Family-based studies

Is your trait heritable?

- Definition of heritability:
  Fraction of phenotypic variability attributable to genetic variation

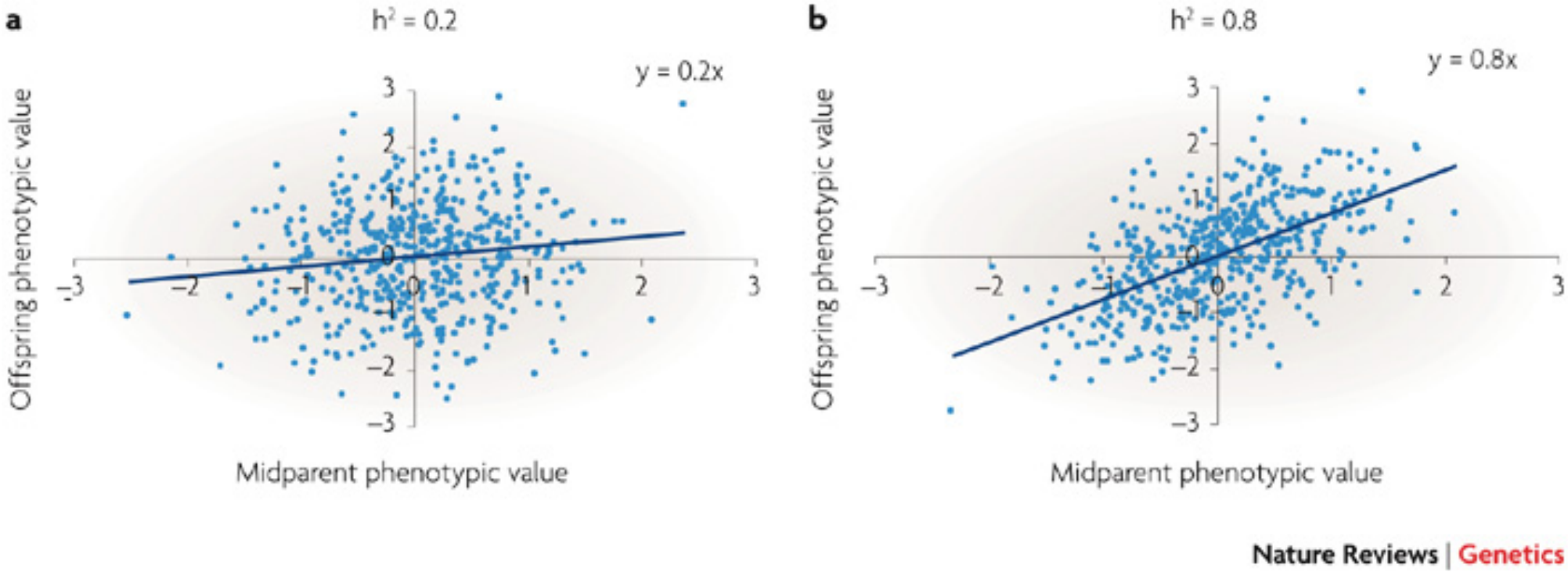
- Aggregation studies:
  - Are relatives of diseased individuals more likely to have the disease compared to the general population?
  - Is the clustering of disease in families different from what you would expect based on the prevalence in the general population?

- Important note: Disease can run in families due to *shared environment* rather than genetics.

Relative Recurrence Risk Ratios, λ_R

- Measures the familial aggregation of disease
- Defined as the risk of disease in relatives of a random individual with disease, divided by the population prevalence of the disease
- Example: λ_S is the ratio of the probability that a sibling to an affected index case is affected over the population prevalence
- Limitation: Both shared genes and shared environment likely contribute to familial aggregation of a complex disease

Evidence of heritability for quantitative traits



Visscher, Nat Rev Genet, 2008

Effect of family history on lifetime prostate cancer risk

Family History	Relative Risk	% Absolute Risk
Negative	1	8
Father affected at 60 yrs. or older	1.5	12
1 Brother affected at age 60 yrs. or older	2	15
Father affected before age 60 yrs.	2.5	20
1 Brother affected before age 60 yrs.	3	25
2 Affected male relatives*	4	30
3 Or more affected male relatives†	5	35–45

* Father and brother, or 2 brothers, or a brother and a maternal grandfather or uncle, or a father and a paternal grandfather or uncle.
† The absolute lifetime risk for mutation carriers is probably 70% to 90% for high penetrance genes.

Bratt, J Urol, 2002

Twin studies

- **Twin studies** allow researchers to examine the overall role of genes in the development of a trait or disorder. Comparisons between monozygotic **twins** and dizygotic **twins** are conducted to evaluate the degree of genetic and environmental influence on a specific trait.

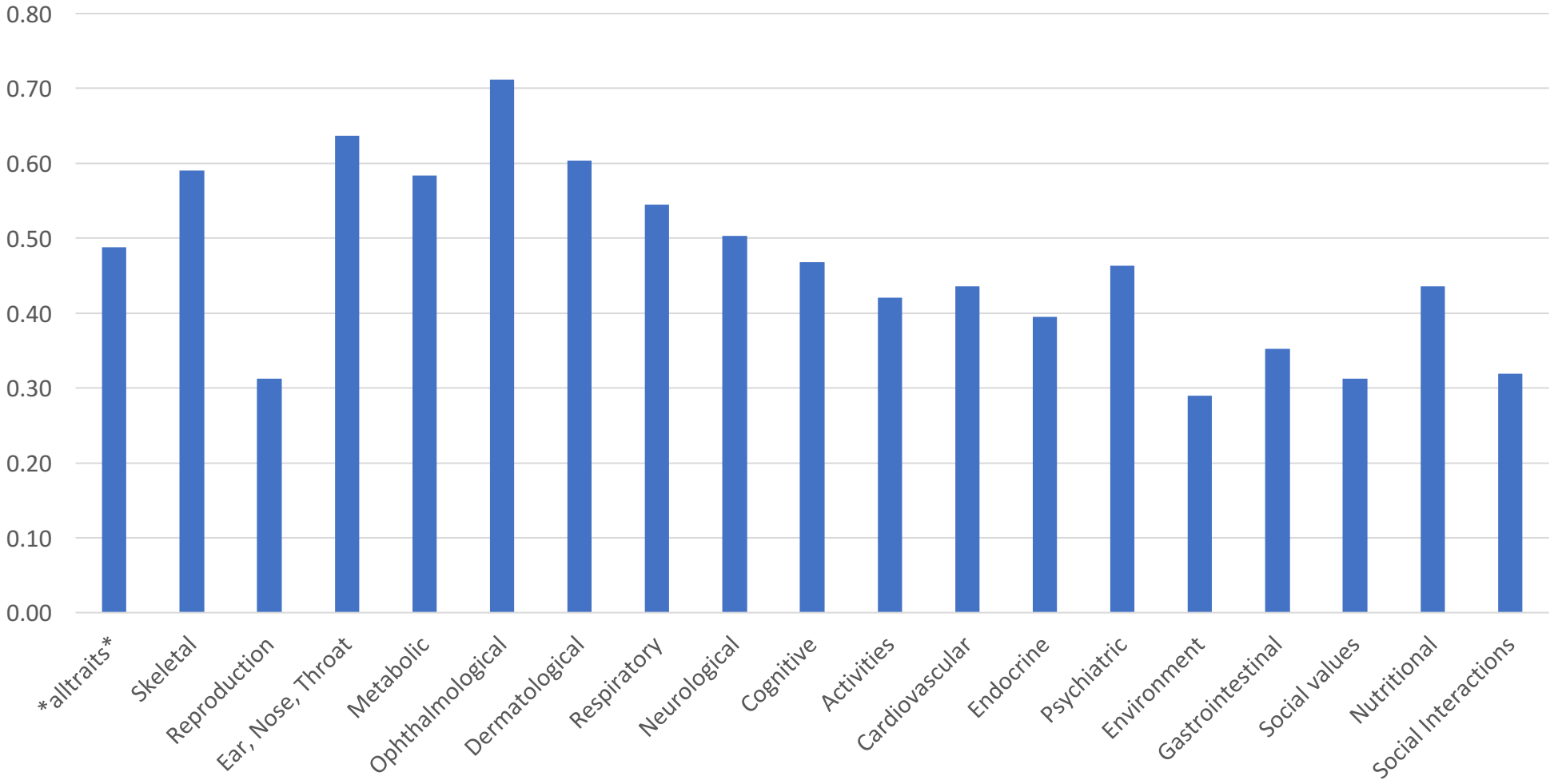
- Since monozygotic (MZ) twins share 100% of their DNA and dizygotic (DZ) twins share 50% of their DNA, we should observe a higher risk in a MZ twin compared to a DZ twin given that one twin is affected.

