Lecture 5: Family-Based Association Analysis

Introduction

Contact Information:

Jingjing Yang
Assistant Professor of Human Genetics
School of Medicine
Whitehead Biomedical Research Building, Suite 305K
jingjing.yang@emory.edu; 404-727-3481

See videos about introductions to the basic principles of genetics as provided by 23&me: https://www.23andme.com/gen101/

Linkage Analysis

- **Key Goals of Linkage Analysis:**
  - Locating the disease susceptibility locus (DSL) for dichotomous traits
    - Parametric linkage analysis: Recombinant counting; LOD Score of recombination rate
    - Nonparametric linkage analysis based on IBD/IBS
  - Quantitative traits: Variance component models (also used to estimate heritability)
- **Procedure of Linkage Analysis:**
  - Recruit families with multiple cases of disease of interest
  - Perform linkage analysis to find candidate chromosomal interval
  - Fine-map by adding additional markers to linkage study

Linkage Analysis

- **Successful** for identifying genes for
  - Monomorphic Mendelian disease: Cystic Fibrosis, Sickle Cell, Huntington's Disease
  - "Mendelian forms" of common diseases: BRCA1 and BRCA2 for breast and ovarian cancer
- **Linkage analysis methods** were developed in the ages when we only had low-resolution genotype data (e.g., Microarray Chip data).
- Even if linkage was identified, the candidate regions were often too large to locate the disease susceptibility locus (DSL).
- **Want to study common/complex traits** with high-resolution genotype data (Next-generation sequencing data)?

Association Analysis

- **Key Goals of Association Analysis**
  - Test associations between each locus and the interested trait
  - Understand the biological function of these associated loci (Challenging)
- **Rationale:**
  - Most traits and diseases have a complex genetic etiology
  - Considerable genetic heterogeneity
  - Genetic effects of common variants on complex traits are often considerably smaller than what you would expect for high-penetrant variants

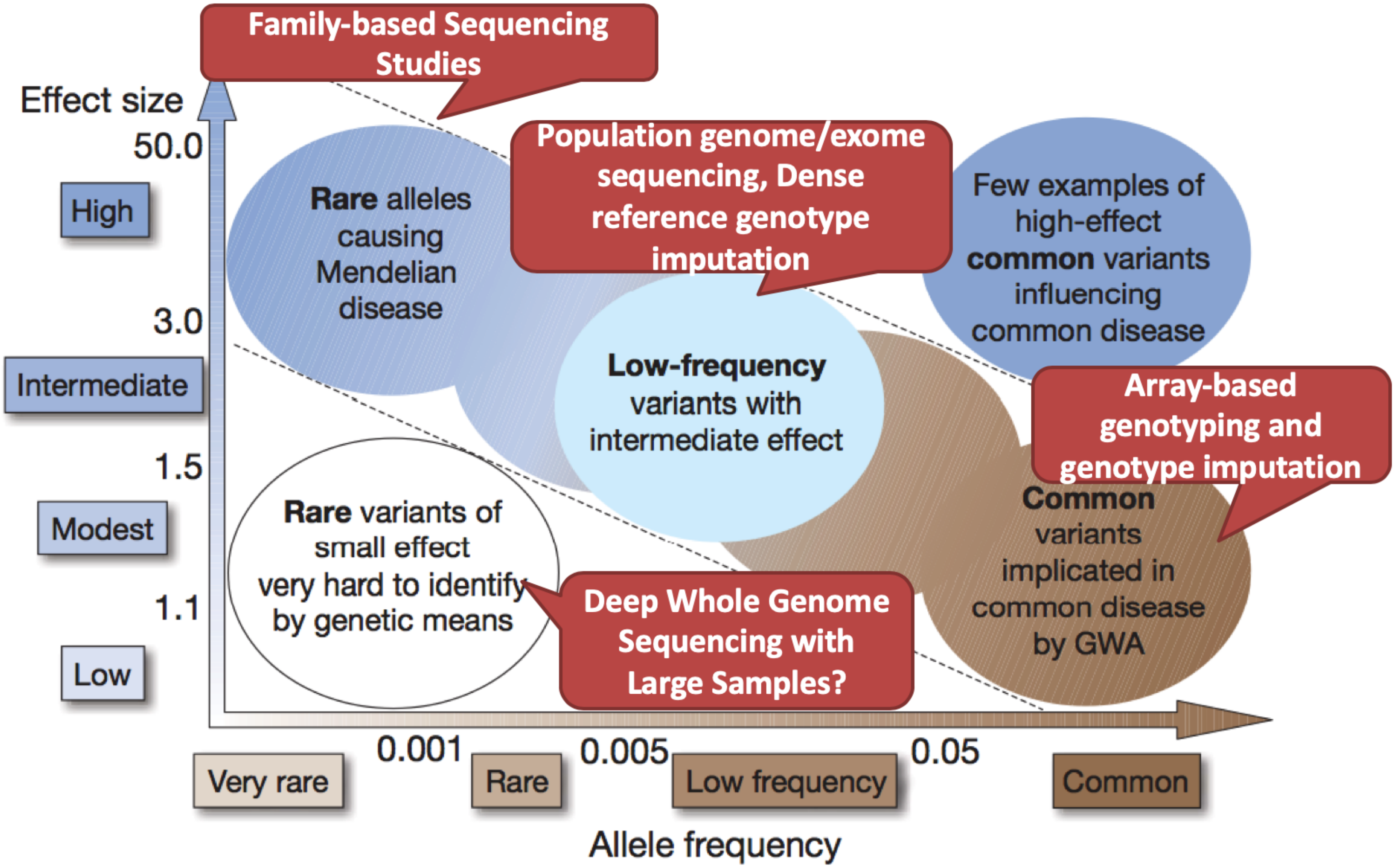
  Common Variants: Generally defined as genetic variants (e.g., SNPs) with minor allele frequency (MAF) $> 5\%$.

