Consequences of Population Structure

Biostatistics 666

Sources of Association

- Causal association
  - Genetic marker alleles influence susceptibility

best

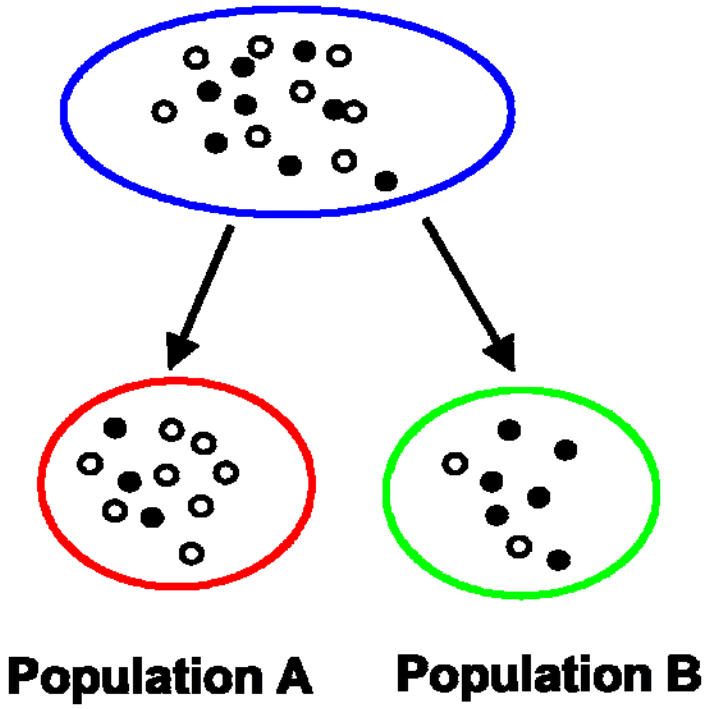
- Linkage disequilibrium
  - Genetic marker alleles associated with other nearby alleles that influence susceptibility

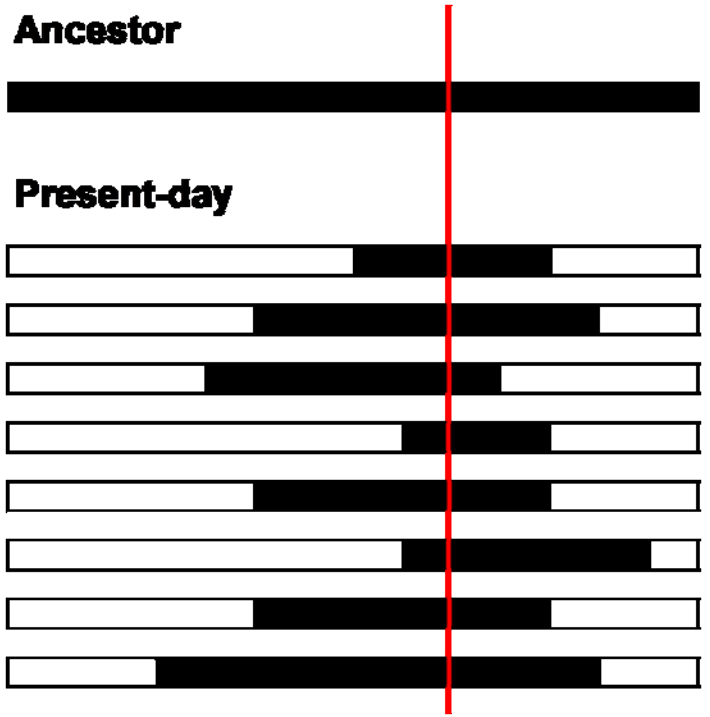
useful

- Population stratification
  - Genetic marker is unrelated to disease alleles

misleading

Stratification *vs* Disequilibrium





Impact of Stratification at One Locus – Numerical Example

	Sample		
	Population 1	Population 2	Combined
Allele Frequencies			
p_1	0.20	0.80	0.50
p_2	0.80	0.20	0.50
Genotype Frequencies			
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_1	B_2
A_1	160	160
A_2	40	40