- Assumes shared environment

Adoption studies

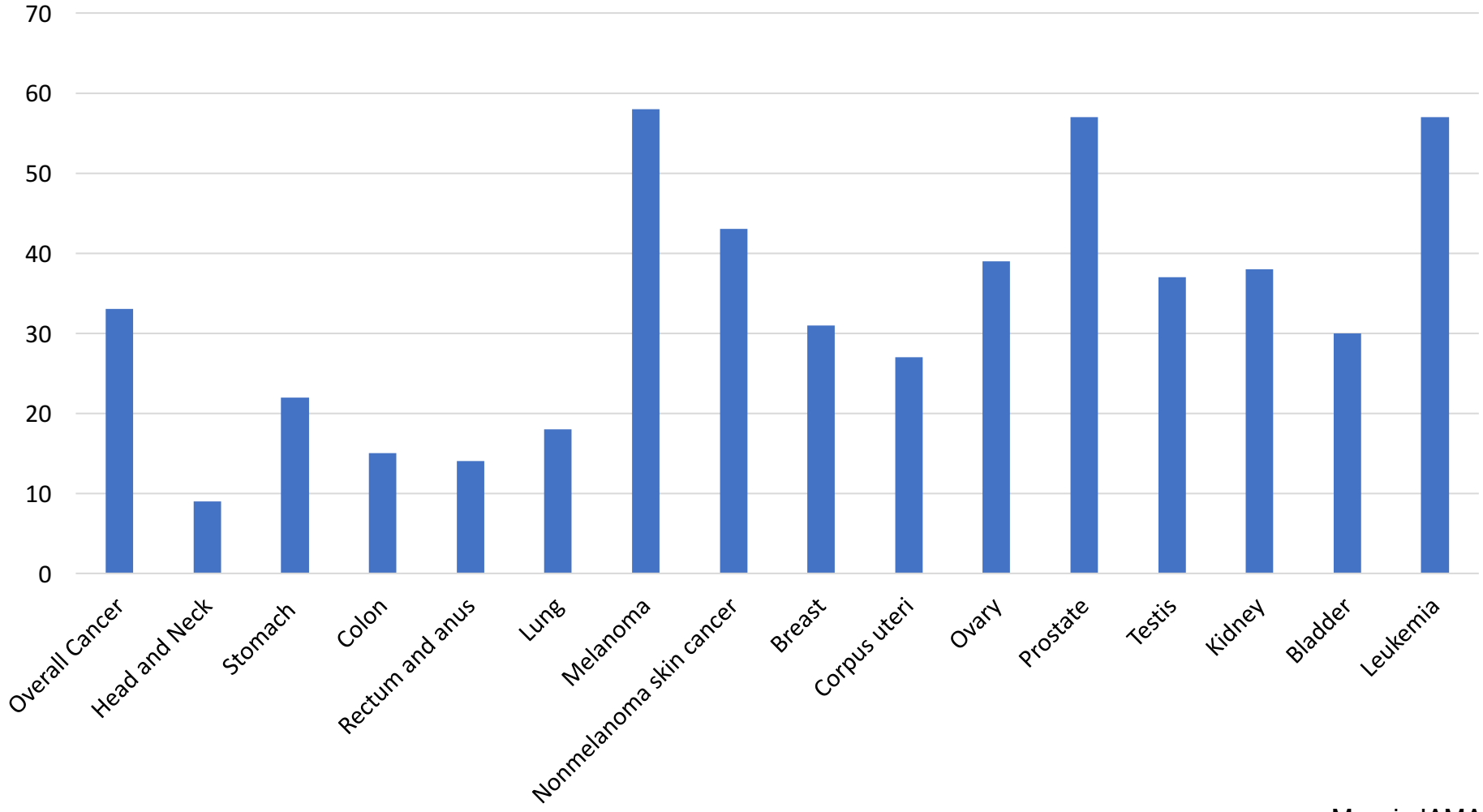
- Idea: Any correlations between offspring and biological parent is due to genetics since there is no shared environment
- Does the disease run in the biological or adopted family?
- If the biological parents are affected, will the adopted offspring get affected? Yes (genetic), No (environmental)

