Introduction to Association

Benjamin Neale

22nd International Workshop on Twin Methodology

2009

Liberally sampled from talks by
Lon Cardon and Shaun Purcell

Outline

1. Definition of terms
2. Population-based association
3. Stratification
4. Family-based association
5. Direct vs. Indirect association

Outline

1. Definition of terms
2. Population-based association
3. Stratification
4. Family-based association
5. Direct vs. Indirect association

Association Studies

Simplest design possible

Correlate phenotype with genotype

Candidate genes for specific diseases

common practice in medicine/genetics

Pharmacogenetics

genotyping clinically relevant samples (toxicity vs efficacy)

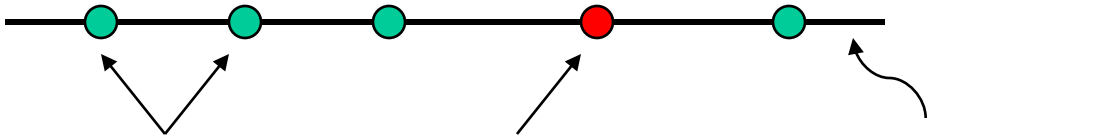
Positional cloning

recent popular design for human complex traits

Genome-wide association

with millions available SNPs, can search whole genome exhaustively

Definitions

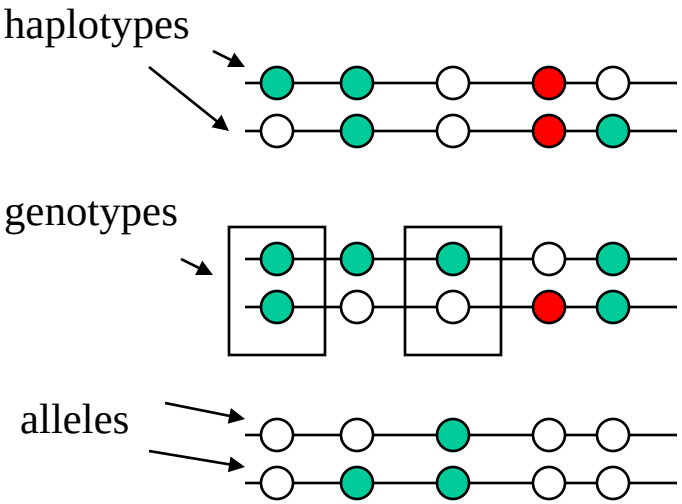


SNPs/ markers

trait variant

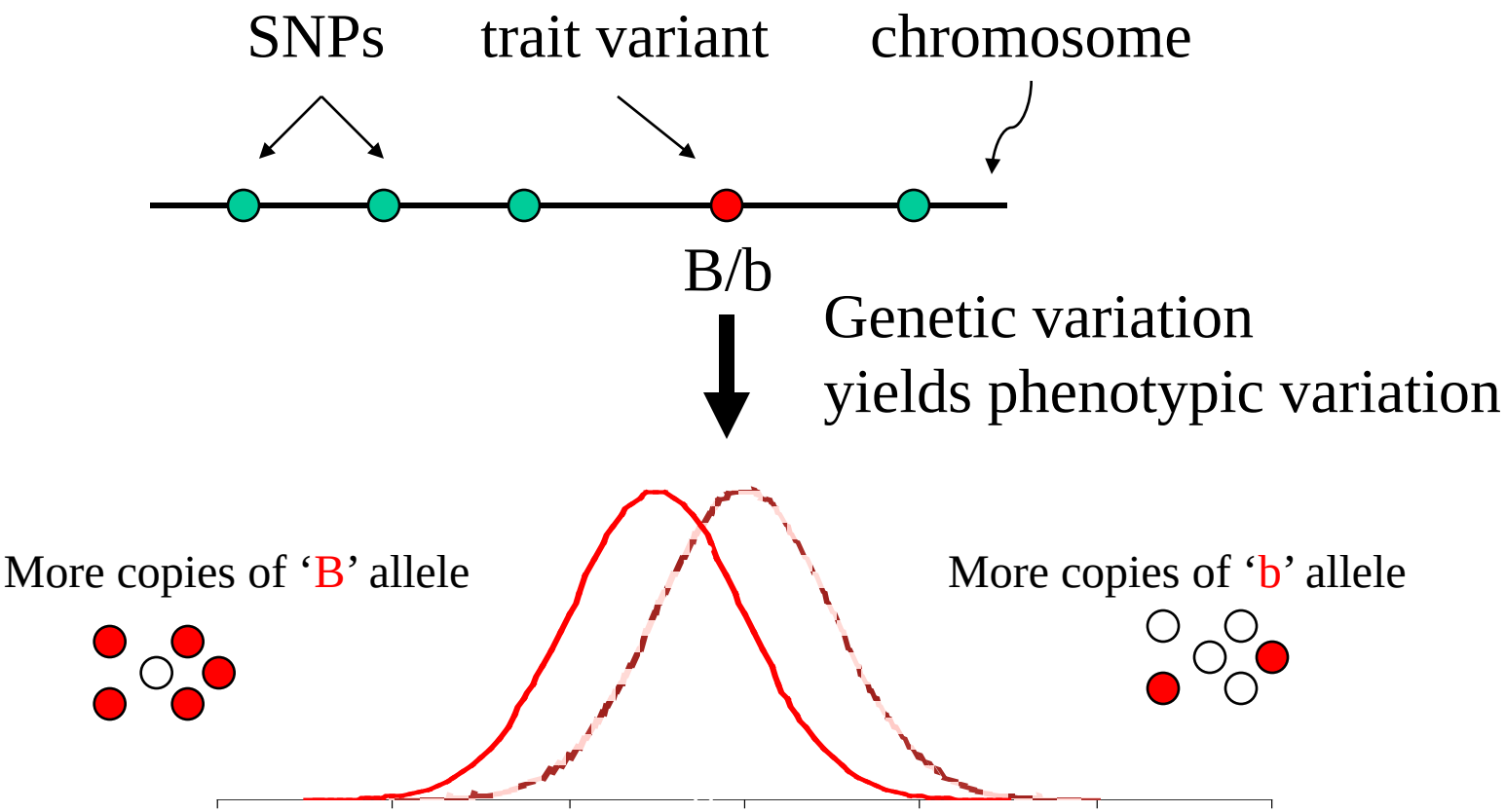
chromosome

Population Data



Correlate any of these with phenotype (continuous trait or affection status)

Allelic Association



Outline

1. Definition of terms
2. Population-based association