$Chi^2 = 0.0$

Population B

	B_1	B_2
A_1	160	40
A_2	160	40

$Chi^2 = 0.0$

Combined Population

	B_1	B_2
A_1	320	200
A_2	200	80

$Chi^2 = 7.83$

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

Family-Based Tests

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

Trio Families

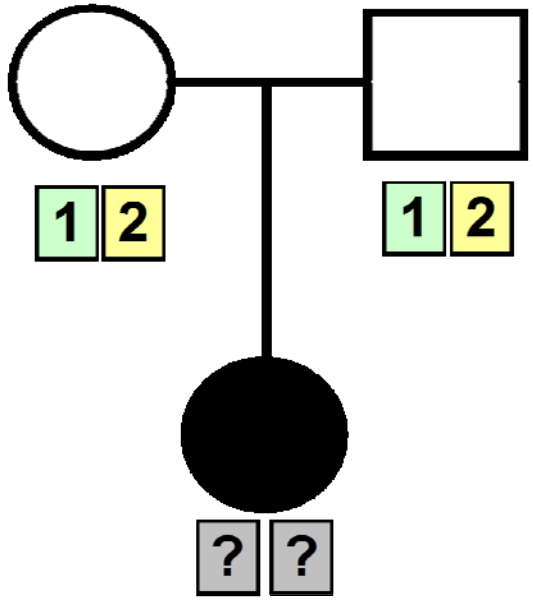
- Families with two genotyped parents
- One affected child
- Calculate distribution of child genotypes conditional on parental data
  - Focus on children with heterozygous parents

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



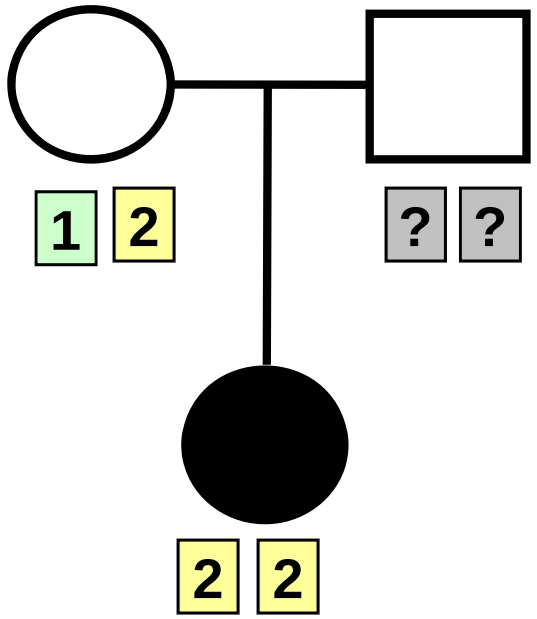
The TDT statistic

	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!

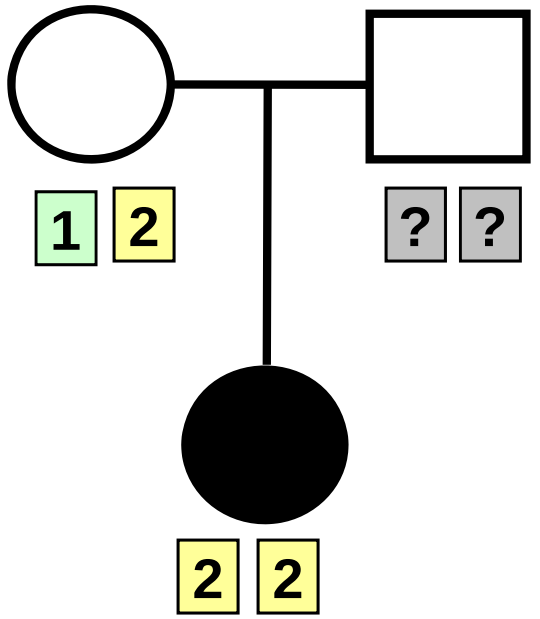
- It seems we can deduce transmitted allele…
- However, this leads to bias…
  - Why?