Heritability of categories of traits from twin studies



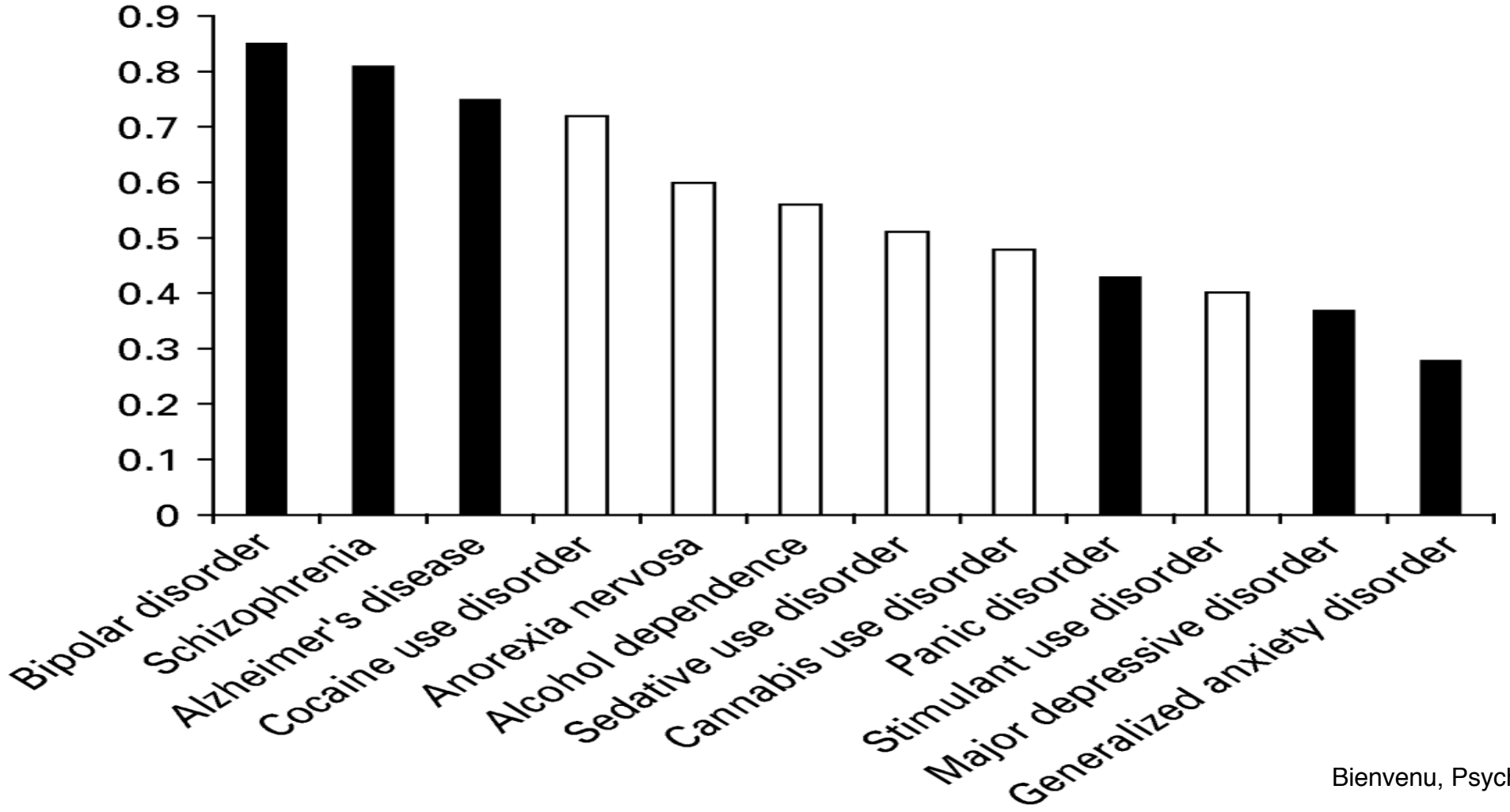
Polderman, Nat Genet 2015

Heritability of cancers



Mucci, JAMA 2016

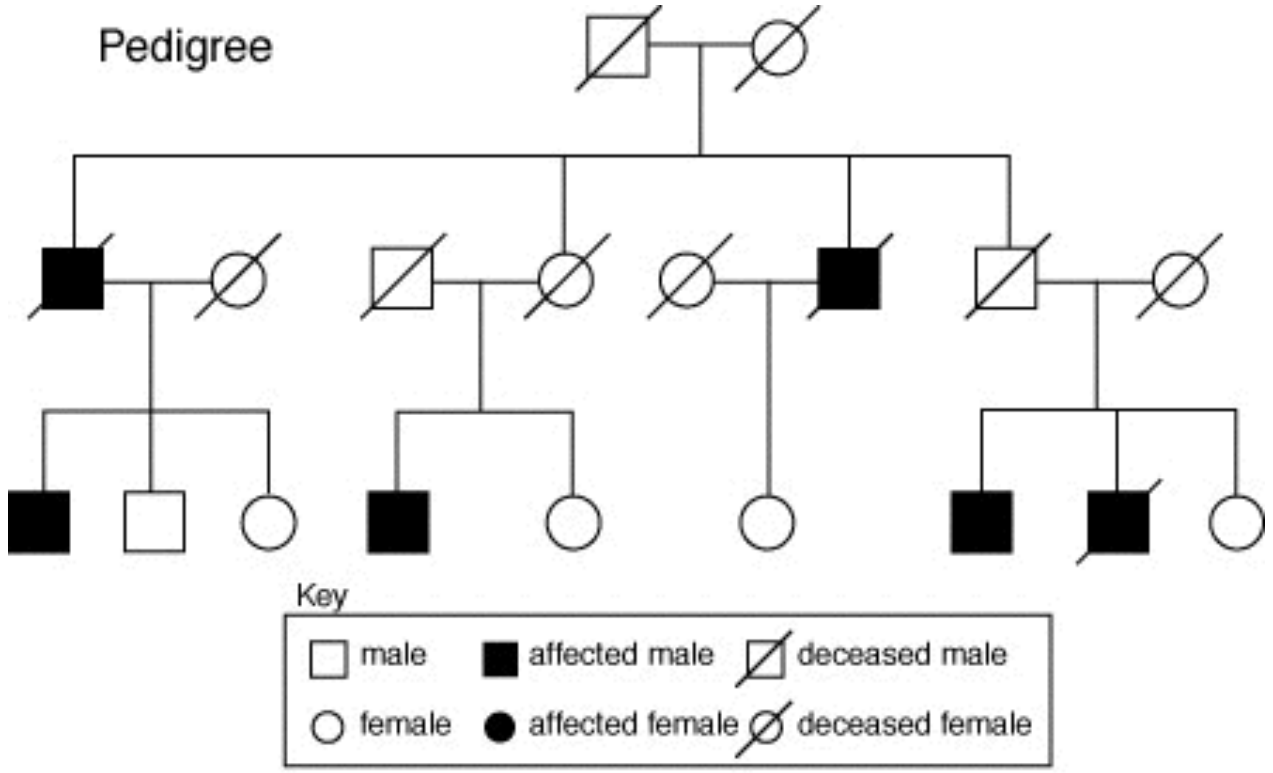
Heritability summary estimates for psychiatric diseases/disease-like conditions and behavioral disorders



Bienvenu, Psych Med 2011

Using family-based designs to identify genetic risk factors

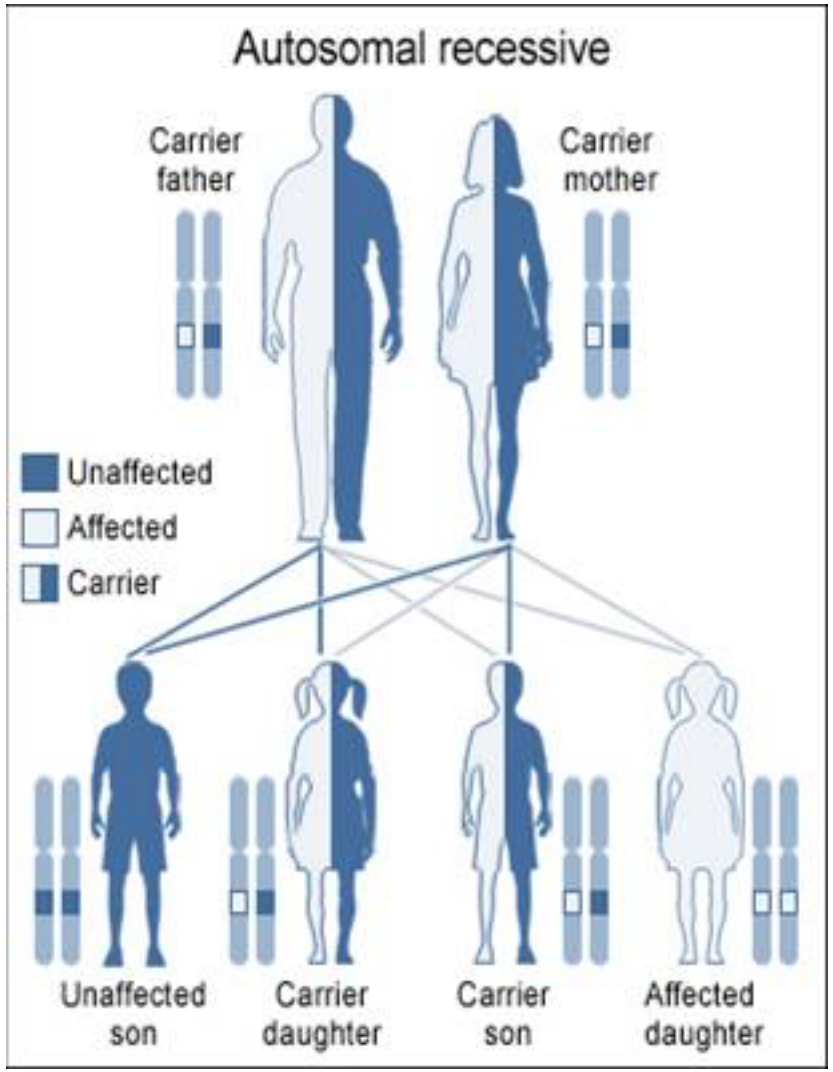
- Linkage analysis
- Family-based association tests



- Penetrance: *The proportion of individuals with a particular genotype that express the phenotype. When 100% of individuals with the causal genotype have the phenotype, penetrance is complete. When less than 100% of individuals with the influential genotype have the phenotype, penetrance is incomplete.*

Modes of inheritance

- Single gene
  - Autosomal
    - Recessive
    - Dominant
    - Codominant
  - X-linked
    - Recessive
    - Dominant
  - Y-linked
  - Mitochondrial, imprinted
- Polygenic



Examples: Cystic Fibrosis, Sickle cell anemia

Segregation analysis

- *The statistical methodology used to determine from family data the mode of inheritance of a particular phenotype, especially with a view to elucidating single gene effects; it is thus a basic tool in human genetics.*

  R.C Elston, 1981

- What proportion of the trait is due to genetic factors?
- What mode of inheritance best represents the genetic factors?
  - Single gene, polygenic, dominant, recessive?