3. Stratification
4. Family-based association
5. Direct vs. Indirect association

Why Do Association?

- Simpler test than linkage
  - Test of mean/frequency differences

Why Do Association?

- Simpler test than linkage
  - Test of mean/frequency differences
- Considerably more powerful
  - Capable of detecting effect sizes an order of magnitude smaller

Why Do Association?

- Simpler test than linkage
  - Test of mean/frequency differences
- Considerably more powerful than linkage
  - Capable of detecting effect sizes an order of magnitude smaller
- **Sexier**

  - When was the last time you saw a linkage study in Nature Genetics?

Simplest Regression Model of Association

$$Y_i = \alpha + \beta X_i + e_i$$

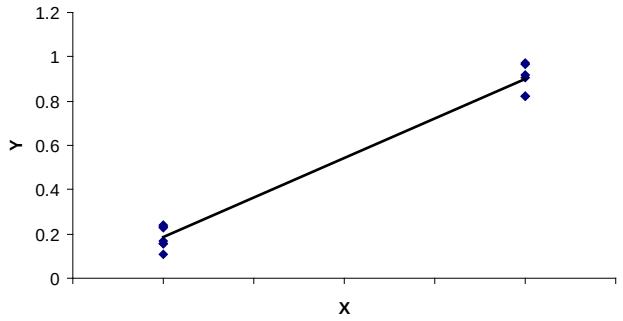
where

Y_i = trait value for individual i

X_i = 1 if individual i has allele ‘A’

0 otherwise

i.e., test of mean differences between ‘A’ and ‘not-A’ individuals



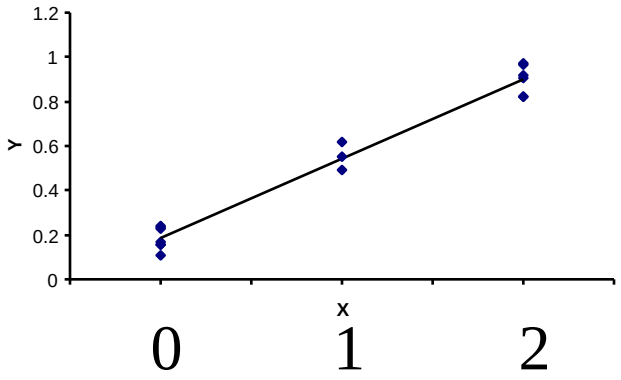
More Sophisticated Test of Association

$$Y_i = \alpha + \beta X_i + e_i$$

where

Y_i = trait value for individual i

X_i = 2 if individual i has genotype ‘AA’
1 if individual i has genotype ‘Aa’
0 if individual i has genotype ‘aa'



