INTRODUCTION TO GENETIC EPIDEMIOLOGY

Chia-Ling Kuo

Associate Professor, Department of Public Health Sciences

University of Connecticut Health

Email: kuo@uchc.edu

WHAT IS GENETIC EPIDEMIOLOGY?

- . . . represents an important interaction between the two parent disciplines: genetics and epidemiology
- Different from **Epidemiology** by its explicit consideration of genetic factors and family resemblance
- Different from **Population Genetics** by its focus on disease
- Different from **Medical Genetics** by its emphasis on population aspects

Ziegler A, König IR. *Genet Epidemiol*. 2014;38(5):379-380.

WHAT IS GENETIC EPIDEMIOLOGY?

- a science that is concerned with the etiology, distribution, and control of disease in groups of relatives, and with inherited causes of disease in populations"

[Morton and Chung, 1978]

- a science that deals with the etiology, distribution, and control of disease-related phenotypes in groups of relatives, and with inherited causes of disease-related phenotypes in populations

[Ziegler and König, 2014]

Ziegler A, König IR. *Genet Epidemiol*. 2014;38(5):379-380.

CLASSICAL STEPS IN GENETIC EPIDEMIOLOGY

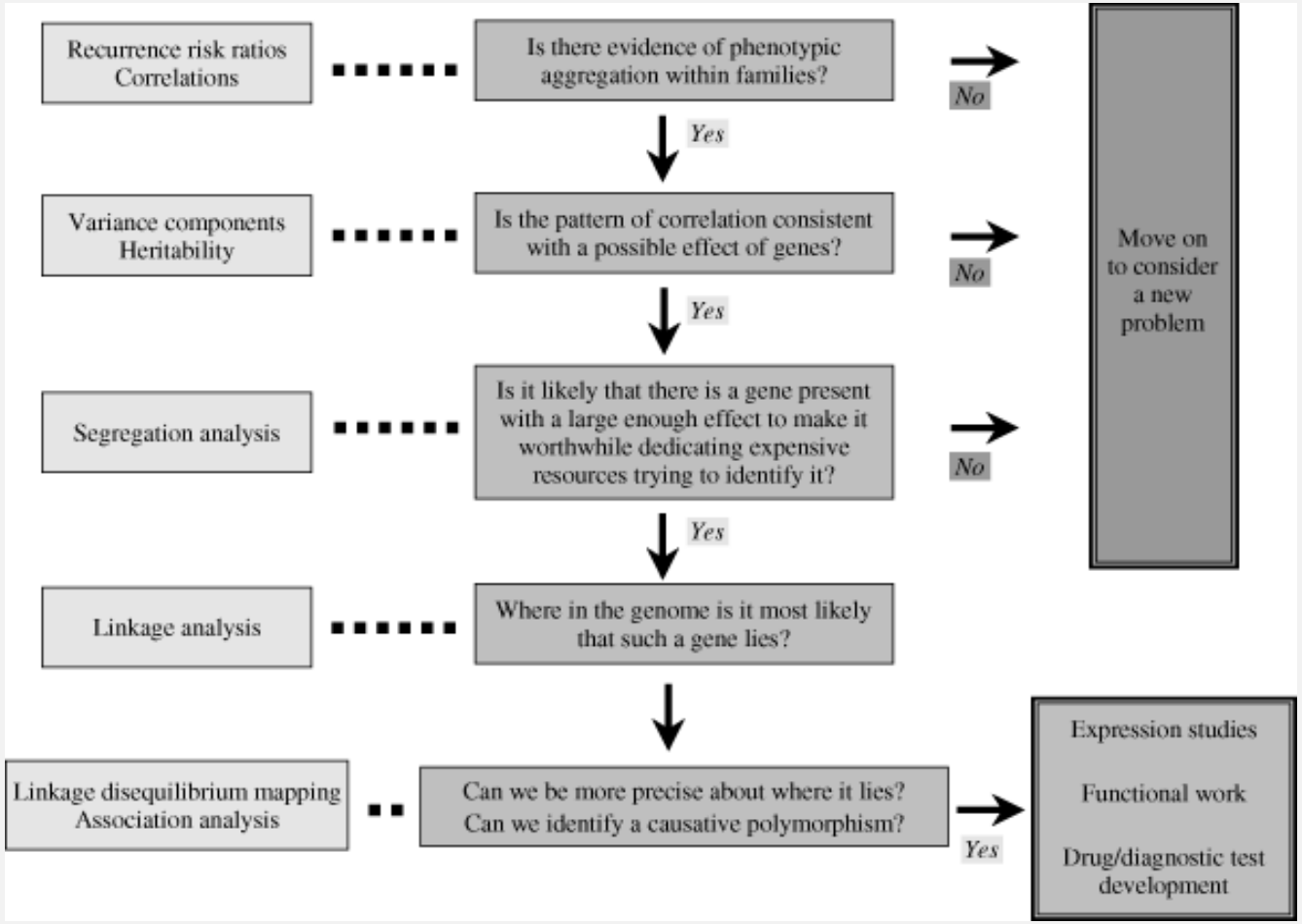
Elston et al. [2007][a]	Thomas [2004]	Ziegler and König [2010]
	1. Descriptive epidemiology	
1. Familiality	2. Familial aggregation	1. Familiality
		2. Heritability
2. Segregation analysis	3. Segregation analysis	3. Segregation analysis
3. Linkage analysis	4. Linkage analysis	4. Linkage analysis
	5. Fine mapping	
4. Association analysis	6. Association analysis	5. Association analysis
	7. Cloning	6. Risk estimation
	8. Characterization	7. Functional studies

