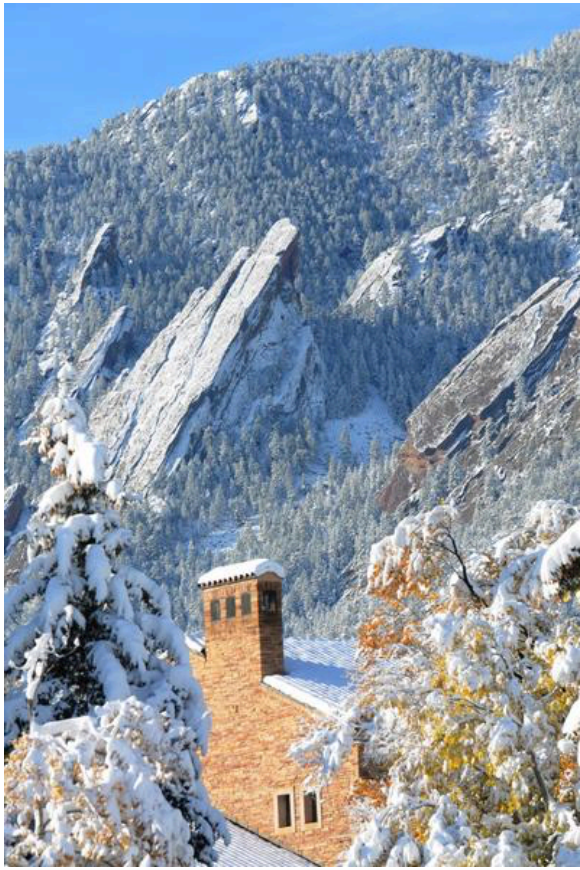
University of Colorado Boulder

Lecture 5: Family-Based Analysis

Matt McQueen | Associate Professor

Department of Integrative Physiology
Institute for Behavioral Genetics
Institute of Behavioral Science
University of Colorado Boulder

Department of Epidemiology (secondary)
Colorado School of Public Health
University of Colorado



Genetic Associations

- Truth
  - Causal locus (direct)
  - In LD with causal locus (indirect)
- Chance
  - If you test 100 times ~ 5 tests < 0.05
  - The association is due to chance - no causal underpinning
- Bias
  - Association is not causal
  - Population stratification

Stratification

- Essentially a confounder!

- How does it happen?

How Does it Happen?

- Two Necessary Components:
  - Subpopulation 1 has higher prevalence (mean) of disease
  - Subpopulation 1 has different allele frequency

Examine the Data

- Allele frequencies in ethnic subgroups
- Prevalence (means) in ethnic subgroups

Famous Example

Knowler et al (1988)