Yet More Sophisticated Test of Association

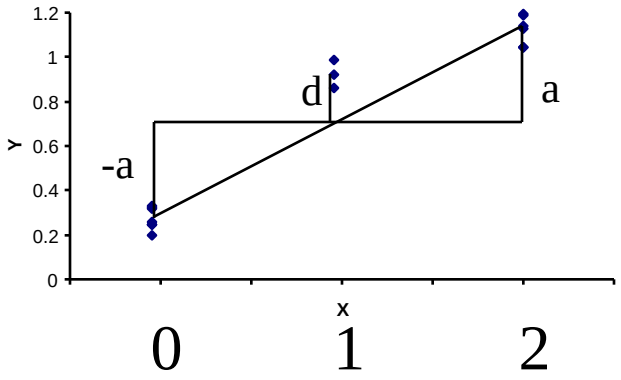
$$Y_i = \alpha + \beta_x X_i + \beta_z Zi + e_i$$

where

Y_i = trait value for individual i

X_i = 1 if individual i has genotype ‘AA’
0 if individual i has genotype ‘Aa’
-1 if individual i has genotype ‘aa‘

Z_i = 0 for ‘AA’
1 for ‘Aa’
0 for ‘aa’



All Parametizations for Association Tests

Model	Additive Only		Additive and Dominance		Major Recessive	
Parameter	1	2	1	2	1	2
AA	-1	0	-1	0	1	0
Aa	0	0	0	1	0	0
aa	1	0	1	0	0	0
Model	Minor Recessive		Heterozygote			
Parameter	1	2	1	2		
AA	0	0	0	0		
Aa	0	0	1	0		
aa	1	0	0	0		

Further extensions

- Multi-allelic (*k*) markers/haplotypes
  - Relative to reference allele
    - *k*-1 dummy variables for the additive effects of *k*-1 alleles
    - *k*-1 dummy variables for the dominance effects of k-1 alleles
- Multi-locus association (*j)* markers/haplotypes with (*k*) total alleles
  - k_j-1dummy variables for each allele for all markers
  - Warning, colinearity is a potential difficulty
  - Grouping by locus may shed further insight

Phenotype Dictates Test

- Disease Outcome/Binary Phenotype
  - Logistic Regression
  - χ^2 on contingency table
    - 2x2 or 3x2
  - Fisher's exact test
  - Armitage Trend Test
- Quantitative Outcome
  - Linear Regression
  - Student's T Test

Outline

1. Definition of terms
2. Population-based association
3. Stratification
4. Family-based association
5. Direct vs. Indirect association

No Association

- I would like to test genetic variation for chopstick use
- I will collect a sample
- Genotype said sample
- Perform basic association using χ^2

My Samples

	Sample 1 Americans	
	$\chi^2=0$	$p=1$
	Use of Chopsticks	
A	Yes	No	Total
A_1	320	320	640
A_2	80	80	160
Total	400	400	800

My Samples

Sample 2 Chinese			
$\chi^2=0$		$p=1$	
Use of Chopsticks			
A	Yes	No	Total
A_1	320	20	340
A_2	320	20	340
Total	640	40	680

My Samples

Sample 3 Americans + Chinese

$\chi^2=34.2$ $p=4.9x10^{-9}$

Use of Chopsticks

A	Yes	No	Total
A_1	640	340	980
A_2	400	100	500
Total	1040	440	1480

So What Happened?

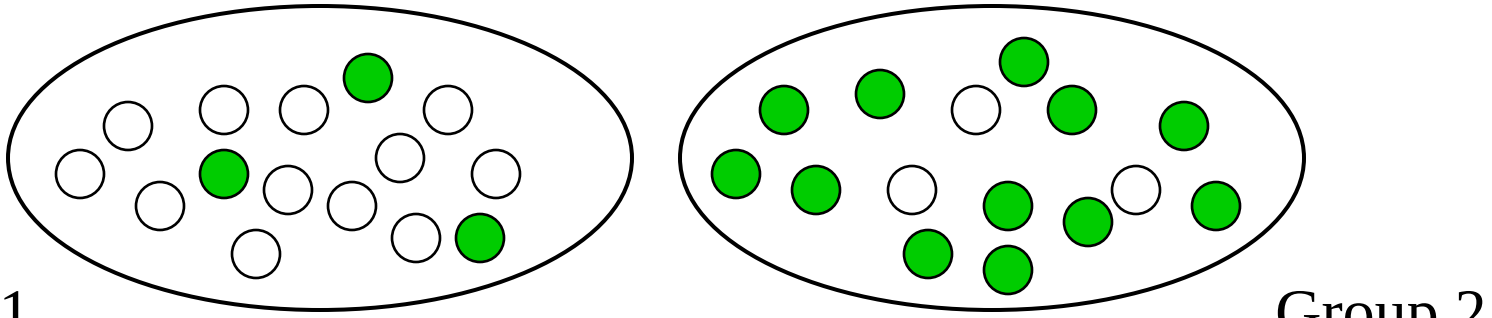
- No association in the American sample
- No association in the Chinese sample
- Combining the two samples yielded a genome-wide association significant result
- Any ideas?

