical to use 100,000 SNPs or more (from genomewide studies) as null

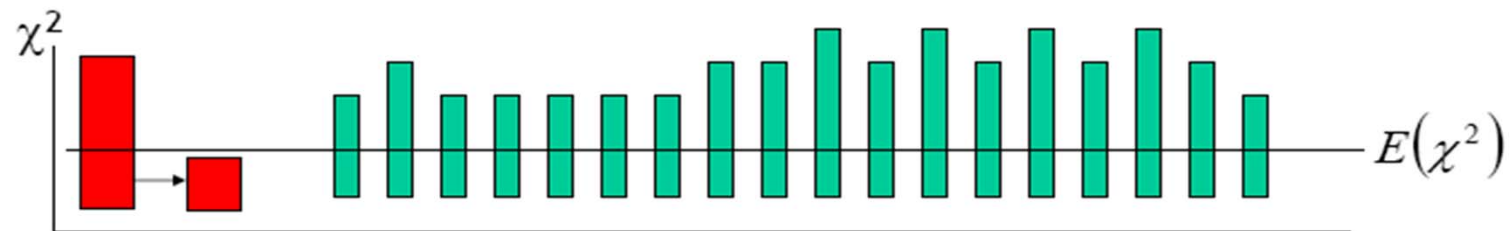
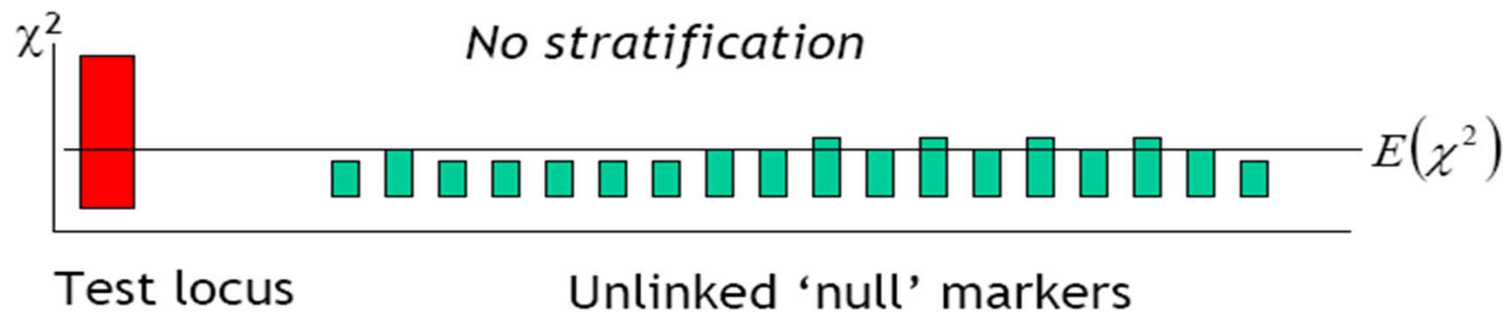
# What if “null markers” reject null?

- Early suggestion:
  - Reject Association
  - Pritchard and Rosenberg (1999)  
*Am J Hum Genet*  
**65:220-228**
- What might be a better approach?



(Figure from Cardon and Bell, *Nature Reviews Genetics*, 2001)

# Genomic Control



*Stratification → adjust test statistic*

(Figure courtesy Shaun Purcell, Harvard, and Pak Sham, HKU)



## Define Inflation Factor

---

- Compute chi-squared for each marker
- Inflation factor  $\lambda$ 
  - Average observed chi-squared
  - Median observed chi-squared / 0.456
  - Should be  $\geq 1$
- Adjust statistic at candidate markers
  - Replace  $\chi^2_{\text{biased}}$  with  $\chi^2_{\text{fair}} = \chi^2_{\text{biased}} / \lambda$