Ziegler A, König IR. *Genet Epidemiol*. 2014;38(5):379-380.

CLASSICAL STEPS IN GENETIC EPIDEMIOLOGY

- **Familial aggregation**: first step to see whether it trends to aggregate in families
- **Heritability** to quantify the inheritance assuming a given mode of inheritance
- **Segregation analysis** to characterize the mode of inheritance at a single locus
- **Linkage analysis** and **association analysis**: two main statistical methods for finding disease susceptibility loci

Relevant questions in genetic epidemiology



(Handbook of Statistical Genetics - John Wiley & Sons; Fig.28-1)

K Van Steen slides

FAMILIAL AGGREGATION

- How do we measure extent to which a trait is genetic?
- Two primary measures
  - Recurrence risk ratio (dichotomous traits)
  - Heritability (originally defined for continuous traits; can be adapted to dichotomous disease traits)

RECURRENCE RISK RATIO

- **Recurrence risk ratio** defined for dichotomous disease trait as
- λ_R = P(relative of type R diseased | proband diseased) /P(disease)
  - **Proband**: Subject selected into sample because of disease status.
  - **P(disease) = K**, population prevalence of the disease
  - Relative of type R (parent, sib, etc.)

EXAMPLE: PROSTATE CANCER

Risk Group	λ_R for Prostate Cancer (95% CI)
Brother(s) with prostate cancer diagnosed at any age	3.14 (2.37–4.15)
Father with prostate cancer diagnosed at any age	2.35 (2.02–2.72)
One affected FDR diagnosed at any age	2.48 (2.25–2.74)
Affected FDRs diagnosed <65 y	2.87 (2.21–3.74)
Affected FDRs diagnosed ≥65 y	1.92 (1.49–2.47)
Second degree relatives* diagnosed at any age	2.52 (0.99–6.46)
Two or more affected FDRs diagnosed at any age	4.39 (2.61–7.39)

CI = confidence interval; FDR = first-degree relative.

*The aunts, uncles, grandparents, grandchildren, nieces, nephews, or half-siblings of an individual.

PDQ Cancer Genetics Editorial Board. 2020.

HERITABILITY

- Phenotypic variation attributed to 1) environmental factors, 2) genes, and 3) interactions between genes and environmental factors
- Heritability is a concept that summarizes how much of the variation in a trait is due to variation in genetic factors.
- High heritability implies strong resemblance between parents and offspring with regard to a specific trait.