Classic case of confounding or lurking variable

- Geneticists call this population stratification
- Required conditions
  - Differences in P(Disease | Population)
  - Differences in P(A_1|Population)

Visualization of stratification conditions

- Suppose that a disease is more common in one subgroup than in another…

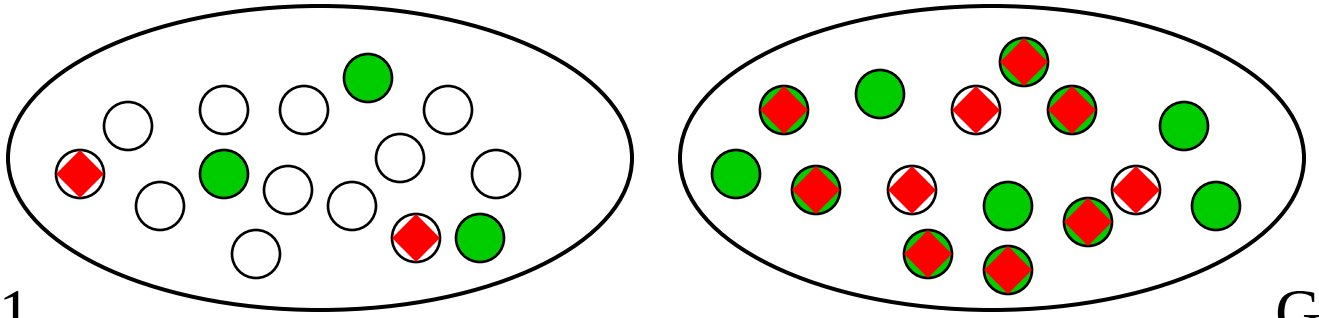


Group 1 Group 2

- …then the cases will tend to be over-sampled from that group, relative to controls.

…and this can lead to false positive associations

- Any allele that is more common in Group 2 will *appear* to be associated with the disease.



Group 1 Group 2

- This will happen if Group 1 & 2 are “hidden” – if they are known then they can be accounted for.
- Discrete groups are not required – admixture yields same problem.

Solutions

- Stratified Analysis
  - Analyze Chinese and American samples separately then meta-analyze
- Model the confounder
  - Include a term for Chinese or American ancestry in a logistic regression model
- Matching
  - Pair Chinese with Chinese and Americans with Americans
  - Family-based analysis

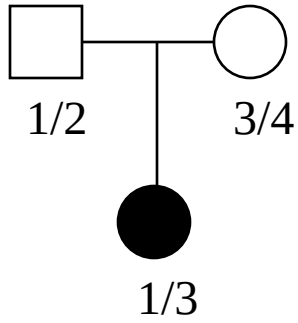
Isn’t this sexy

Outline

1. Definition of terms
2. Population-based association
3. Stratification
4. Family-based association
5. Direct vs. Indirect association

Family-based association

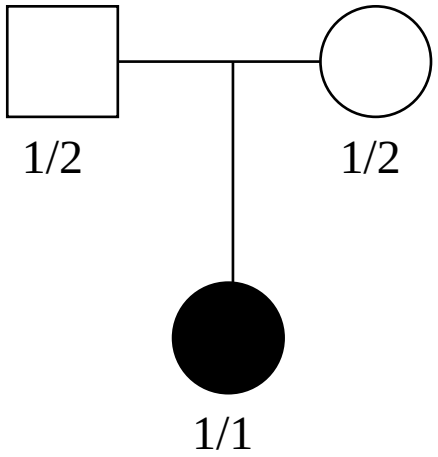
- Because of fear of stratification, complex trait genetics turned away from case/control studies
  - *fear may be unfounded*
- Moved toward family-based controls (flavor is TDT: transmission/disequilibrium test)



"Case"	= transmitted alleles
	= 1 and 3
"Control"	= untransmitted alleles
	= 2 and 4

TDT: Spielman, McGinnis & Ewens AJHG (1993)

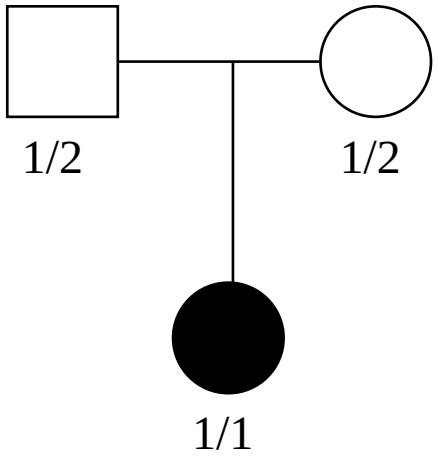
TDT more formally



According to Mendel's law of independent segregation either allele ought to be equally likely to be transmitted from parent to offspring