  Association Analysis:

  - Quantitative traits
  - Dichotomous traits (i.e., Case-control studies)

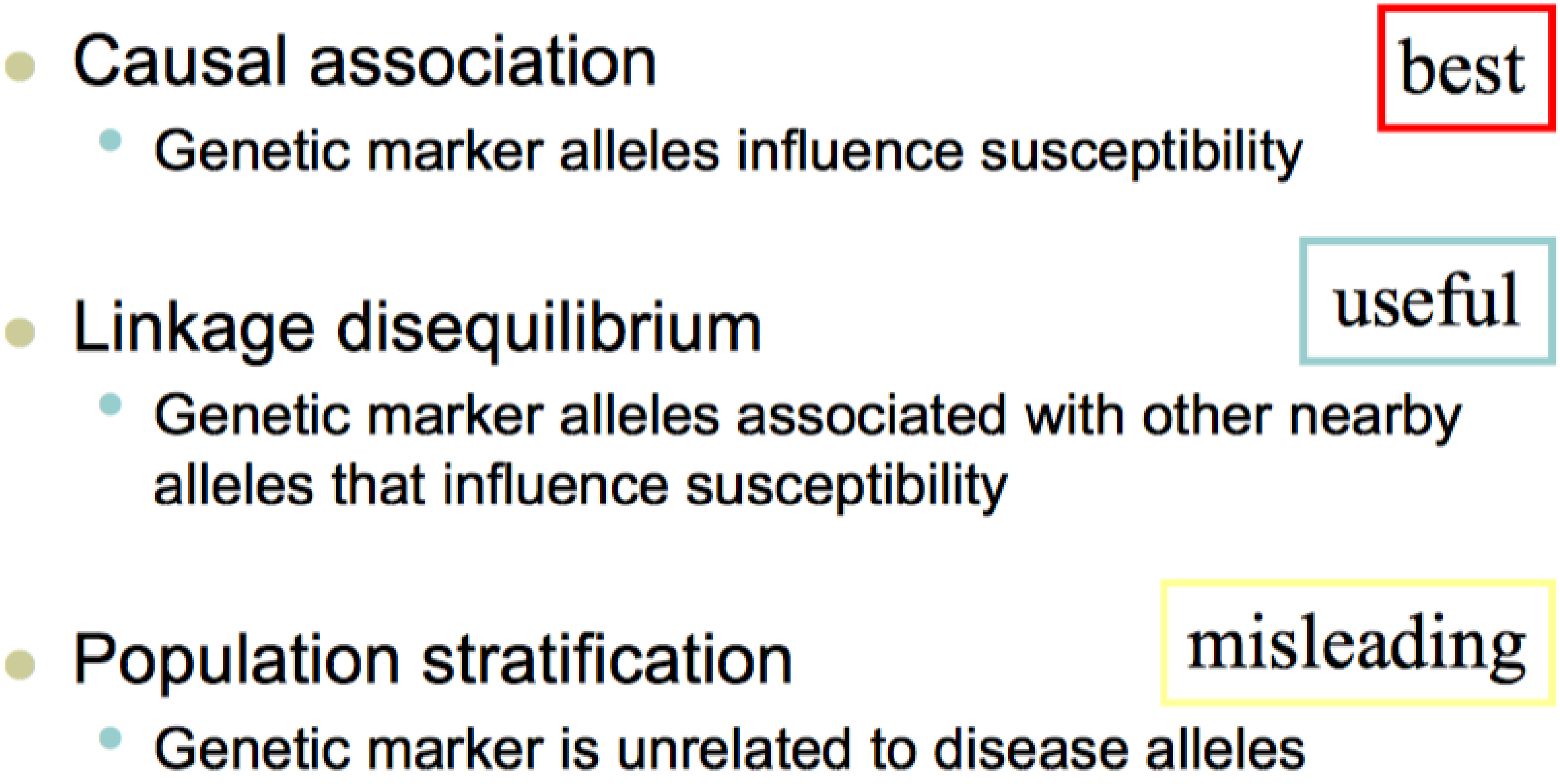
MAF Spectrum and Genetic Effect Sizes

Genetic architecture of complex traits


Family-based Sequencing Studies
Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Population genome/exome sequencing, Dense reference genotype imputation
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Array-based genotyping and genotype imputation
Common variants implicated in common disease by GWA
Rare variants of small effect very hard to identify by genetic means
Deep Whole Genome Sequencing with Large Samples?
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

GWA (or GWAS): genome-wide association studies.