## Questions

---

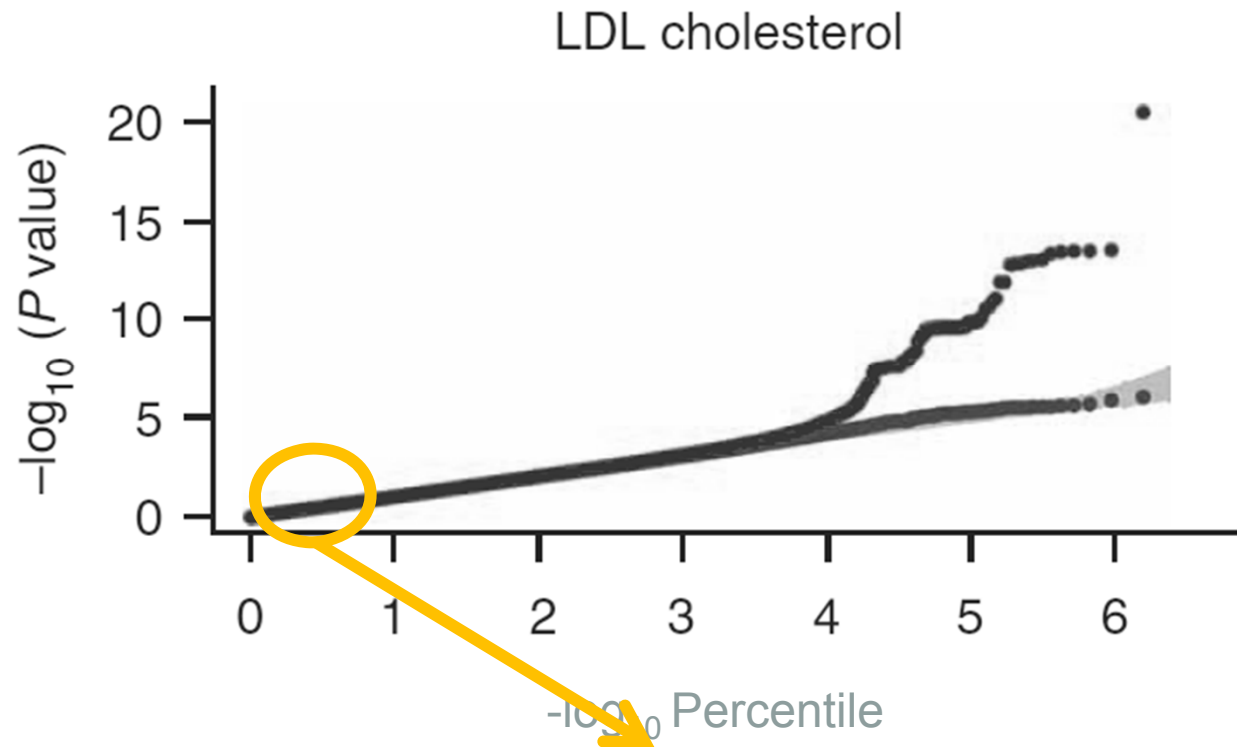
- When defining the inflation factor  $\lambda$  ...
- Why do we use a lower bound of 1?
- What might be the advantages of using the median rather than the mean?

# Applying Genomic Control

---

- Simple and convenient approach...
  - Easily adapted to other test statistics, such as those for quantitative trait and haplotype tests
- Under the null, stratification always inflates evidence for association...
  - Is this also true under the alternative?
  - What might be the consequences?