Caution: Parental genotypes crucial!

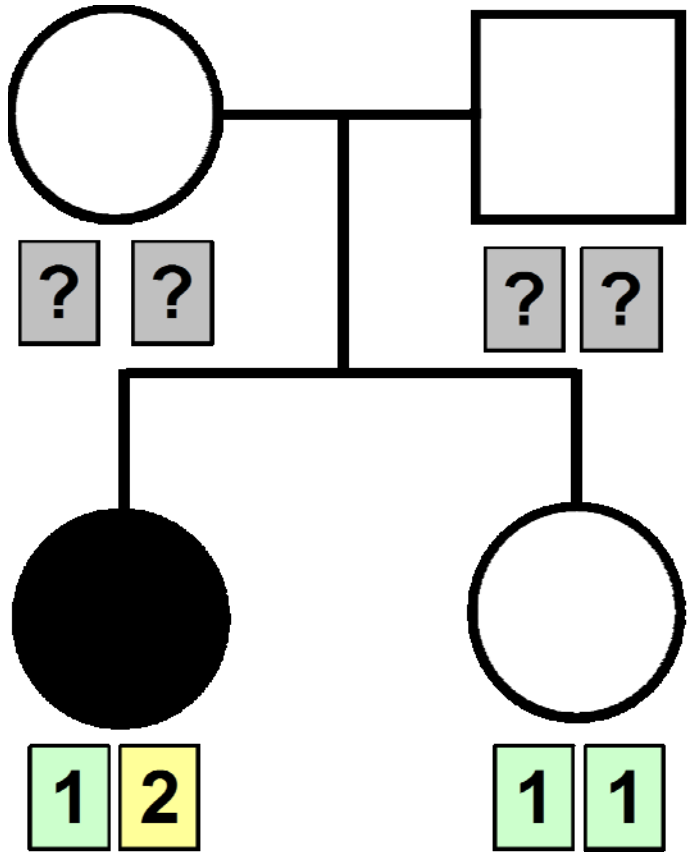
- Probability of inferring transmitted genotype depends on population allele frequencies

- 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

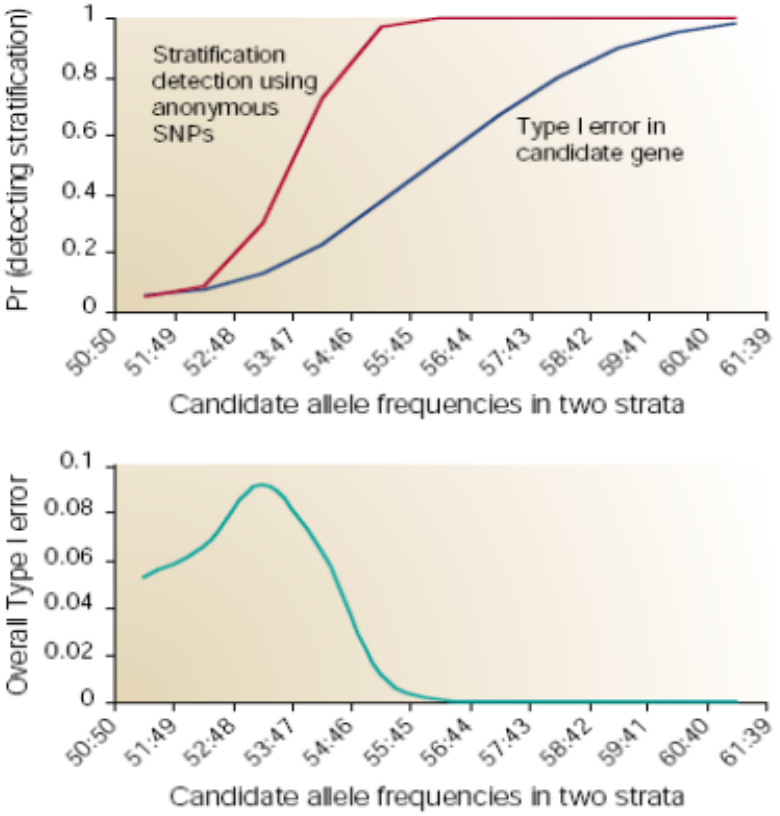
- The TDT can be further extended to model genotype, rather than allele, distributions
  - Schaid (1999) *Genet Epidemiol* **16**:250-260
- The TDT can be extended to accommodate different family structures