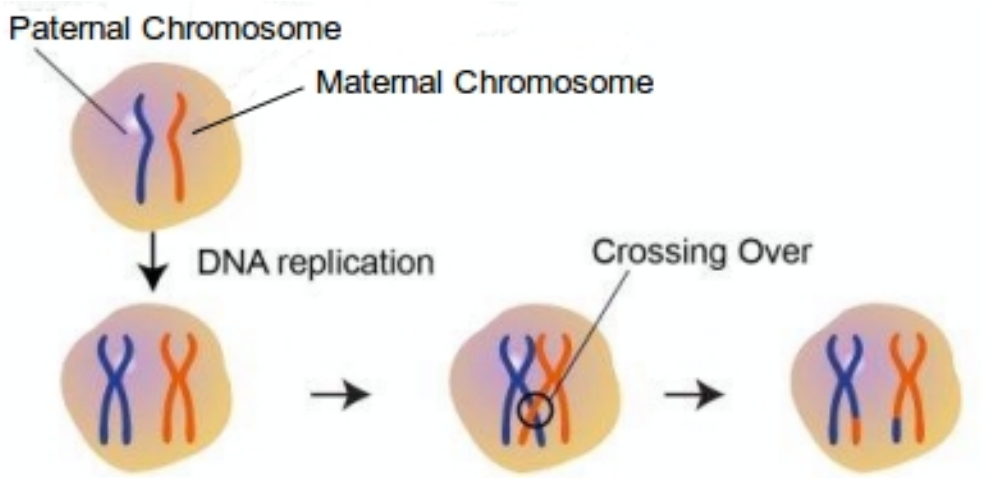
- Based on pedigrees, not genetic data
- Maximum-likelihood approach to identify the best genetic model
- For complex traits, segregation results have shown heterogeneous results
- Ascertainment issues: If pedigrees are collected through a proband (affected person), the probability that a family is sampled will increase with number of affected.

Linkage Analysis (I)

- Widely used method for mapping disease genes using families
- Look for evidence of co-segregation of disease marker and genetic marker within a family
- Likelihood-based method
- Builds on the concept of recombination

Recombination

- Alleles on the same chromosome are inherited together unless *recombination (crossing over)* occurs
- The probability of recombination between two alleles increases with the distance between them
- The distance between two loci with an expected average number of 0.01 recombinations in a single generation is called a **centimorgan** (cM). On average 1 cM corresponds to 1 million base pairs
- The parameter θ estimates the probability of observing a recombinant gamete (the *recombination fraction*)
- $0<\theta<0.5$ where $\theta=0$ means complete linkage and $\theta=0.5$ means no linkage



Recombination between 2 homologous chromosomes

Linkage analysis (II)

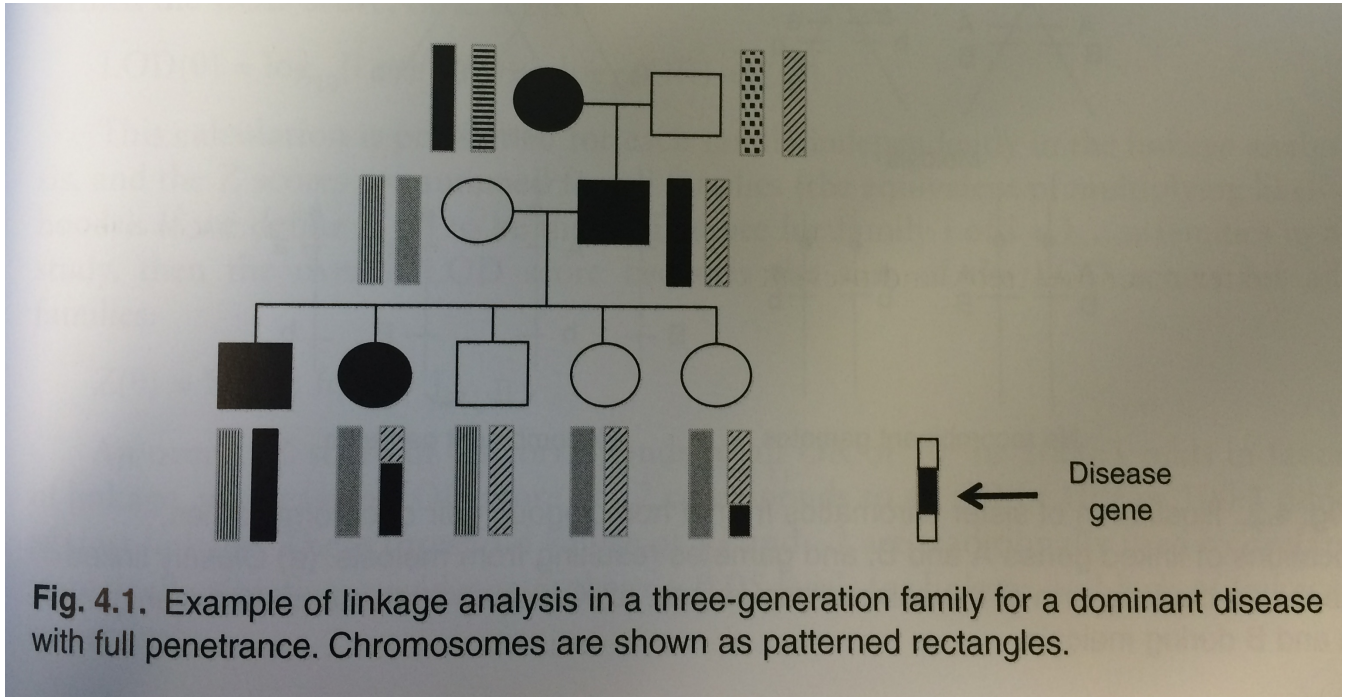


Fig. 4.1. Example of linkage analysis in a three-generation family for a dominant disease with full penetrance. Chromosomes are shown as patterned rectangles.

- Families are only fully informative when one parent is a double heterozygote (for disease and genetic marker)
- Linkage analysis is parametric (model-based) since you make an assumption about a particular mode of inheritance that explains the pattern in the family

Microsatellites – tool used in linkage analysis

- Most popular markers for linkage analysis
  - String of repetitive DNA where a sequence of 1-6 base pairs is repeated
  - Large number of alleles (average 5-10)
  - Can distinguish and track individual chromosomes in families

- Relatively abundant in the genome
  - ~15,000 mapped loci
  - 500-1,500 are often used in linkage analysis

LOD score method

- The LOD score compares the likelihood of observing your data if the two loci (your marker and disease loci) are linked, to the likelihood of observing the same data purely by chance.

$$LOD = Z = \log_{10} \frac{L(\hat{\theta})}{L(\theta_0)} = \frac{\text{Likelihood of the data if loci are linked at a particular } \theta}{\text{Likelihood of the data if loci are unlinked } (\theta = 0.50)}$$