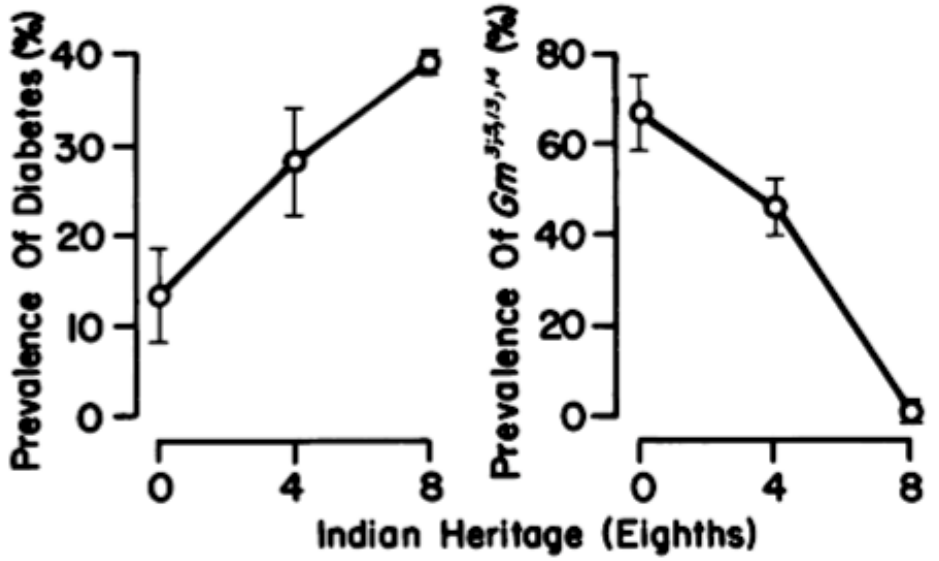
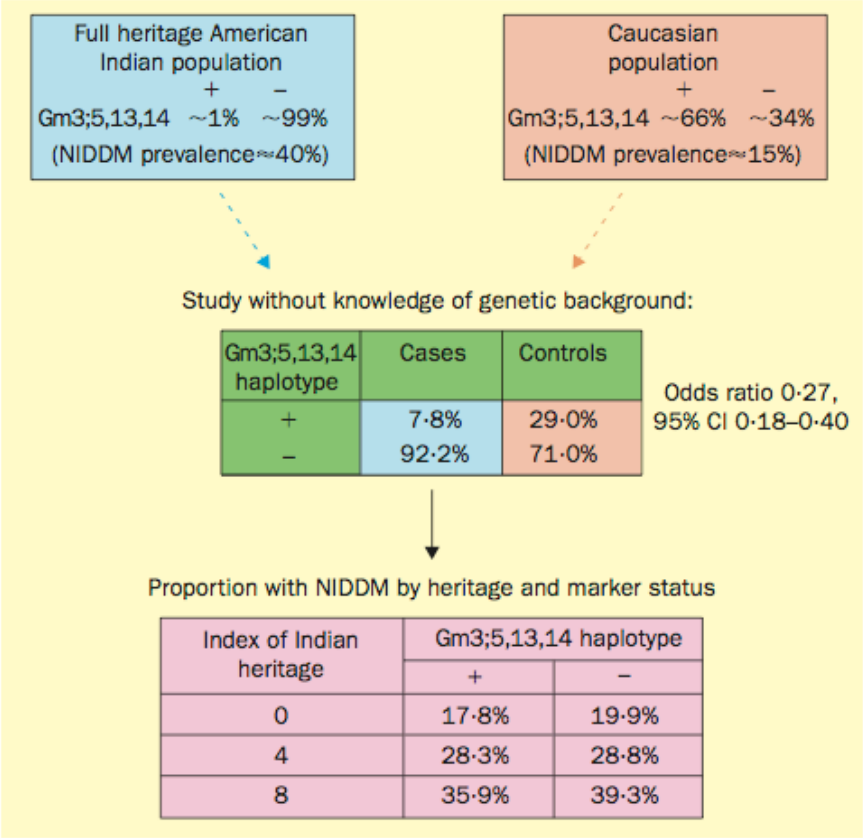
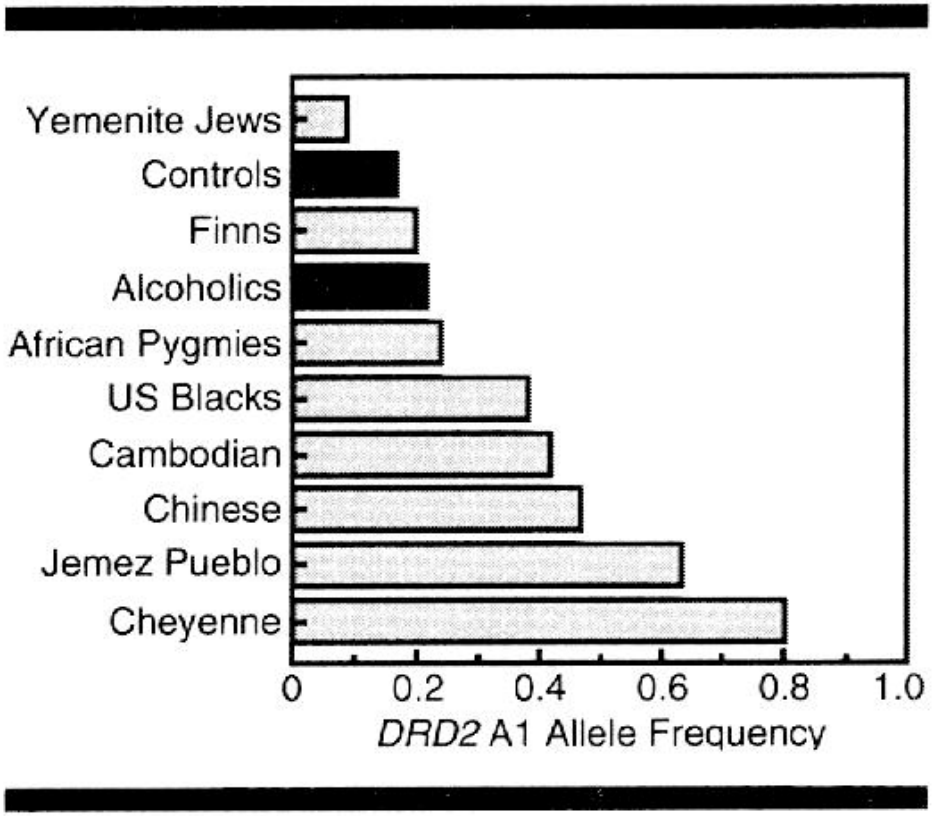