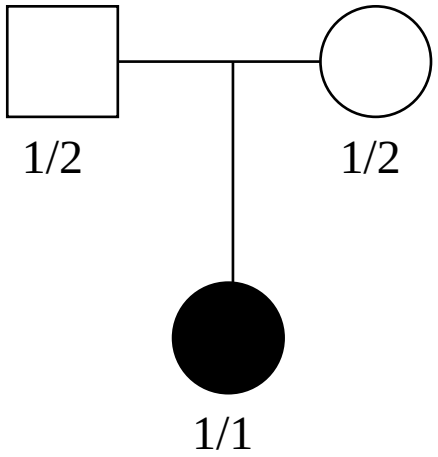
In the presence of association, P(Transmission of risk allele) increases for affected offspring

TDT Testing



The TDT is effectively a matched-pair design. We only count transmission from heterozygous parents to offspring. Thus, the non-transmitted allele is the matched pair for the transmitted allele. We can then invoke the McNemar χ^2

TDT Testing



McNemar $\chi^2 = (T_1 - NT_1)^2/(T_1 + NT_1)$

Where T_1 is the number of transmissions of allele 1 and NT is the number of non-transmissions of allele 1

Note that $T_1 = NT_2$ and $NT_1 = T_2$

TDT Advantages/Disadvantages

Advantages

Robust to stratification
Genotyping error detectable via Mendelian inconsistencies
Estimates of haplotypes improved

Disadvantages

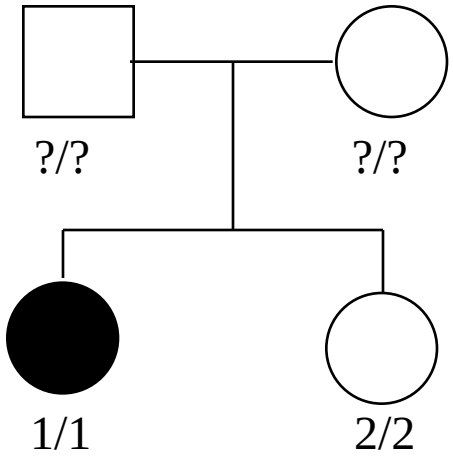
Detection/elimination of genotyping errors causes bias (Gordon et al., 2001)
Uses only heterozygous parents
Inefficient for genotyping

3 individuals yield 2 founders: 1/3 information not used

Can be difficult/impossible to collect

Late-onset disorders, psychiatric conditions, pharmacogenetic applications

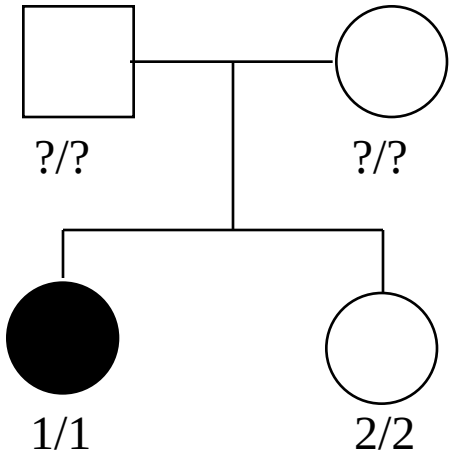
Sib-TDT



If we have siblings discordant for a disease, we can obtain association information from comparing the genotypes at the risk loci. Affected offspring should carry more risk variation than unaffected offspring.

Spielman and Ewens, AJHG 1998

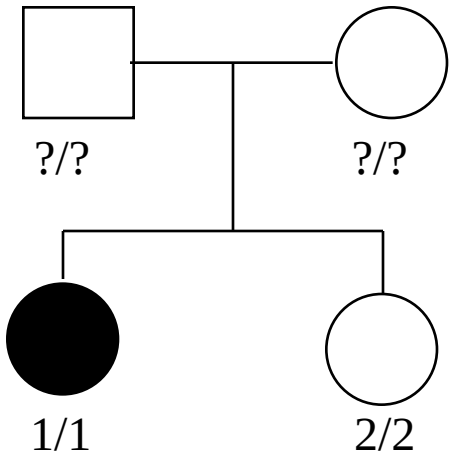
Sib-TDT



Two conditions:

1) There must be at least 1 affected and 1 unaffected offspring
2) The members of the sibship must not have the same genotype

Sib-TDT



The test:

- Calculate the allele frequency in affected offspring and in unaffected offspring
- Generate the difference in allele frequencies
- Permute affected and unaffected labels within sibship
- Generate an empirical distribution of allele frequency differences

Other Disease Tests

- HHRR (Terwilliger and Ott Hum Hered 1992)
  - Generates ‘pseudocontrols’
- PDT (Martin et al. AJHG 2000)
  - Incorporates TDT and sib-TDT in a unified framework
- Transmit (Clayton, AJHG, 1999)
- FBAT (Laird NM, Horvath S and Xu X Gen Epi 2000)
- DFAM implemented in PLINK