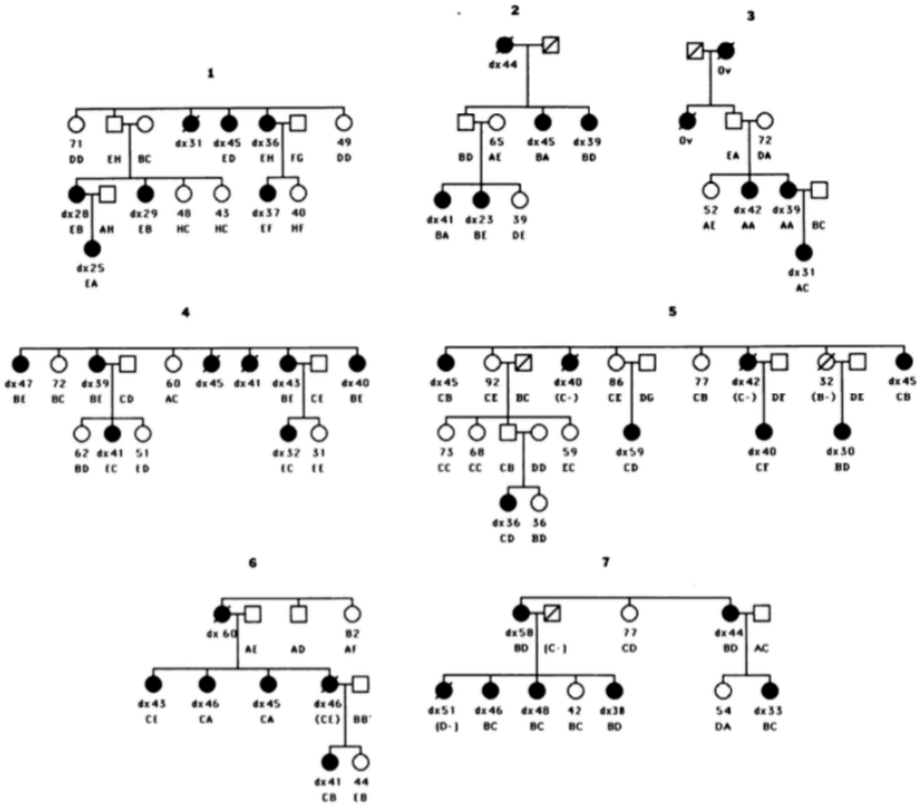
- Positive LOD scores favor the presence of linkage, whereas negative LOD scores indicate that linkage is less likely. A LOD score of 3.3 or higher has been shown to correspond to a genome-wide significance level of 0.05.

- LOD scores can be added across families

Segregation analysis of a population-based series of 1,500 families indicated a rare autosomal dominant allele with major effect