Sources of Association



- Causal association
  - Genetic marker alleles influence susceptibility

best

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

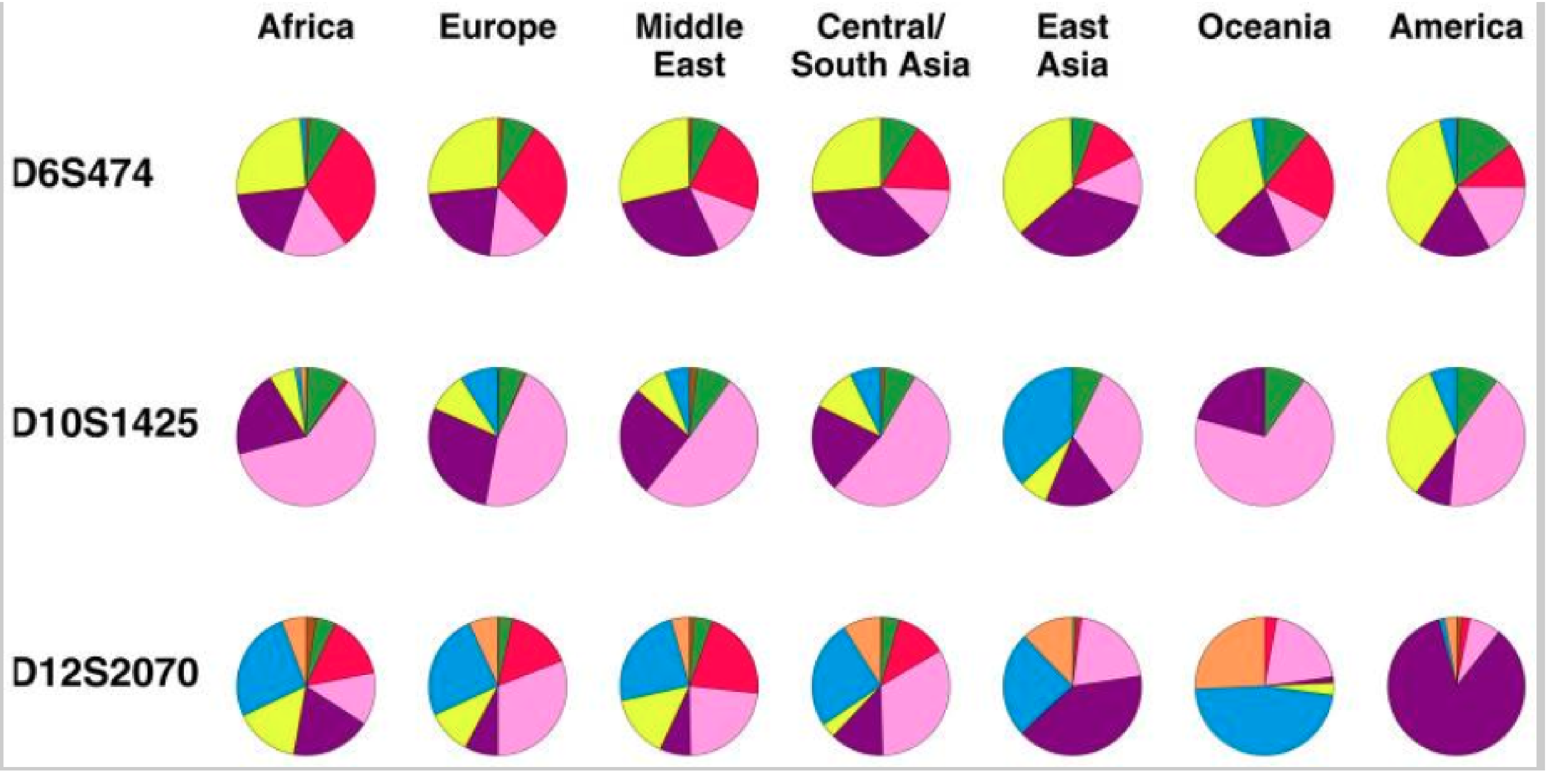
useful

- Population stratification
  - Genetic marker is unrelated to disease alleles

misleading

Population Stratification

Population stratification (or population structure) is the presence of a systematic difference in allele frequencies between subpopulations, possibly due to different ancestry.