Figure 3 Age-adjusted prevalence (±1 standard error) of diabetes (left) and of $Gm^{3;5,13,14}$ (right), according to Indian heritage, among residents of the Gila River Indian Community.

Cardon et al (2003)



DRD2


Yemenite Jews
Controls
Finns
Alcoholics
African Pygmies
US Blacks
Cambodian
Chinese
Jemez Pueblo
Cheyenne
0
0.2
0.4
0.6
0.8
1.0
DRD2 A1 Allele Frequency

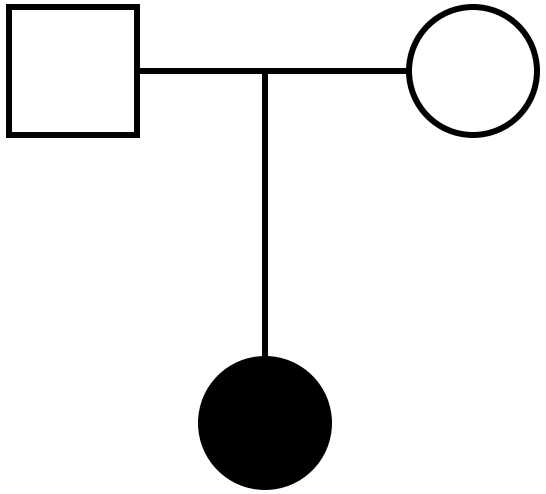
Stratification Happens

- Strategies to deal with it
  - Self-Reported Ancestry
    - Match (design) or Adjust (analysis)
  - Use other genetic markers (ancestry informative)
    - Genomic Control (Devlin – U of Pittsburgh)
    - STRUCTURE (Pritchard – U of Chicago)
    - Eigenstrat (Reich – Broad Institute/Harvard)
  - ***Use a family-based design***