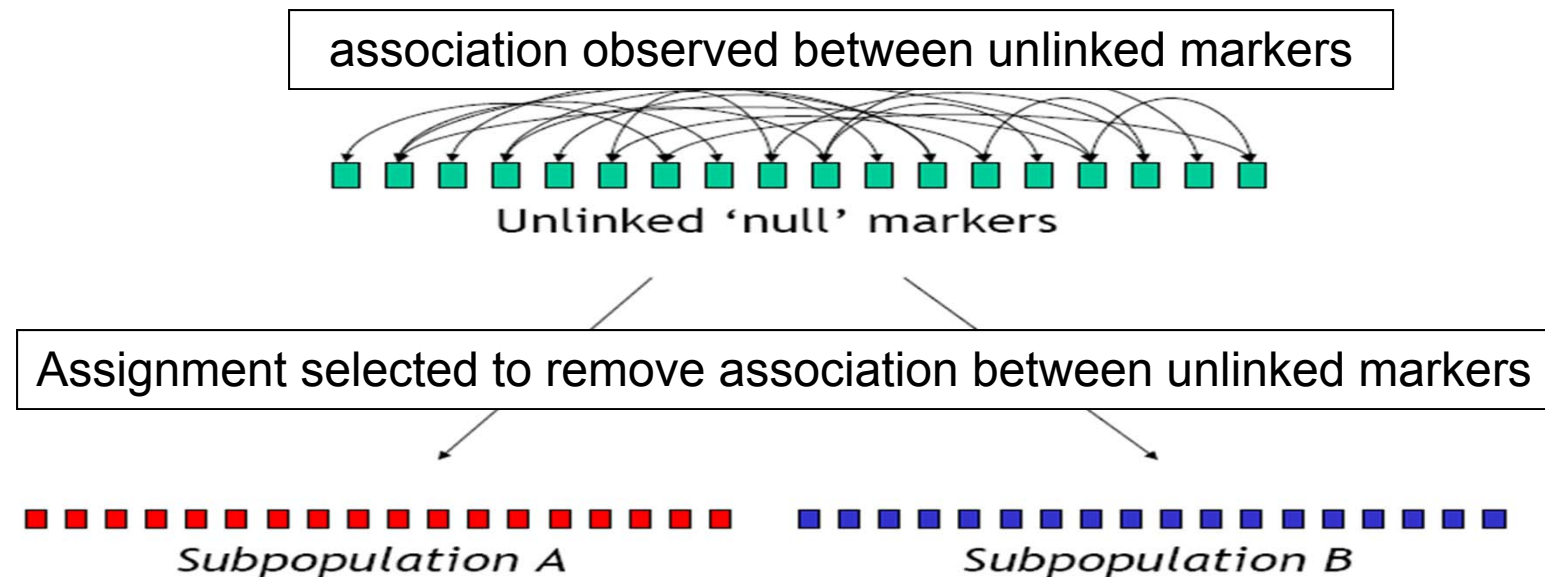
# Q-Q Plots: A Useful Diagnostic



The genomic control value examines markers with little evidence for association. If these large p-values were to deviate from expected, there is a problem! In this case,  $\lambda=1.02$ .

# Structured Association

- Use unlinked markers to assign individuals to subpopulations then...
  - Test for association within each subpopulation
  - Test for association while conditioning on subpopulation



(Figure courtesy of Shaun Purcell and Pak Sham)

## Some Attractive Features

---

- Allows for flexibility in association test
- Describing subpopulations can be useful
- Does not assume constant population differentiation across the genome

# Simple Mixture Distribution

---

$$p(x|\boldsymbol{\pi}, \boldsymbol{\Phi}) = \pi_1 p(x|\phi_1) + \dots + \pi_k p(x|\phi_k)$$

- $p()$  are the probability functions
- $x$  are the observed genotypes
- $\pi$  are the mixture proportions for subpopulations
- $\phi$  are allele frequencies for each subpopulation
- $k$  is the number of components

# Maximum Likelihood Approach

---

- Find the parameters that maximize the likelihood for the entire sample

$$L = \prod_j p(x_j | \boldsymbol{\pi}, \boldsymbol{\Phi})$$
$$\ell = \sum_j \log p(x_j | \boldsymbol{\pi}, \boldsymbol{\Phi})$$

- Prior for the allele frequencies may also be included in the likelihood
- Likelihood can be maximized using a Gibbs sampler or E-M algorithm