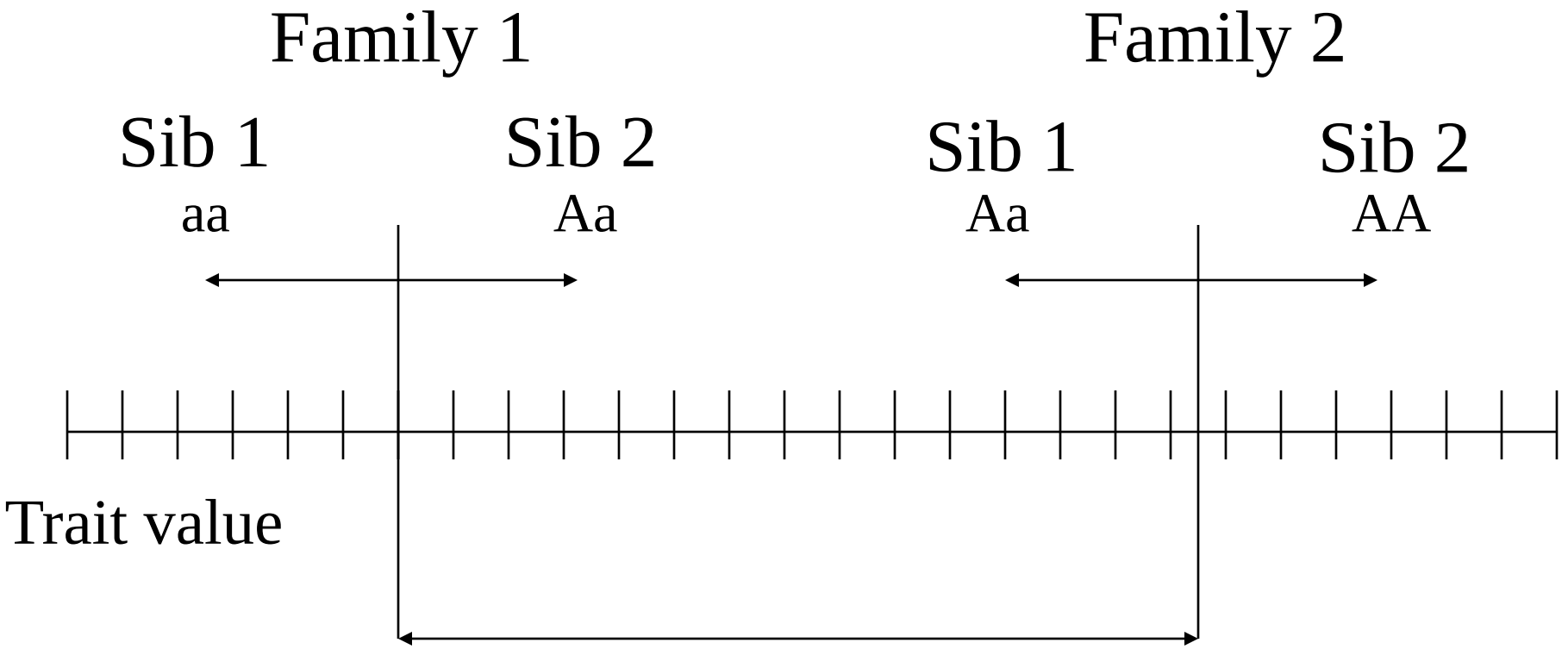
But What about Quantitative Measures?

- Family-based association of quantitative measures
- Initial methods by Allison (1997) and Rabinowitz (1997)
- Rabinowitz partitioned the association evidence in families into between and within components
- Unification of linkage and association information in Fulker et al. 1999
- Generalization to any pedigree information in Abecasis et al. 2000