246 cases from 23 families

LOD score: 5.98

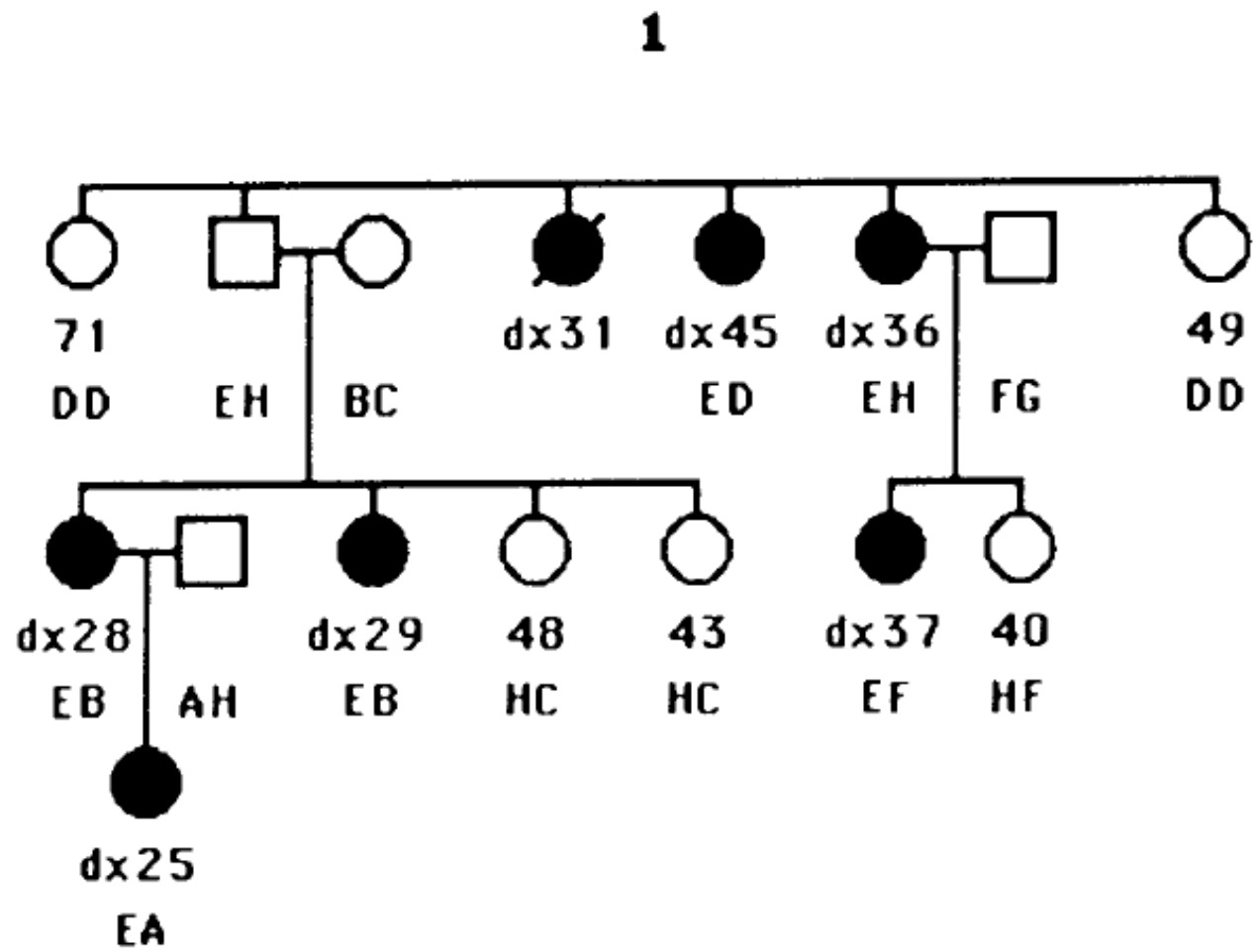


Research Articles

Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21

JEFF M. HALL, MING K. LEE, BETH NEWMAN, JAN E. MORROW, LEE A. ANDERSON, BING HUEY, MARY-CLAIRE KING

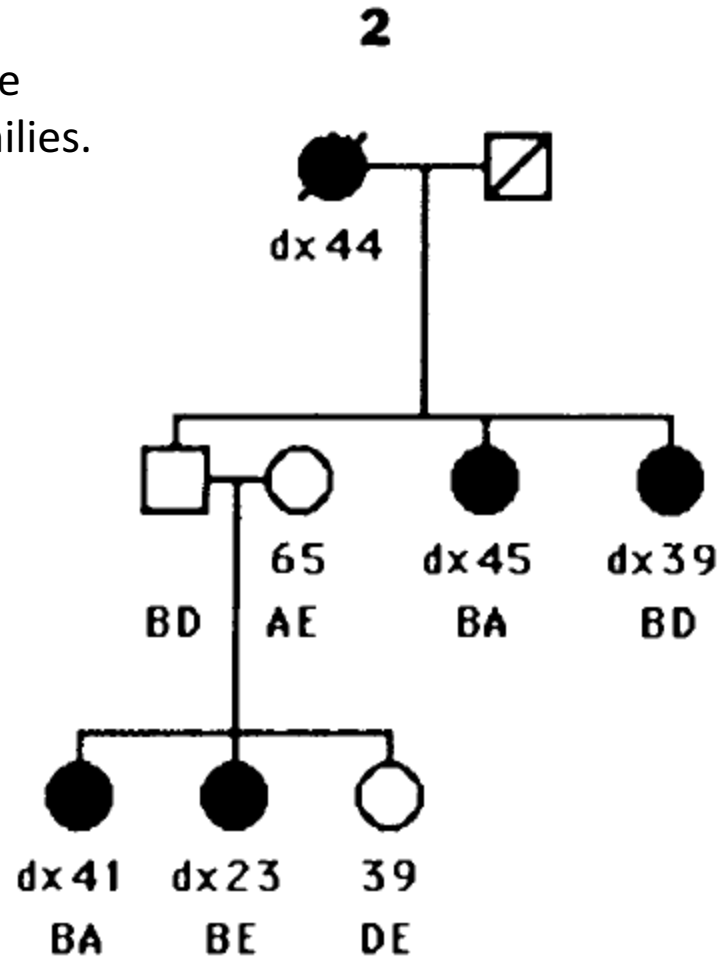
Hall, Science 1990



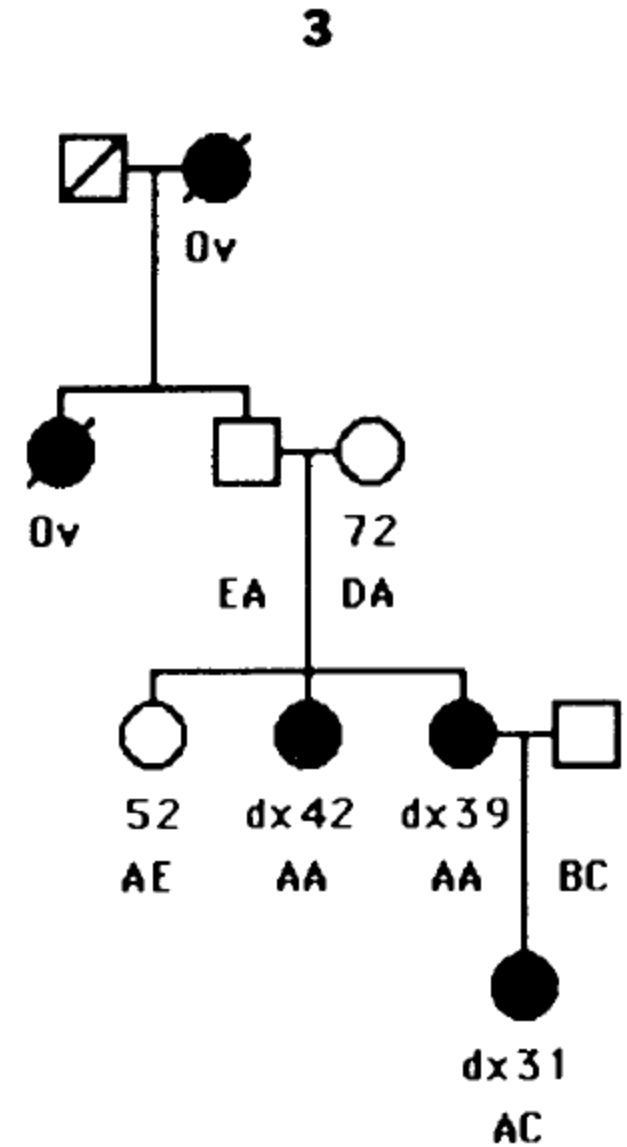
BRCA1 (DS17S74) and the E allele cosegregate

Hall, Science 1990

Different alleles can be linked in different families.



BRCA1 (DS17S74) and the B allele cosegregate



BRCA1 (DS17S74) and the A allele cosegregate

Hall, Science 1990

1) Which marker (A,B,C) is linked to the disease?
2) Which allele (1-5) is linked to disease?

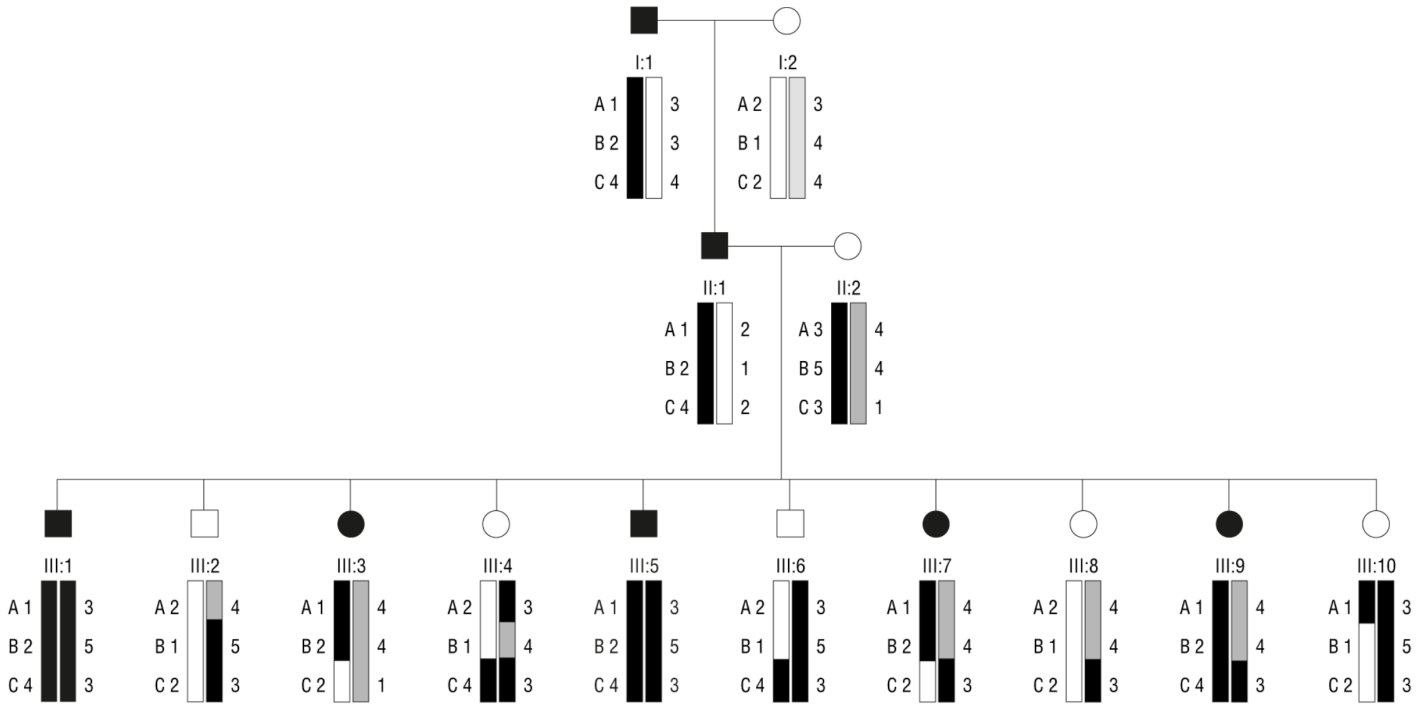


Figure Legend:

Three-generation pedigree segregating an autosomal dominant trait. Alleles at 3 marker loci designated A, B, and C are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.

Arch Neurol. 1999;56(6):667-672.