Wray, N. & Visscher, P. (2008) *Nature Education* 1(1):29

QUANTIFYING HERITABILITY

$V_P = V_G + V_E$ *variance attributed to genetic and environmental sources*

$= (V_A + V_D + V_I) + V_E$ *genetic: additive, dominance, and epistatic*

- **Broad-sense heritability $H^2 = V_G/V_P$**

  prop. of phenotypic variation due to genetic values that may include allelic interactions within loci (dominance) and between loci (epistasis).

- **Narrow-sense heritability $h^2 = V_A/V_P$**

  prop. of genetic variation that is due to additive genetic values (V_A)

Wray, N. & Visscher, P. (2008) *Nature Education* 1(1):29

QUANTIFYING HERITABILITY

- Given its definition as a ratio of variance components, the value of heritability always lies between 0 and 1.
- Resemblance between relatives is mostly driven by additive genetic variance (Hill *et al.*, 2008)
- For instance, for height in humans, narrow-sense heritability is approximately 0.8 (Macgregor et al., 2006).

Wray, N. & Visscher, P. (2008) *Nature Education* 1(1):29

ESTIMATING HERITABILITY

- Traditionally, heritability was estimated from simple and often balanced designs.
  - Simple functions of the regression of offspring on parental phenotypes, the correlation of full or half sibs
  - The difference in the correlation of monozygotic (MZ) and dizygotic (DZ) twin pairs

Visscher, P., Hill, W. & Wray, N. *Nat Rev Genet* **9,** 255–266 (2008).