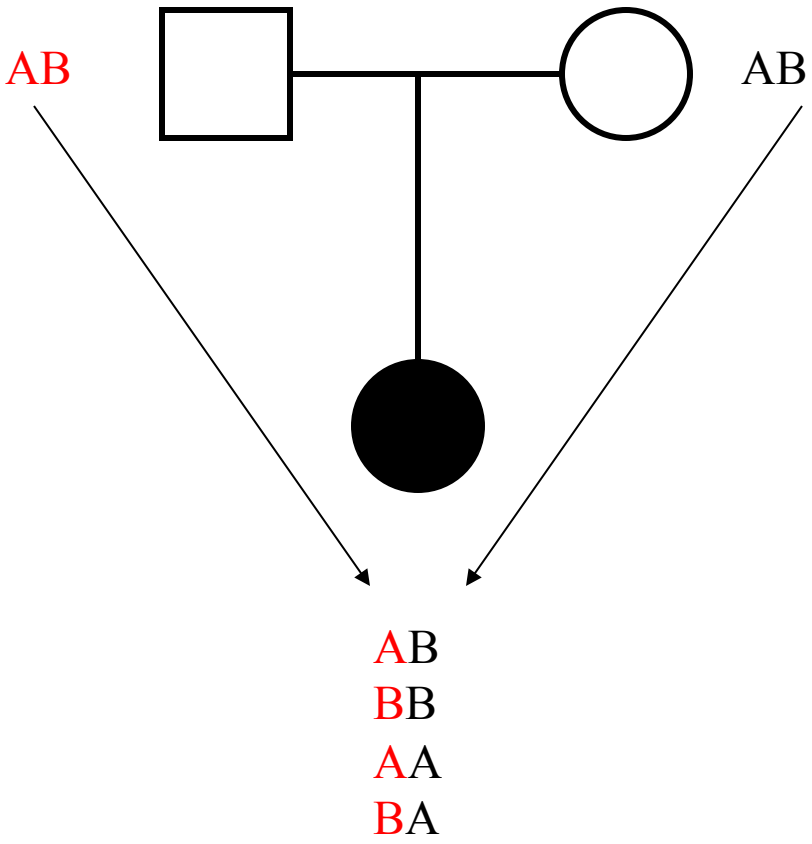
Family Based Designs

- Cases and their parents (Trios)
- Test for both linkage and association
  - rejection implies marker is close to DSL
- Robust to population substructure and failure of HWE

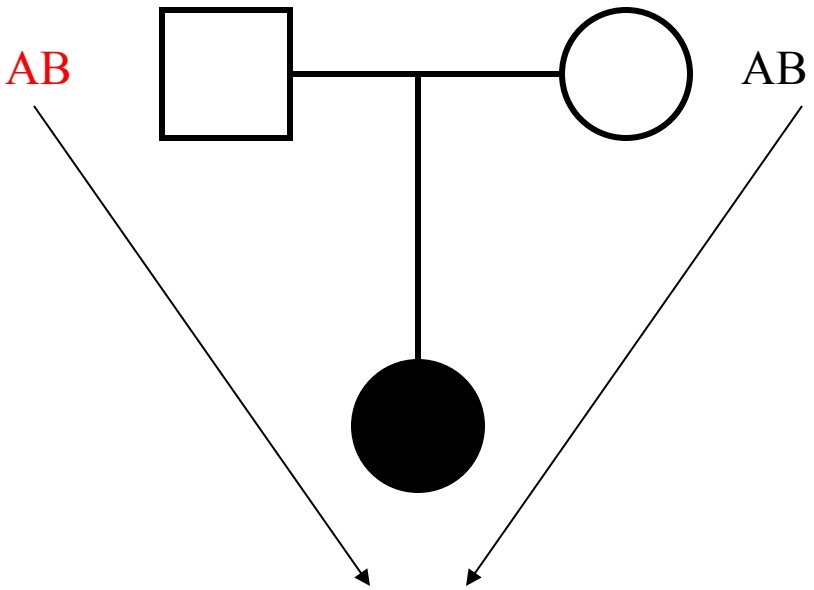
Transmission Disequilibrium Test (TDT)



Transmission Disequilibrium Test (TDT)