# Classifying Individuals

---

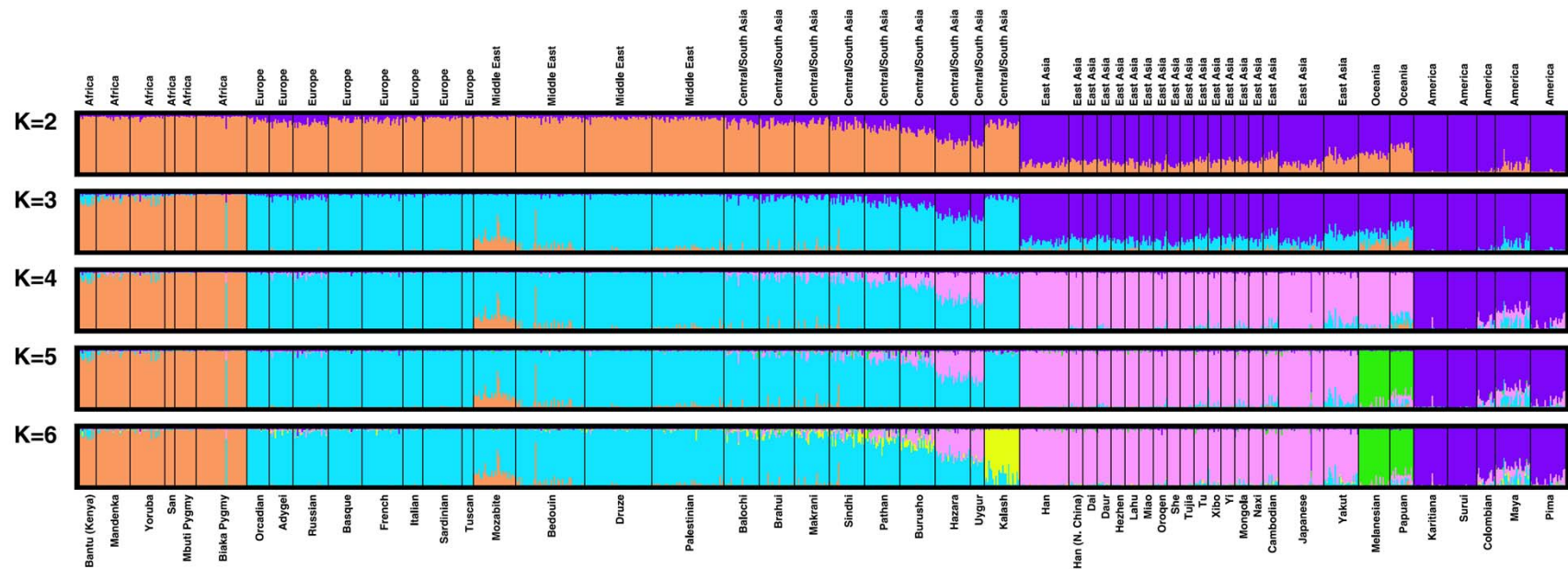
- Let  $Z_j$  be the population membership for individual  $i$

$$\Pr(Z_j = i | \boldsymbol{\pi}, \boldsymbol{\Phi}) = \pi_i$$

$$\Pr(Z_j = i | x_j, \boldsymbol{\pi}, \boldsymbol{\Phi}) = \frac{\pi_i p(x_j | \phi_i)}{\sum_l \pi_l p(x_j | \phi_l)}$$

- Results from the application of Bayes' theorem

# Classification of Individuals Using Genetic Data



Genetic classification of individuals across several populations, as the number of modeled subpopulations (K) changes.

(from Rosenberg et al, *Science*, 2002)

# Testing for Association

---

- Once individuals are classified, there is leeway in selecting association test:
  - Test within each subpopulation
  - Test within each subpopulation, combine results
  - Model effect of population membership
    - For example, using covariates in a regression model