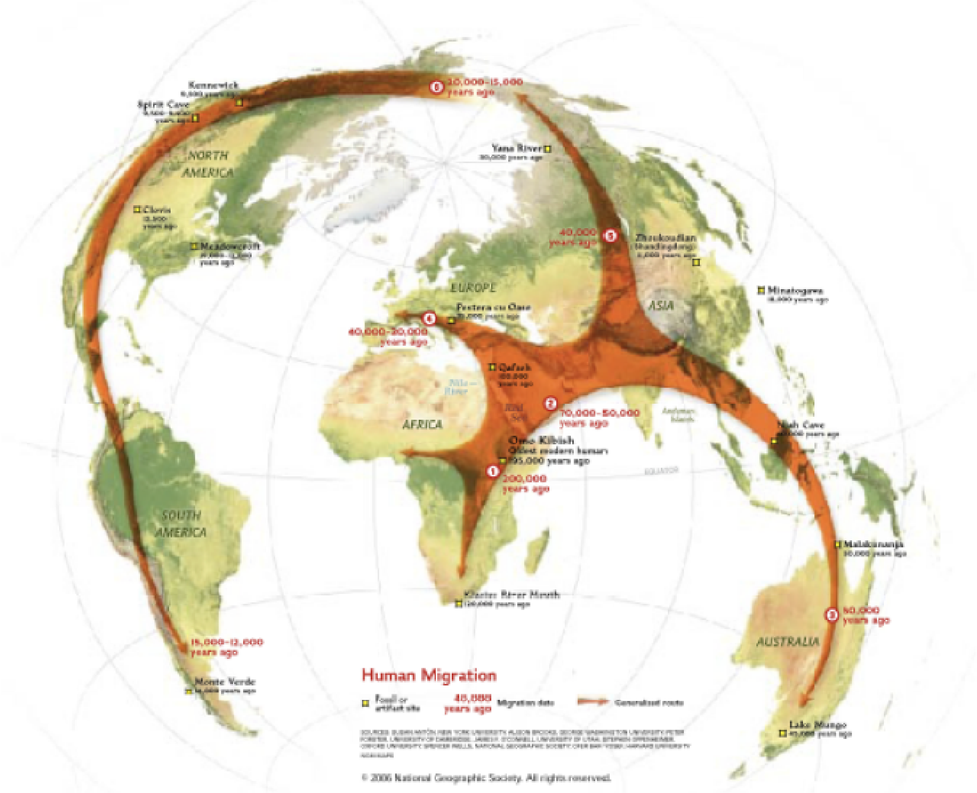


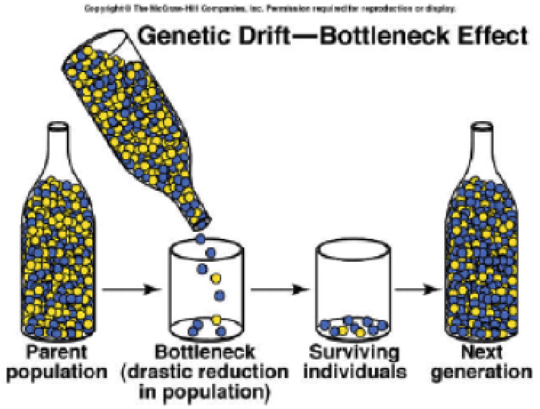
Allele frequencies at three microsatellite loci (Rosenberg N.A., Hum Biol. 2011). Each of the three loci has exactly eight alleles. In most of the pie charts, one or more alleles is rare or absent.

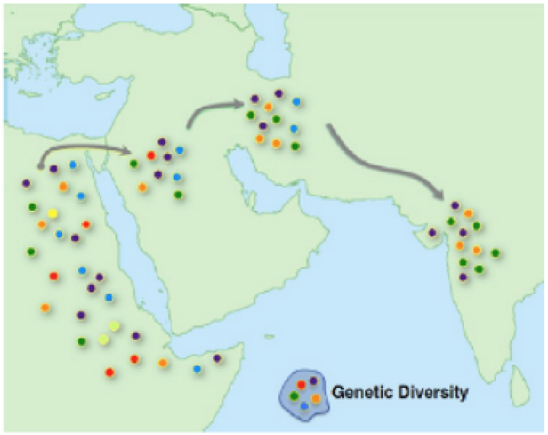
Population Stratification Causes

Basic cause of population stratification is non-random mating, often due to human migration and physical separation.

Human migration:




Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.
Genetic Drift—Bottleneck Effect
Parent population
Bottleneck (drastic reduction in population)
Surviving individuals
Next generation



National Geographic

Henn *et al.* (2012) *PNAS*

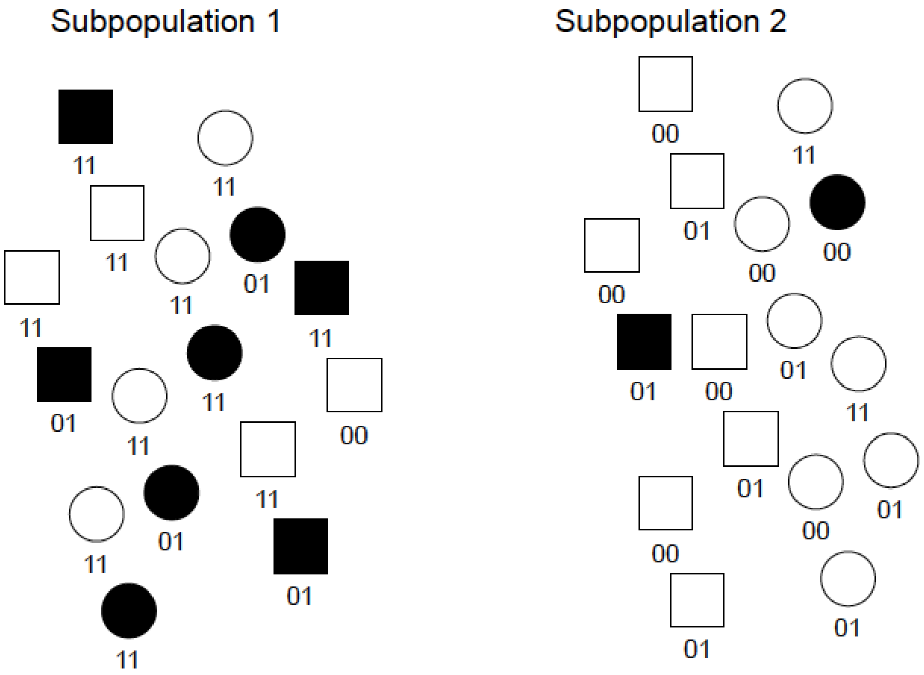
4

Population Stratification Consequences

1. Population stratification is a major confounder in genetic association studies, which can lead to false significant association that are not due to a disease locus;
2. Often lead to inflated false positive findings for studies including a mixture of different subpopulations;
3. Often seen when case-control ratio (or disease prevalence) is different across subpopulations, or when phenotypes differ among subpopulations.