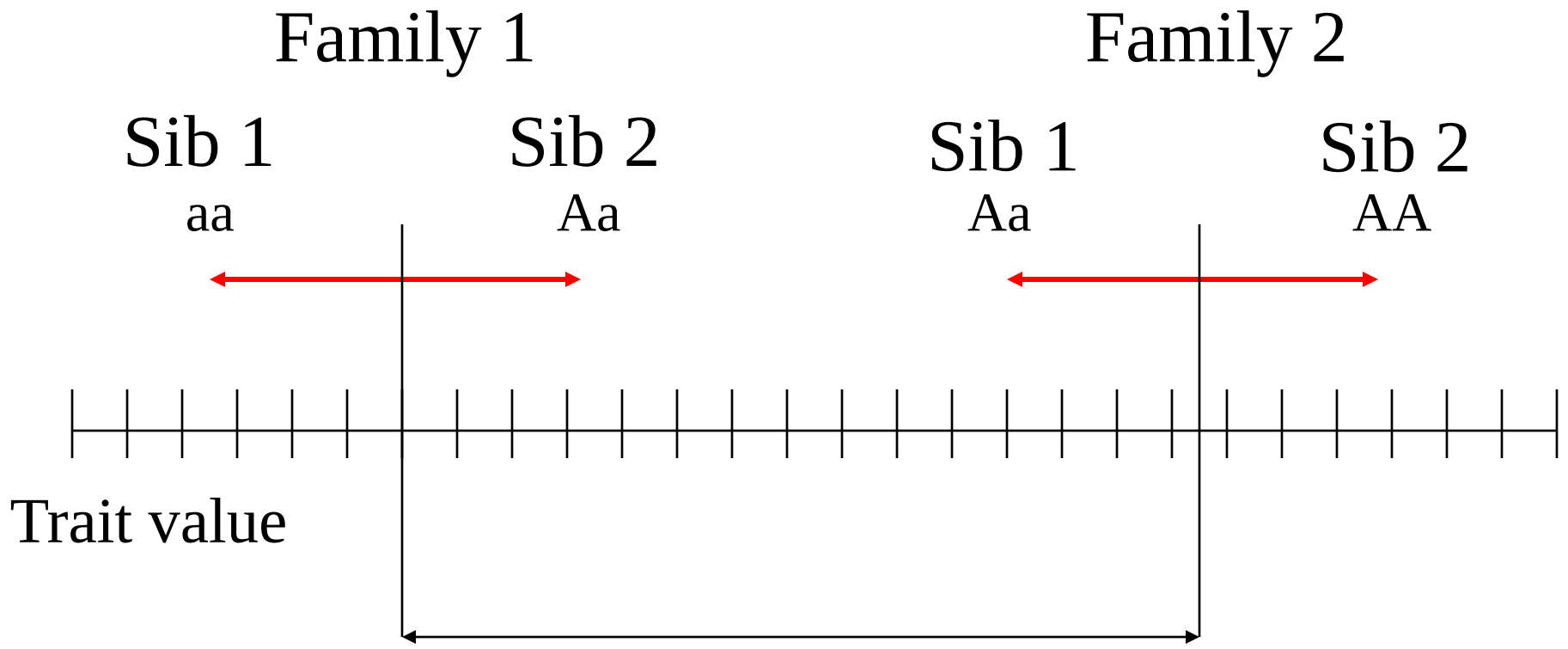
Fulker model of association

- Sibling analysis of quantitative phenotypes
- Breaks down association information into ‘between’ and ‘within’
- Between information compares the average genotype against the average phenotype across all families
- Within information compares the difference in genotype against the difference in phenotype within each family