Linkage analysis is especially powerful for Mendelian traits

- Very effective for finding rare, highly penetrant mutations
  - These are hard to identify using large-sample association methods
  - Pr(carry risk allele|disease) is high

- For complex traits:
  - Pr(carry risk allele | disease) is low.
  - Low resolution: Linkage peaks typically > 1 Mb
  - Sample size considerations: For low-to-moderate effect estimates, you will need very large sample sizes

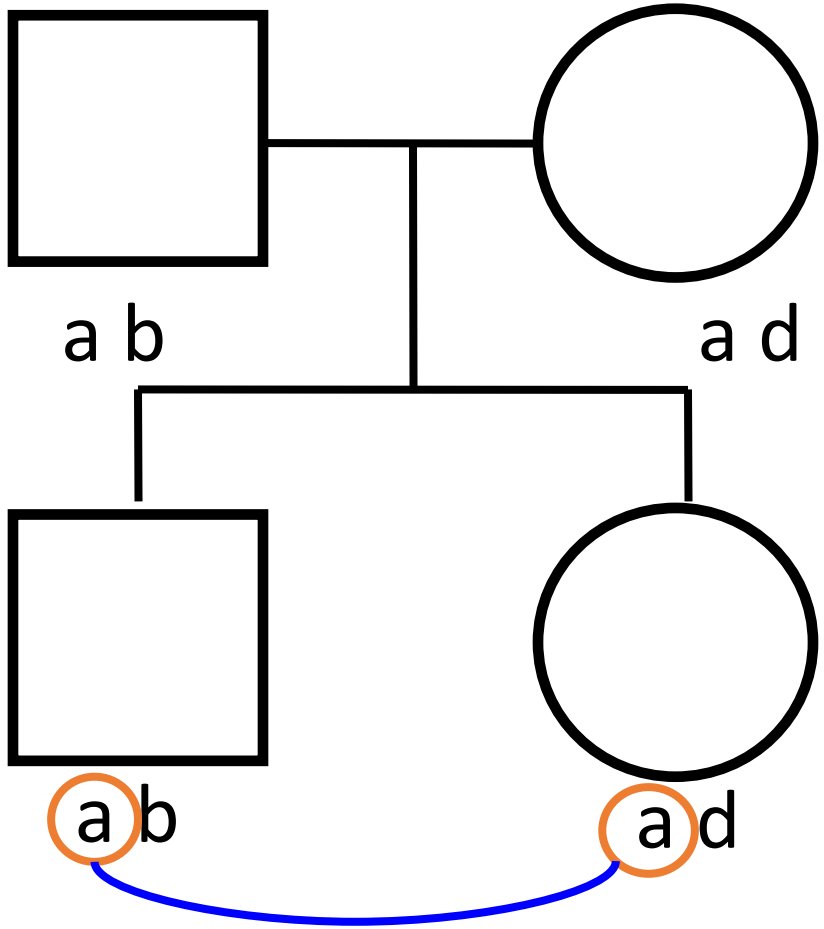
- Note: There are examples of complex traits with “Mendelian-like” subsets
  - Breast cancer, colon cancer, Alzheimers

Model-free linkage analysis

- Based on the assumption on *similarity:*
  - Relatives that are phenotypically similar should also be genotypically similar and relatives that are phenotypically dissimilar should be genotypically dissimilar
- Based on calculating IBS and IBD
- Two examples: Affected sib-pair analysis, Affected pedigree member analysis

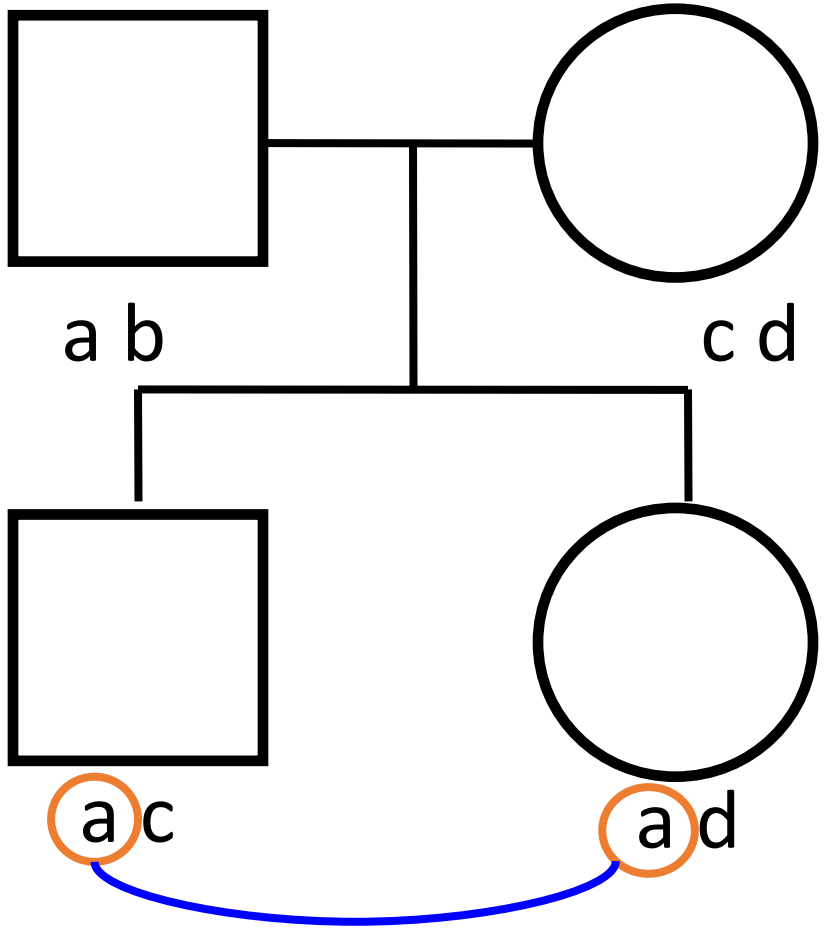
Identical by state - IBS

- Two alleles are said to be IBS if they are identical copies of the same allele.



Identical by descent - IBD

- Two alleles are said to be IBD if they are identical copies of the same allele and inherited from a common ancestor



Affected sib-pair tests (ASP)

- Expected IBD between siblings: $(0, 1, 2) = (0.25, 0.5, 0.25)$
- Under the alternative hypothesis of linkage, this relationship will not hold true for disease locus.
- If we can determine IBD for all affected sib-pairs and let n_i be number of sib-pairs that share *i* alleles IBD, we can test for linkage.

$$\text{Test statistic: } \sum \frac{(n_i - E(n_i))^2}{E(n_i)}$$

- There are multiple approaches (likelihood ratio tests, score tests) and software to conduct ASP.

Affected Pedigree Member (APM) method

- Uses extended pedigrees and IBS instead of IBD (IBD would be difficult to estimate for distant relatives)
- Aims to detect departures from independent segregation of disease and marker phenotypes.
- Does not require any explicit assumptions about the mode of inheritance of the disease.
- For details, see *Weeks and Lange, AJHG 1988*

Family-based association studies

- For rare diseases, family-based association tests are more powerful than case-control studies
- Robust against population stratification