Population Stratification: An Intuitive Example

Consider genotypes (coded as 00, 01 and 11) at a marker locus



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 study tends to show association, even though there is no association within each subpopulation.

Population Stratification: A Numerical Example

Consider this scenario:

- Two subpopulations: European Americans (EA), African Americans (AA)
- A SNP (B/b) has MAF: $\Pr_{\text{EA}}(b) = 0.1$, $\Pr_{\text{AA}}(b) = 0.4$
- Disease prevalence: $\Pr_{\text{EA}}(D = 1) = 0.05$, $\Pr_{\text{AA}}(D = 1) = 0.1$
- 1000 cases, 1000 controls from a population with equal prop. of EAs & AAs
- The SNP is NOT associated with the disease.

By Bayes' Theorem,

$$\begin{aligned}
\Pr(\text{AA}|\text{Case}) &= \frac{\Pr(\text{Case}|\text{AA})\Pr(\text{AA})}{\Pr(\text{Case}|\text{AA})\Pr(\text{AA}) + \Pr(\text{Case}|\text{EA})\Pr(\text{EA})} = \frac{0.1 \times 0.5}{0.075} = 0.667 \\
\Pr(\text{EA}|\text{Case}) &= 1 - 0.667 = 0.333 \\
\Pr(\text{AA}|\text{Control}) &= 0.486 \\
\Pr(\text{EA}|\text{Control}) &= 0.514
\end{aligned}$$

Association testing: any difference in number of b alleles between cases & controls

$$\begin{aligned}
2000 \times \Pr(b|\text{Case}) &= 2000 \times \{\Pr_{\text{AA}}(b)\Pr(\text{AA}|\text{Case}) + \Pr_{\text{EA}}(b)\Pr(\text{EA}|\text{Case})\} \approx 600 \\
2000 \times \Pr(b|\text{Control}) &= 2000 \times \{\Pr_{\text{AA}}(b)\Pr(\text{AA}|\text{Control}) + \Pr_{\text{EA}}(b)\Pr(\text{EA}|\text{Control})\} \approx 492
\end{aligned}$$

Association test would be statistically significant! $\Rightarrow$ False positive result

How to Address Population Stratification

Straitforward approach:

- Carefully select samples such that cases and controls are ethnically matched
- Stratify analyses by ethnicity and then combine results by meta-analysis

Potential problems:

- Self-report is not always reliable
- Considerable variability exists even within race

Widely used approach:

- Account for inflated false-positive rate (genomic control factor, our later lecture)
- Adjust for ancestry quantified by genetic markers (Principle Components Analysis, our later lecture)

Alternative approach:

- Family-based association analysis (this lecture)

Family-Based Association Methods

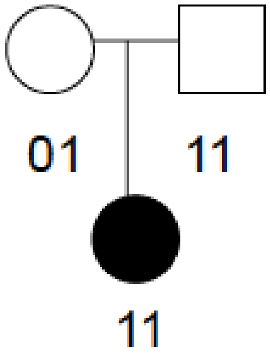
- **Intuition:** Under the null hypothesis of no association, an affected child is equally likely to inherit either allele at the tested marker locus; allele not inherited by the affected child serves as a matched control.
  - **Transmission disequilibrium test (TDT):** Father-Mother-AffectedChild trios
  - **Discordant alleles tests:** Affected-Unaffected siblings
  - **Family-based association test (FBAT):** General tests that can be used for both dichotomous and quantitative traits
  - **Quantitative TDT (QTDT):** Variance-Components based test of transmission distortion for quantitative traits

TDT

Spielman *et al.* 1993. *American Journal of Human Genetics*

- **Study design:** Consider trio families with two genotyped parents and one affected child: father + mother + one affected child
- **Rationale:** Compare the distribution of alleles transmitted to the affected offspring to the distribution of alleles not transmitted to the affected offspring
  - Under the null hypothesis, a heterozygous parent is equally likely to pass either allele to affected offspring, according to Mendel's Law
  - Under the alternative hypothesis, a heterozygous parent is more likely to pass one allele (disease allele) to affected offspring than the other allele
  - Focus on children with heterozygous parents
- **Matched design:** Transmitted alleles (cases) are matched with non-transmitted alleles (controls)
- **Advantage:** Conditioning on parents' genotypes makes the study free of population stratification

TDT scoring



Contribution of a single trio to an overall table

- Mother: transmitted allele – 1; not transmitted allele – 0.
- Father: transmitted allele – 1; not transmitted allele – 1.

		Transmitted Allele	
		1	0
Not Transmitted Allele	1	+1	
	0	+1	

TDT scoring (continued)

With multiple trios, we can generate the following contingency table:

		Transmitted Allele	
		1	0
Not Transmitted Allele	1	N_{11}	N_{10}
	0	N_{01}	N_{00}

N_{ab}: number of times allele a is not transmitted and allele b is transmitted when parents have genotype ab.