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

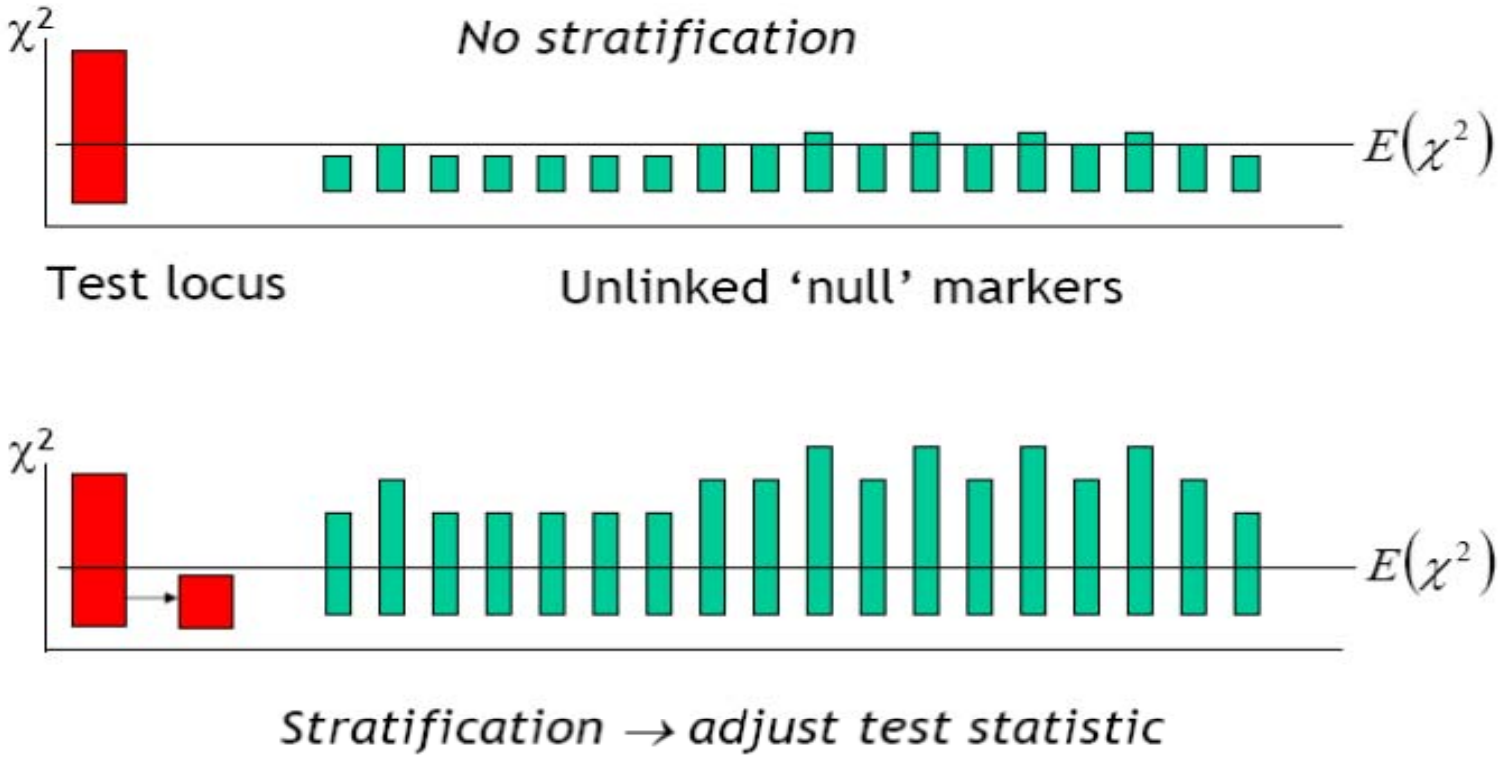
What if “null markers” reject null?