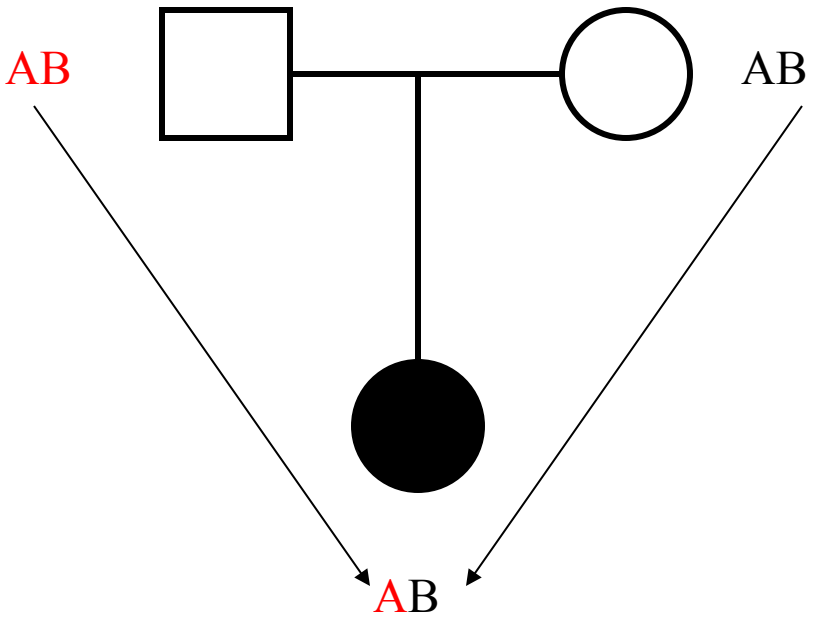


Transmission Disequilibrium Test (TDT)



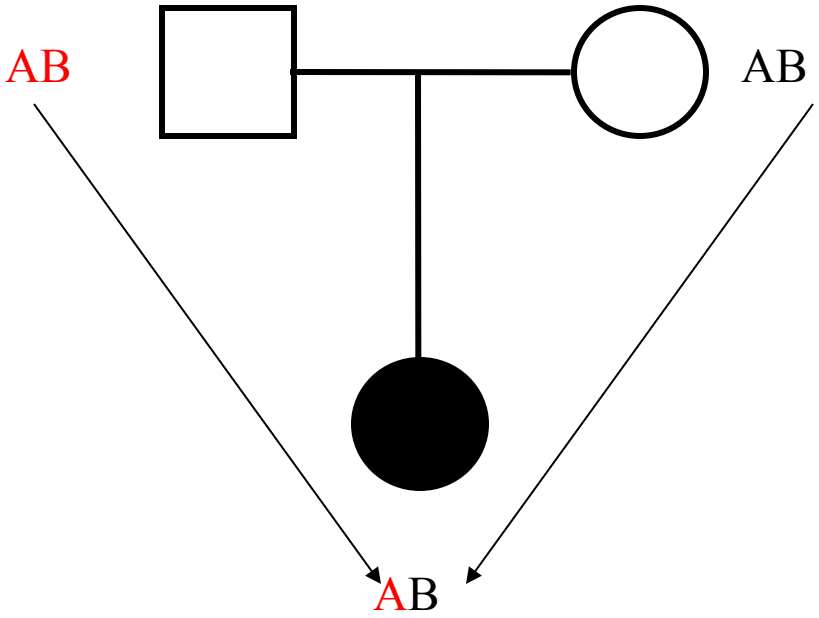
Under the null:
Equally probable!

Transmission Disequilibrium Test (TDT)



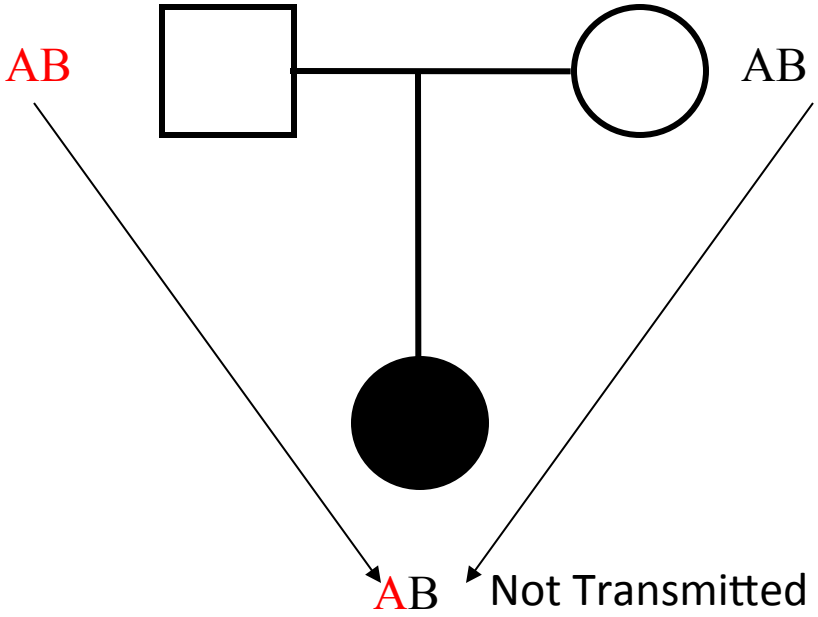
Father - “A” was transmitted and “B” wasn’t
Mother - “B” was transmitted and “A” wasn’t