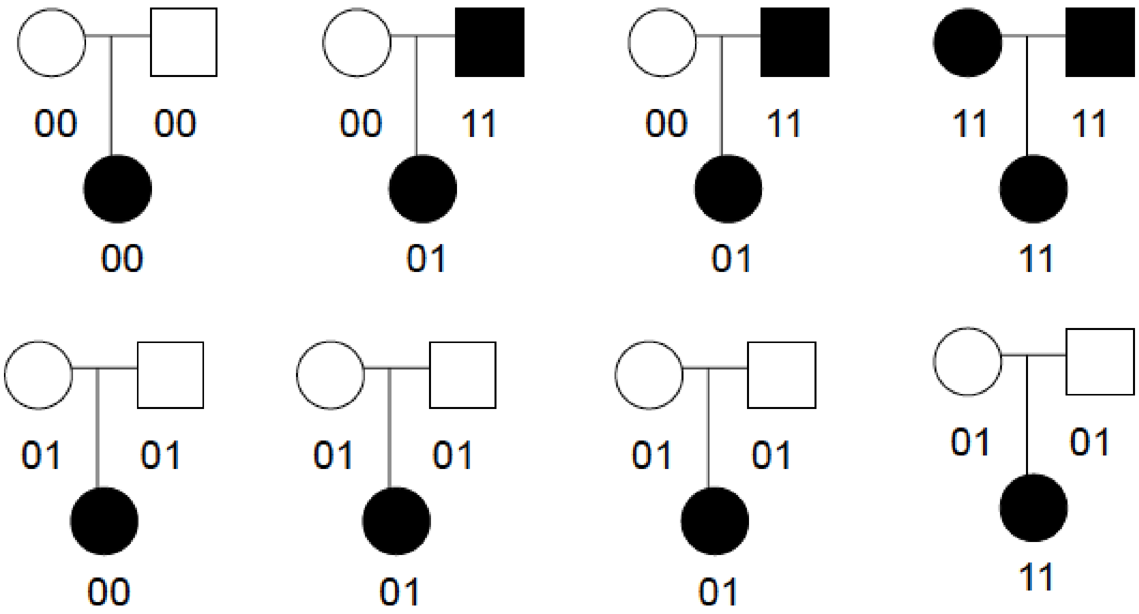
Test statistic:

$$X^2 = \frac{(N_{10} - N_{01})^2}{N_{10} + N_{01}} \sim \chi^2_1$$

Note:

- This is the McNemar's test of association for matched cases and controls
- Only heterozygous parents contribute to the test statistic
- Null hypothesis: a heterozygous parent is equally likely to transmit either allele to affected offspring

TDT Example – No Association



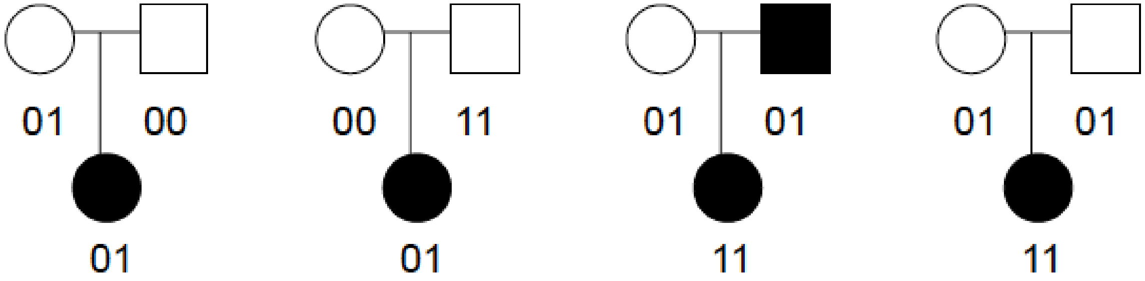
- In the parental population, affected cases are always 11, suggesting association
- TDT test:

		Transmitted Allele	
		1	0
Not Transmitted Allele	1	4	4
	0	4	4

$$X^2 = 0$$

- Affected children are not more likely to inherit allele 1 than allele 0; the parental population exhibits spurious association.

TDT Example – True Association



- TDT test:

		Transmitted Allele	
		1	0
Not Transmitted Allele	1	1	0
	0	5	2

- Association is significant because p-value = .025.

TDT Hypotheses

Three possible null-hypotheses:

- H_0 : No linkage and no association (subject to candidate-gene studies without having obtained a previous linkage signal, or population based GWAS)
- H_0 : Linkage but no association (subject to follow up studies of linkage signals)
- H_0 : Association but no linkage (subject to further test in population based case-control association studies, potentially due to population stratification)

Only one alternative hypothesis:

- H_a : The marker is both linked and associated with a disease-susceptibility locus (DSL)

Features of TDT

- Robust to population stratification for conditioning on parental genotype
- Robust to potential misspecification of the disease models and phenotype distributions for conditioning on the trait
- The validity of the test statistic relies only on Mendel’s law of random transmission of each parental allele with equal probability to each offspring (under null hypothesis)
- Sensitive to genotyping errors (Mitchell and Cutler, 2003)

FBAT as Extension of TDT

- Family-based association test (FBAT) is an extension of TDT that allows for
  - missing data in parents (due to late-onset disease)
  - quantitative traits
  - covariate adjustment
  - general pedigrees (not limited to trios)
- The test statistic uses a natural measure of association between two variables, a covariance between the trait T_{ij} and the genotype X_{ij} variables. Consider covariance

$$U = \sum_{i,j} T_{ij}(X_{ij} - E[X_{ij}|S_i]),$$

  - i indexes family and j indexes non-founders (offsprings) in the family
  - T_{ij} is typically centered
  - $E[X_{ij}|S_i]$ is the expected genotype value conditioning on the sufficient statistic for parental genotype and under Mendel's law.