Visual Representation



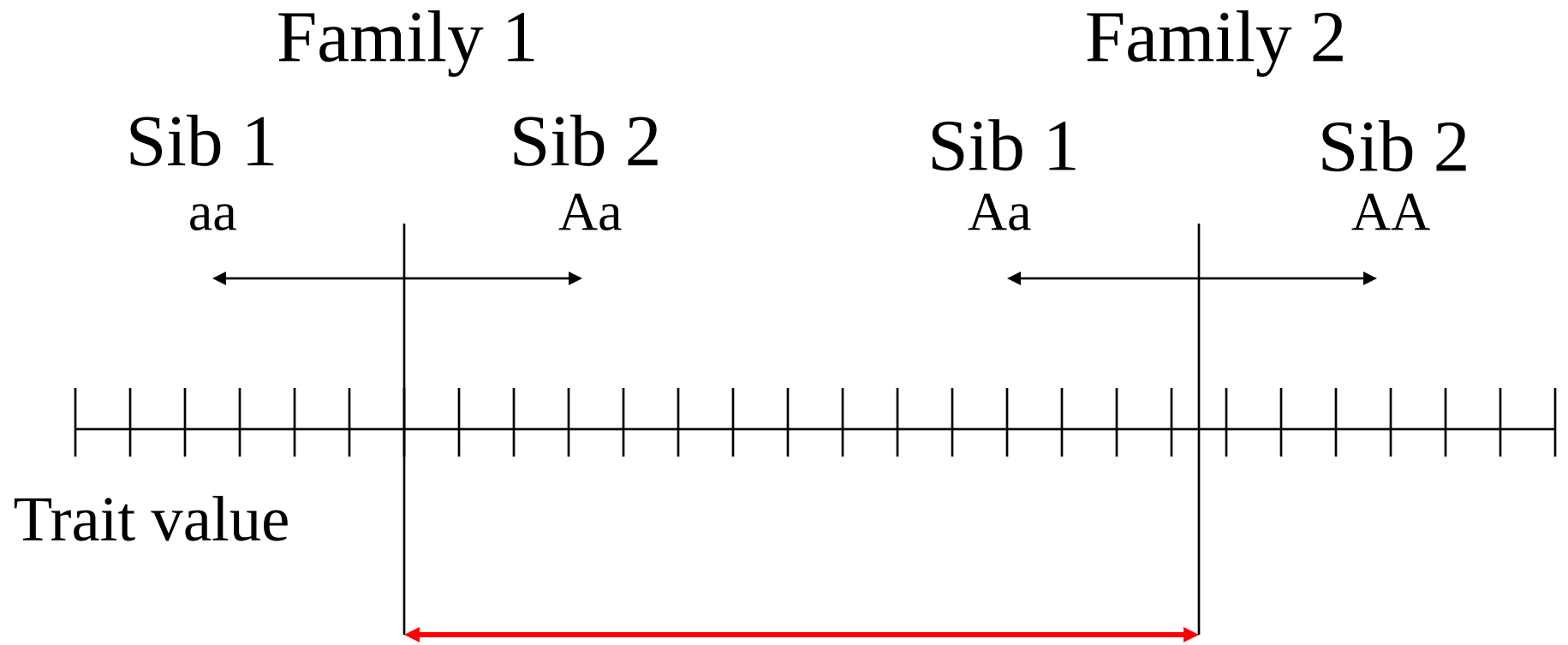
Visual Representation



Within component

Difference in phenotype within sibship against difference in genotype [here we see association to A]

Visual Representation



Between component

Average phenotype per sibship against difference in genotype [Here we see association to A]

Potential tests

- B = Between component
- W = Within component

H_a	H_a	Test
BW	B	Within test
W	0	Within test
BW	W	Between test
B	0	Between test
B=W	0	Combined test
BW	B=W	Stratification Test

Outline

1. Definition of terms
2. Population-based association
3. Stratification
4. Family-based association
5. Direct vs. Indirect association

Allelic Association
Three Common Forms

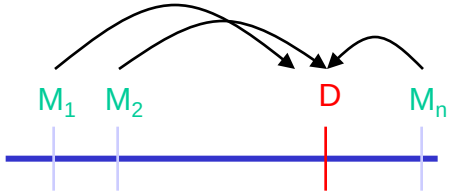
- Direct Association
  - Mutant or ‘susceptible’ polymorphism
  - Allele of interest is itself involved in phenotype
- Indirect Association
  - Allele itself is not involved, but a nearby correlated marker changes phenotype
- Spurious association
  - Apparent association not related to genetic aetiology (most common outcome…)

Indirect and Direct Allelic Association

Direct Association

Measure disease relevance (*) directly, ignoring correlated markers nearby

Indirect Association & LD



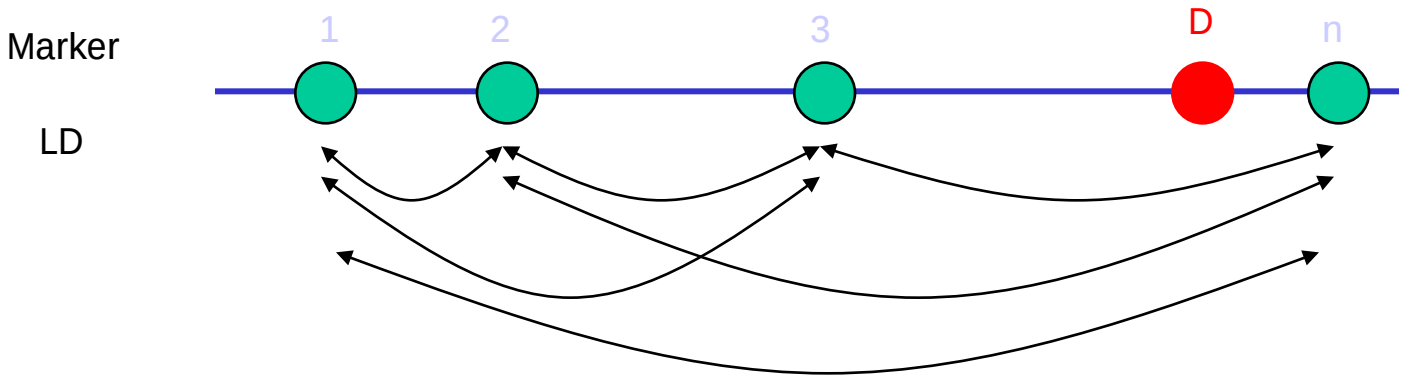
Assess trait effects on D via correlated markers (M_i) rather than susceptibility/etiologic variants.

Semantic distinction between

Linkage Disequilibrium: correlation between (any) markers in population

Allelic Association: correlation between marker allele and trait

Linkage Disequilibrium Maps & Allelic Association



Primary Aim of LD maps: Use relationships amongst background markers (M_1, M_2, M_3, …M_n) to learn *something* about D for association studies

Something =

* Efficient association study design by reduced genotyping
* Predict approx location (fine-map) disease loci
* Assess complexity of local regions
* Attempt to quantify/predict underlying (unobserved) patterns

…