Transmission Disequilibrium Test (TDT)



	Offspring		
Parent	*AA*	*AB*	*BB*
AAxAA			
AAxAB			
AAxBB			
ABxAB	*0*	*1*	*0*
ABxBB			
BBxBB			

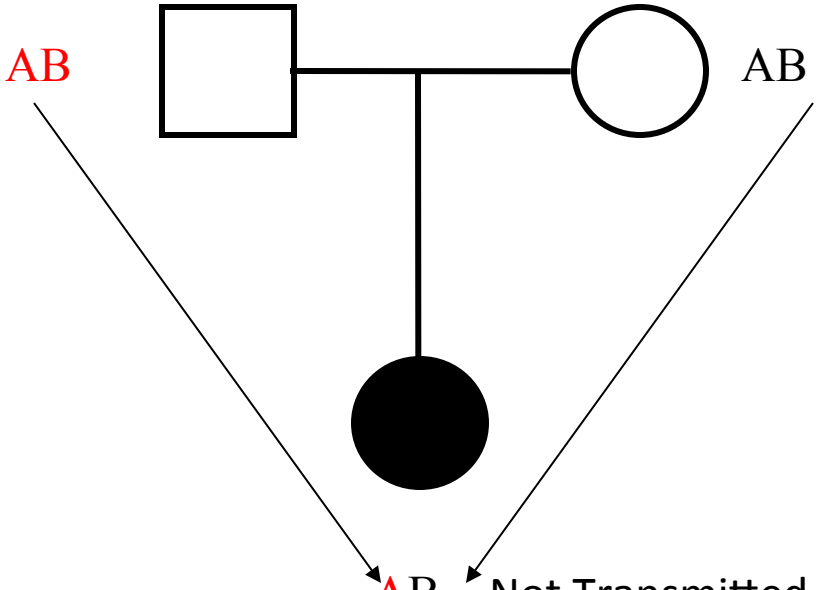
Transmission Disequilibrium Test (TDT)



	Transmitted	
Not Transmitted	A	B
A	n_{AA}	n_{AB}
B	n_{BA}	n_{BB}

	Offspring		
Parent	AA	AB	BB
AAxAA			
AAxAB			
AAxBB			
ABxAB	0	1	0
ABxBB			
BBxBB			

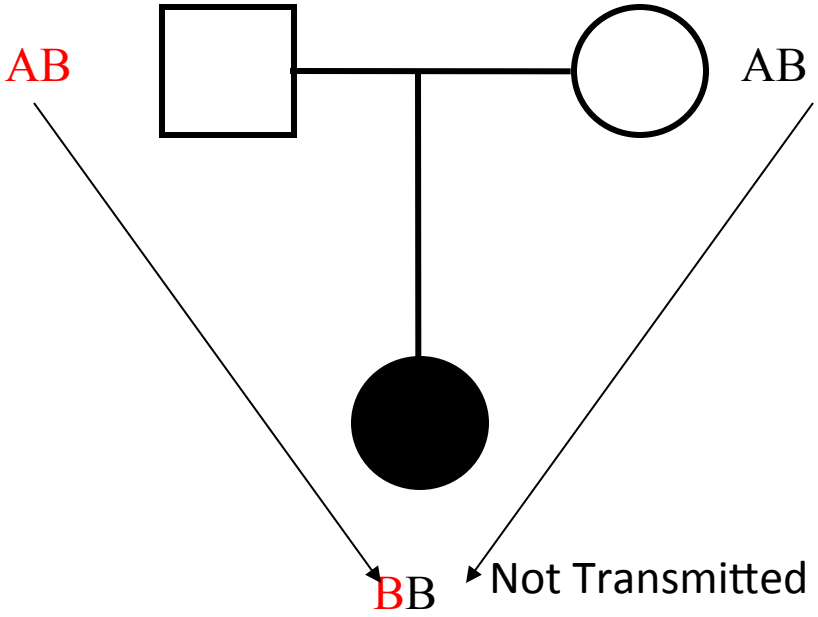
Transmission Disequilibrium Test (TDT)