The key is the definition of the coded traits and the coded genotypes and how the distribution is computed under the null (no association).

- X_{ij} denotes coded genotype, e.g., the number of minor alleles (additive disease model)
  - $(X_{ij} - E[X_{ij}|S_i]) = 0$ if both parents are homozygous
  - With one homozygous parent, $(X_{ij} - E[X_{ij}|S_i]) = 1/2$ if the minor allele is transmitted, or $(X_{ij} - E[X_{ij}|S_i]) = -1/2$ if the minor allele is not transmitted
  - With two heterozygous parents, $(X_{ij} - E[X_{ij}|S_i]) = 1, 0, -1$, respectively for the number of transmitted minor alleles $2, 1, 0$.
- In the special case where $T_{ij} = 1$ for all i, j, U simply counts the total number of transmitted minor alleles from heterozygous parents, minus their expected number
- X_{ij} can be encoded to reflect the model of inheritance, e.g., $X_{ij} = 1$ with genotype aa and $X_{ij} = 0$ otherwise for recessive disease model

Coding the trait:

- Dichotomous traits: Cases are encoded as $T_{ij} = 1 - \mu$, controls are encoded as $T_{ij} = -\mu$,
  - μ is often taken as the disease prevalence (also referred as offset)
  - $\mu < 0.5$ can also be taken to minimize $Var(U)$
  - $\mu = 0$ will only take affected individuals into account
- Quantitative traits: $T_{ij} = Y_{ij} - \mu$ with sample mean μ

FBAT

- Test statistic

$$\chi^2_{FBAT} = \frac{U^2}{Var(U)}$$

- Under the NULL hypothesis H_0 of no association and no linkage:

$$Var(U) = \sum_{i,j} T^2_{ij} Var(X_{ij}|S_i),$$

where $Var(X_{ij}|S_i)$ is computed conditional on the sufficient statistics of parental genotypes.

- Under the NULL hypothesis with large sample size:

$$\chi^2_{FBAT} \sim \chi^2_1$$

FBAT vs. TDT

- The FBAT statistic is exactly the same as the TDT statistic under the following conditions:
  - both parents are genotyped, bi-allelic marker
  - $T_i = 1$ for affected offspring and $T_i = 0$ for others
  - X_i counts the number of a specific allele (additive disease model)
  - NULL hypothesis of no association and no linkage
- $U = N_{10} - N/2$, where N_{10} denotes the number of heterozygous transmissions of $10 \rightarrow 0$ to affected offsprings, and N denotes the number of heterozygous parent-child pairs
- Multiple offsprings are independent because of no linkage under NULL, and $Var(U) = \sum_{i,j} Var(X_{i,j}|S_i) = N(1/2)^2$, because each transmission has variance equal to $(1/2)^2$
- Thus,

$$\chi^2_{FBAT} = \frac{(N_{10} - N/2)^2}{N(1/2)^2} = \frac{(N_{10} - N_{01})^2}{N},$$

where N_{01} is the number of heterozygous transmissions of $10 \rightarrow 1$ to affected offsprings, and $N = N_{10} + N_{01}$.

Transmission-Based vs. Non Transmission-Based

Transmission-based association tests

- TDT, FBAT
- Control for population stratification
- Lose power because they have to condition on parental genotypes

Non transmission-based association tests

- QTDT (the variance component methods as discussed for linkage analysis of quantitative traits)
- Account for family relatedness in the variance (e.g., genetic component in the variance induces family relatedness)
- More powerful because of not conditioning on parental genotypes
- Not necessarily robust for population stratification

Challenges For Family-Based Association Studies

Family-based designs

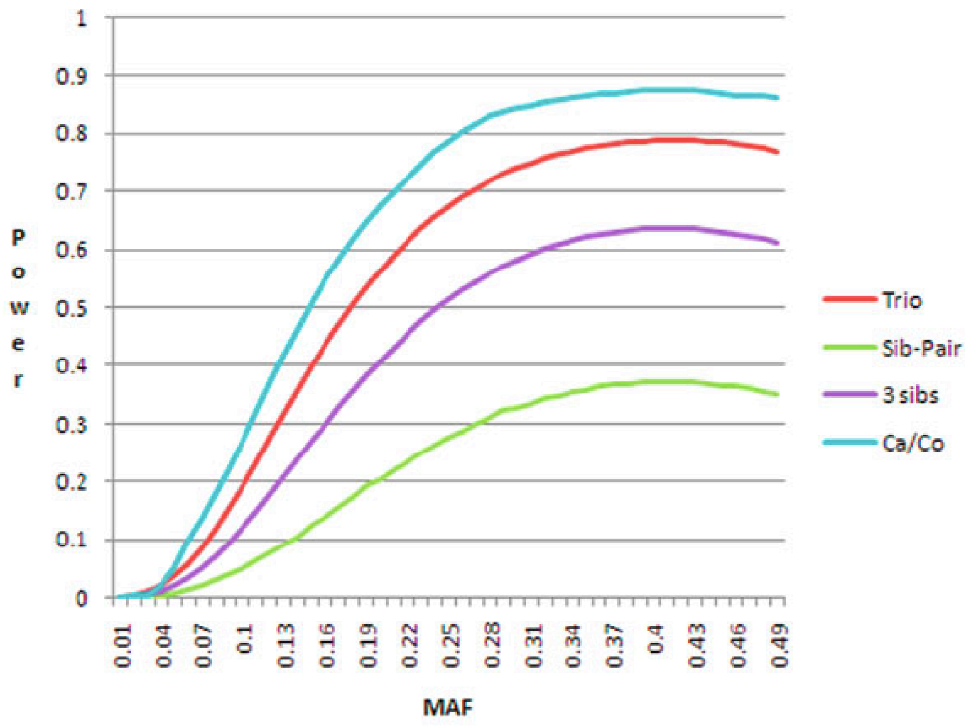
- Popular because of convenience; at the time most genetic studies were family-based linkage studies
- Difficult to ascertain parents/discordant siblings
- (particularly discordant sibling design) less powerful than case-control design
- Less used nowadays.