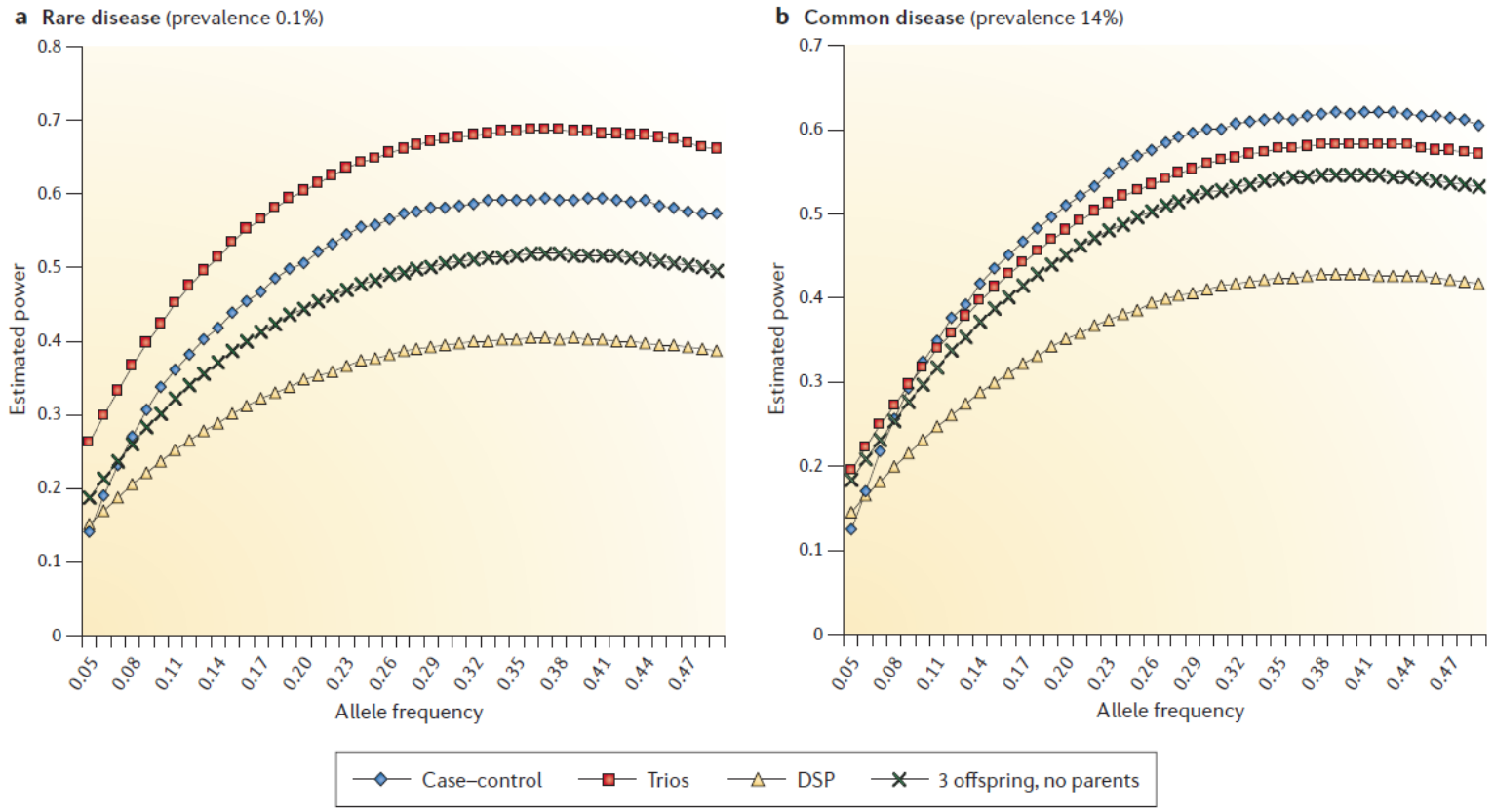


Figure 1 | **Power comparison between case–control studies and family-based designs.** The estimated power levels for a case–control study with 200 cases and 200 controls are compared with those for various family-based designs: 200 trios (of an affected offspring plus parents); 200 discordant sibling (sib) pairs (DSPs; one affected and one unaffected) without parents; 200 '3 discordant offspring (at least 1 affected, at least 1 unaffected) and no parents'. Discordant-sib pair designs have 50% less power than case–control designs, as has been previously noted[75]. For the rare diseases (**a**), trio designs are more powerful than case–control designs. For common diseases (**b**), case–control designs are slightly more powerful than trio designs and designs with 3 discordant sibs. Although it is not shown here, for larger-effect sizes (for example, odd ratios greater than 2), unaffected probands contain more information, and the DSP design can achieve power levels that are similar to those of trios designs[14]. The power calculations for both the family designs and the case–control designs were done in PBAT (v3.3) using Monte-Carlo simulations.

Discordant Sib Pairs (DSP)

- Compare the genotype of affected sibling to the genotype of unaffected sibling
- Use conditional logistic regression
  - Match set is sib pair
- Because the pairs are perfectly matched on genetic ancestry, population stratification is not an issue

Trios: Study Design of Affected Offspring and Both Parents

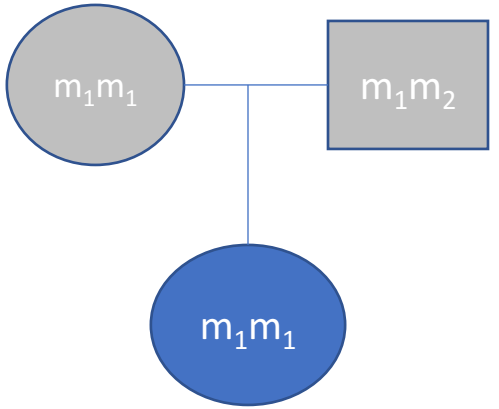
- Phenotypic assessment only in affected offspring.
- Genotyping in both parents and affected offspring.
- Compare the genotype of the affected offspring to that expected under the null, given the parents' genotypes
- Advantage: Not susceptible to population stratification

Transmission disequilibrium test (TDT)

- Tests whether an allele at given locus (linked to disease) is transmitted to affected offspring more frequently than expected by chance.

- Heterozygous parents transmit alleles m_1 and m_2 at given locus with equal frequency (50%); affected offspring should receive disease-associated allele more frequently.

- Eliminates the need for control group.

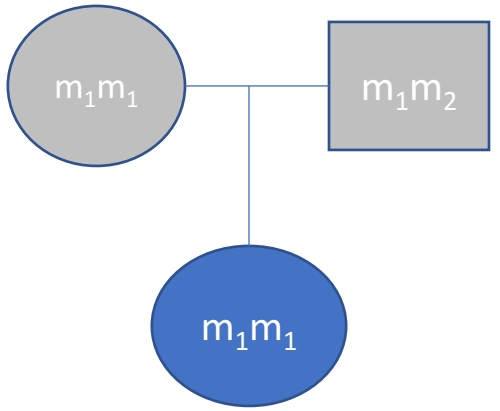


Transmitted Allele

Non-Transmitted Allele	m_1	m_2
m_1	a	b
m_2	c	d

Transmission disequilibrium test (TDT)

- Each family contributes two observations to this table, one for each parent.
- The example family to the right contributes one count to cell a and one count to cell c.
- Estimated per-allele odds ratio is c/b; chi-squared test statistic is $(b-c)^2/(b+c)$.
- Note! Only heterozygous parents are informative

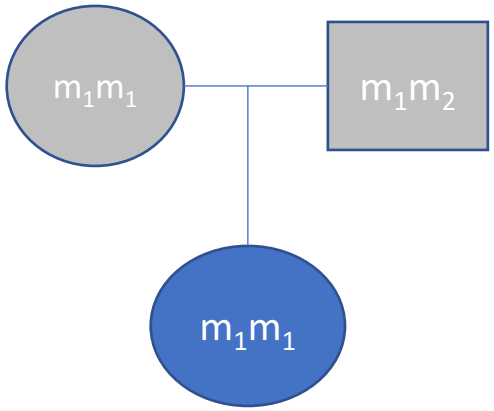


Transmitted Allele

Non-Transmitted Allele		m_1	m_2
	m_1	**a=1**	b=0
	m_2	**c=1**	d=0

Often, only ‘b’ and ‘c’ are given in publications

- ’b’ – ‘Not transmitted’ or “Untransmitted”
  - Number of times the disease allele (here m_1) was not transmitted from parent to offspring
- ‘c’ – Transmitted
  - Number of times the disease allele (here m_1) was transmitted from parent to offspring



Transmitted Allele

Non-Transmitted Allele

	m_1	m_2
m_1	a	**b**
m_2	**c**	d

TDT in Type I Diabetes: Excess Transmission of D18s487 Allele 4

	Transmitted	Not Transmitted	% T	P-value
Affected	c=348	b=276	55.8	0.004
Not affected (healthy siblings)	c=101	b=98	50.8	NS

Merriman T et al. Hum. Mol. Genet 1997

Limitations of Trios

- Difficult to assemble trios if late onset of disease in affected child
- Sensitive to small degrees of genotyping errors which can distort transmission proportions between parents and offspring
- For TDT, only heterozygous parents are informative