- 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



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

Define Inflation Factor

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

Questions

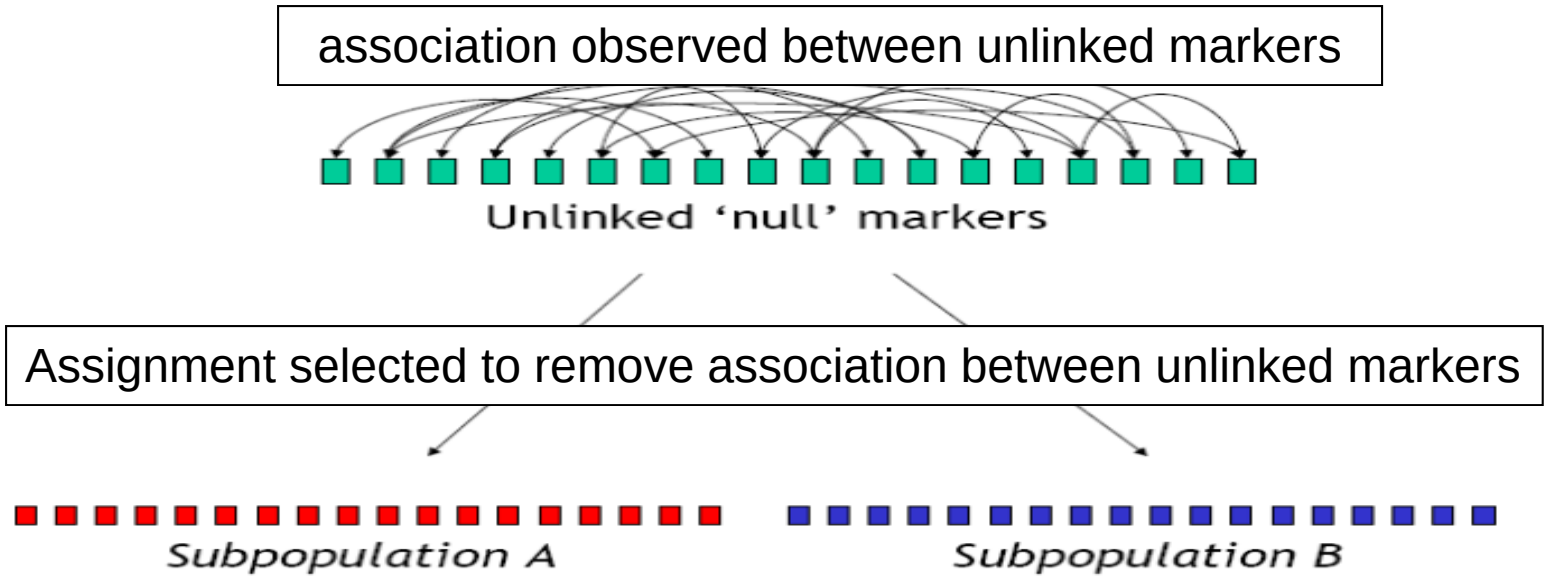
- When defining the inflation factor λ …
- Why do we use a lower bound of 1?
- What might be the advantages of using the median rather than the mean?

Applying Genomic Control

- Simple and efficient method…
  - 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?

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

$$p(x|\boldsymbol{\pi},\boldsymbol{\Phi}) = \pi_1 p(x|\phi_1) + ... + \pi_k p(x|\phi_2)$$

- $p()$ are the probability functions
- x are the observed genotypes
- π are the mixture proportions for subpopulations
- ϕ 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

$$L = \prod_j p(x_j \mid \boldsymbol{\pi}, \boldsymbol{\Phi})$$

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

- 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

Testing for Association

- Once individuals are classified, there is leeway in selecting association test:
  - Test within each subpopulation
  - Test within each subpopulation, then combine results
  - Use a test that conditions on population membership
    - e.g. a regression model with appropriate covariates

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

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
- 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)