	Offspring		
Parent	*AA*	*AB*	*BB*
AAxAA			
AAxAB			
AAxBB			
ABxAB	*0*	*1*	*0*
ABxBB			
BBxBB			

Transmitted	*A*	*B*
A	*0*	*1*
B	*1*	*0*

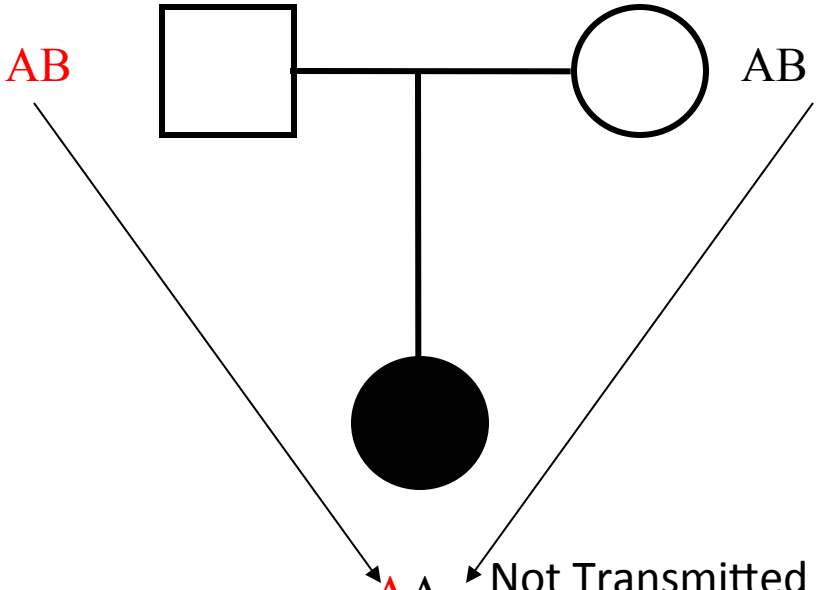
Transmission Disequilibrium Test (TDT)



	Offspring		
Parent	*AA*	*AB*	*BB*
AAxAA			
AAxAB			
AAxBB			
ABxAB	*0*	*0*	*1*
ABxBB			
BBxBB			

Transmitted \	*A*	*B*
A		
B		

Transmission Disequilibrium Test (TDT)



	Transmitted	
	A	*B*
A		
B		

	Offspring		
Parent	*AA*	*AB*	*BB*
AAxAA			
AAxAB			
AAxBB			
ABxAB	*1*	*0*	*0*
ABxBB			
BBxBB			

TDT

Transmitted \ Not Transmitted	A	B
A	n_{AA}	n_{AB}
B	n_{BA}	n_{BB}

$$TDT = \frac{(n_{BA} - n_{AB})^2}{n_{BA} + n_{AB}} \sim \chi_1^2$$

McNemar's Test for Matched-Pair Data

Attraction of TDT

H_0 relies on Mendel's laws, not on control group

H_A marker and disease locus are linked *and* the disease allele is associated with a marker allele

Intuition: If association among parents, but no linkage, alleles at disease locus are not transmitted with alleles at marker so there will be no association in offspring. If linkage, but no association, different alleles will be transmitted with disease alleles in different families.

Consequence: TDT is robust to population stratification, admixture, other forms of confounding (model free).