## Refinements to the Model

---

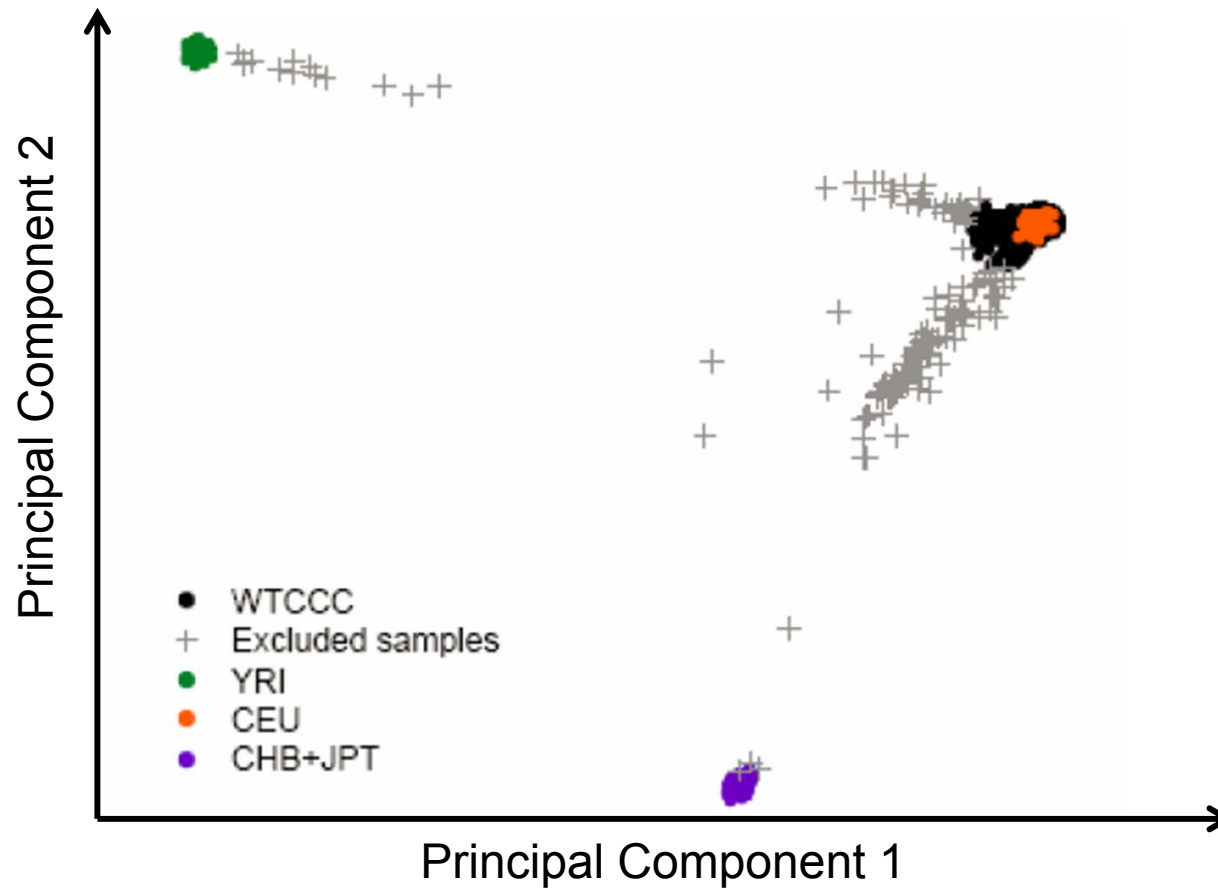
- Allowing for admixture within individuals
- Setting up a prior for allele frequencies that favors similar frequencies across populations
- Allowing for different tiers of population structure

# Principal Components Analysis

---

- Model each genotype as quantitative variable
  - Number of copies of the minor allele
- Identify small number of principal components (PC)
  - Linear combinations of observed genotype scores
  - Selected to explain variation in genotype scores
  - Typically, one to ten PC are modeled
- Allow population structure to be visualized
- Can be used as covariates in association analysis

# Principal Components For a Large Association Study



# Principal Components vs. Structured Association

---

- Both use genotypes to find similar individuals
  - Both can be combined with genomic control analysis
- Tuning structured association methods
  - Select number of subpopulations
  - Fit model parameters
  - Minimize genomic control value
- Tuning principal component analysis
  - Decide how many PC to include as covariates
  - Computationally less demanding
  - Minimize genomic control value

# References

---

- **Genomic Control for Association Studies**
  - Devlin and Roeder (1999) *Biometrics* **55**:997-1004
  - Pritchard and Rosenberg (1999) *Am J Hum Genet* **65**:220-228
- **Methods for Inferring Population Structure**
  - Pritchard, Stephens and Donnelly, 2000. *Genetics* **155**: 945-959
- **Principal Components Analysis**
  - Price et al (2006) *Nature Genetics* **38**:904-909
- **Transmission Disequilibrium Tests**
  - Spielman et al (1993) *Am J Hum Genet* **52**:506-16 (trios)
  - Curtis (1997) *Ann Hum Genet* **61**:319-33 (sibling pairs)