Population-based association studies

- Mostly used nowadays, given the development of statistical methods to control for population stratification and next-generation sequencing technology

Power for Family-Based & Population-Based Designs

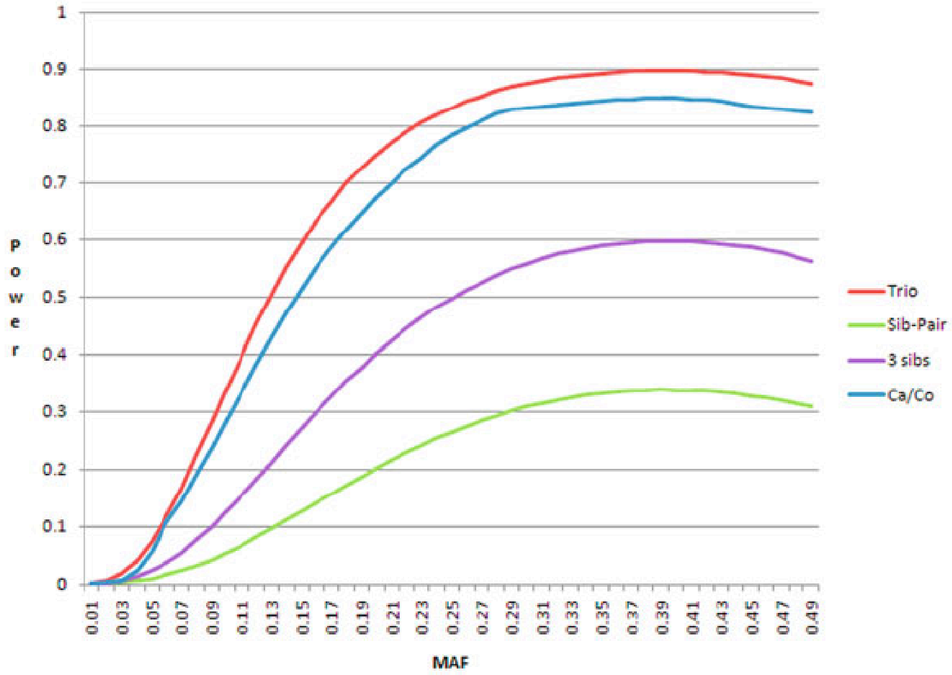
Common disease (prevalence of 10%)



- Sib-Pair: discordant sib pairs
- 3 sibs: discordant sib trios (one discordant sib pair and one additional sibling)
- Power is estimated for 1500 families or 1500 cases and 1500 controls under an additive mode of inheritance and an odds-ratio of 1.4.

Power for Family-Based & Population-Based Designs

Rare disease (prevalence of 0.1%)



- Power is estimated for 1500 families or 1500 cases and 1500 controls under an additive mode of inheritance and an odds-ratio of 1.4.
- For a common disease, case-control design achieves the highest power
- For a rare disease, the trio design is slightly more powerful than case-control

Study Design

Factors to be considered:

- Power
- Robustness to confounding factors such as population stratification
- Genotyping cost
  - Trio design requires genotyping of 1500×3 subjects
  - Case-control requires genotyping of 1500×2 subjects

Homework

Install PLINK: https://www.cog-genomics.org/plink/2.0/

- Download example dataset “89 HapMap samples and 80K random SNPs” from http://zzz.bwh.harvard.edu/plink/tutorial.shtml
- Test PLINK command line usage: Windows can use Cygwin (https://www.cygwin.com/) to get a linux type terminal; MAC uses Terminal
- Read through analysis tutorials with this example dataset
- Midterm project will be about using PLINK to do association analysis

References

TDT

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 52:506–516.

FBAT

Laird NM, Horvath S, and Xu X. 2000. Implementing a unified approach to family-based tests of association. *Genetic Epidemiology* 19: S36–S42.

QTDT

Abecasis GR, Cardon LR, and Cookson WOC. 2000. A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics* 66: 279–292.

Abecasis GR, Cookson WOC, and Cardon LR. 2001. Pedigree tests of transmission disequilibrium. *European Journal of Human Genetics* 8: 545–551.