Limitations of TDT

- Only affected offspring
- Only dichotomous phenotypes
- Biallelic markers
- Single genetic model (additive)
- No allowance for missing parents/pedigrees
- Incorporating siblings must assume no linkage
- Does not address multiple markers or multiple phenotypes

FBAT

- standard genetic models
- other phenotypes, multiple phenotypes
- multiple alleles
- additional siblings; pedigrees
- missing parents
- multiple markers and haplotypes
- based on same conditioning principles as TDT

Key features of TDT

- Random variable in the analysis is the offspring genotype
- Parental genotypes are fixed (condition on the parental genotypes)
- Trait is fixed (condition on all offspring being affected)
- FBAT maintains these general principles

FBAT: General Approach

- Test statistic
  - works for any phenotype, any marker, genetic model
  - use covariance between offspring trait and genotype
- Test Distribution:
  - computed assuming Mendel's laws of transmission; random variable is offspring genotype
  - condition on parental genotypes when available, extend to family configurations otherwise (avoid specification of allele distribution)
  - condition on offspring phenotypes (avoid specification of trait distribution)

Test Statistic: Coding the Genotype

- Use a model to "score" genotypes.
  - Assume '1' allele of a marker is of interest.

- Dominant:

  X = 1 if genotype has a '1' allele (11 or 12)

  X = 0 otherwise

- Recessive:

  X = 1 if genotype = (11)

  X = 0 otherwise

- Additive:

  X = number of '1' alleles in the genotype, 0, 1, 2

Test Statistic: Coding the Trait

- *T* denotes the coded trait of an offspring.
- Define $T = Y - \mu$
  - where *Y* is a phenotype of interest
- *Y* disease status (1,0)
  - *Y* measured phenotype (BMI)
- μ is an 'offset' in FBAT terminology
- Choice of μ depends strongly on study design and will influence the power of the test

The FBAT Test Statistic

$$U = \sum T\left(X - E\left(X|P\right)\right)$$

$$U = \sum \left(Y - \mu\right)\left(X - E\left(X|P\right)\right)$$

Not a typical Covariance:

Centering X with E(X|P) provides robustness to population substructure;

Centering Y by μ (offset) accounts for ascertainment of offspring

Note: $U = S - E(S)$, where $S = \sum T \cdot X$ (used in FBAT output)

Informative Families

- Recall
  - $FBAT = \sum T(X-E(X|P))/ \sqrt{\sum T^2 var(X|P)}$
- Informative Family: $var(X|P) \neq 0$
- Two homozygous parents contribute nothing to *U* or variance
- Double heterozygous parents contribute *-1,0,1* to *U* and variance of *1/2*
- Single heterozygous parent contributes *+1/2* or *-1/2* and variance of *1/4*

Why use affecteds only?

	Marker Allele		
disease	1	2	
+	70	30	100
-	499,930	499,970	999,900
	500,000	500,000	10^6

With rare disease, population differential is in diseased cases; using affected only most informative. With common disease, there will be differential in the unaffected.

Including Unaffected

$T = (Y - \mu)$ μ is an offset between 0 and 1

$T = (1 - \mu)$ if $Y = 1$ (affected)

$T = -\mu$ if $Y = 0$ (unaffected)

$T = 0$ if disease status unknown

Note: Use a contrast because if *X* is over-transmitted to affected, it is under-transmitted to unaffected

- μ is a 'weight'
  μ = 0 TDT (affected only)
  μ = 1 (uses unaffected only)
  μ = ½ count equally

- Optimal choice μ = disease prevalence for a population sample.

- Power of FBAT tests depends on sample design and on offset μ .

Choosing the Offset

- Poor choice can lead to very low power
- Choice depends on ascertainment of offspring
- Sample mean works well if no ascertainment on trait of interest; may not be able to estimate *E(Y)* from a selected sample
- With highly selected traits, dichotomizing may be preferable

Quantitative traits vs dichotomous trait: Rules of thumb

- Use dichotomous trait if have ascertained sample
- Analyzing such a trait as a continuous variable can be difficult because results are highly sensitive to choice of offset;
  - Sample mean is usually a bad choice.
- Use continuous trait in absence of ascertainment on trait
  - Sample mean good choice for offset.
- Transforming the trait to a dichotomous trait will reduce the power.

Extensions to FBAT

- Missing Parents
- Siblings, nuclear families
- Multivariate (GEE)
- Time to onset (Cox-PH)
- See Tutorial 5 for more information