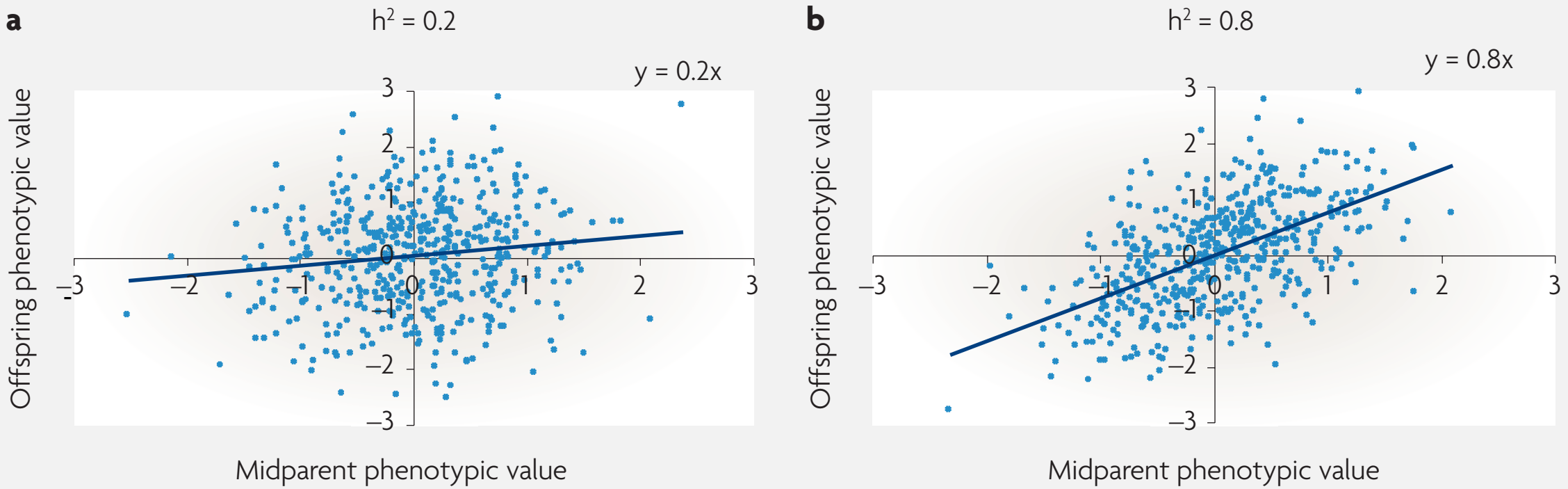


Figure 2 | **Estimation of heritability from the regression of offspring phenotype on the average phenotype of the parents.** The slope of the regression line is an estimate of the narrow-sense heritability for traits with a heritability of 0.2 (**a**) and 0.8 (**b**) and phenotypic variance of 1. The variances of the observations about the regression line are 0.98 (**a**) and 0.68 (**b**), demonstrating that the average phenotypic value of the parents (midparent phenotypic value) is a better predictor of the offspring phenotypic value if heritability is high.

Visscher, P., Hill, W. & Wray, N. *Nat Rev Genet* **9,** 255–266 (2008).

A BETTER STUDY DESIGN…

- If the resemblance of parents and offspring is partly due to common environmental effects, then an estimate of heritability that is based on their resemblance will be biased upwards.
- Phenotypic concordance of monozygotic (MZ, identical) twins versus dizygotic (DZ, fraternal) twins
  - Shared environment for MZ and DZ twins
  - All (MZ) versus half (DZ) shared genome to phenotypic concordance

Visscher, P., Hill, W. & Wray, N. *Nat Rev Genet* **9,** 255–266 (2008).

EXAMPLE: INTELLIGENCE QUOTIENT HERITABILITY

- Cross many studies, the average MZ and DZ correlation of IQ was 0.86 and 0.60, respectively, based on 4,672 MZ and 5,546 DZ twin pairs.
- Falconer's formula:
  - $H^2 = 2(r_{MZ} - r_{DZ}) = 2(86\%-60\%) = 52\%$
- May be overestimated if 1) there is a correlation between genes and environment or 2) there are strong maternal effects on the IQ.

Visscher, P., Hill, W. & Wray, N. *Nat Rev Genet* **9,** 255–266 (2008).

HERITABILITY ESTIMATION USING USING GENOME-WIDE GENETIC DATA

- $Y_{ij} = \mu + F_i + A_{ij} + E_{ij}$, with μ the fixed effects of the mean and F, A, and E the random effects of non-genetic family, additive genetic, and residual factors, respectively.
- The covariance between the phenotypes of two siblings is modeled as $cov(Y_{i1}, Y_{i2}) = var(F_i) + cov(A_{i1}, A_{i2}) = \sigma_F^2 + \pi_{a(i)}\sigma_A^2$, and $cov(Y_{ij}, Y_{kl}) = 0$ if $i \neq k$, with $\pi_{a(i)}$ the estimate of the genome-wide actual additive relationship of the sibling pair.
- $h^2 = \sigma_A^2/(\sigma_A^2+\sigma_F^2 +\sigma_E^2)$

Visscher et al. PLoS Genet. 2006;2(3):e41.

HERITABILITY ESTIMATION USING GENOME-WIDE GENETIC DATA

Linkage disequilibrium score regression

- $\phi = X\beta + \epsilon$ where ϕ is an N × 1 vector of (quantitative) phenotypes, X is an N × M matrix of genotypes normalized to mean zero and variance one.
- The expected $\chi 2$ statistic of variant *j* is:

$$E[\chi^2 \mid \ell_j] = Nh^2\ell_j / M + Na + 1$$

- where *N* is the sample size; *M* is the number of SNPs, such that h^2/M is the average heritability explained per SNP; *a* measures the contribution of confounding biases, such as cryptic relatedness and population stratification; and $\ell_j = \Sigma_k r_{jk}^2$ is the LD Score of variant j, which measures the amount of genetic variation tagged by j.

Bulik-Sullivan, B., Loh, P., Finucane, H. *et al. Nat Genet* **47,** 291–295 (2015).

SEGREGATION ANALYSIS

Chia-Ling Kuo

Associate Professor, Department of Public Health Sciences

University of Connecticut Health

CLASSICAL STEPS IN GENETIC EPIDEMIOLOGY

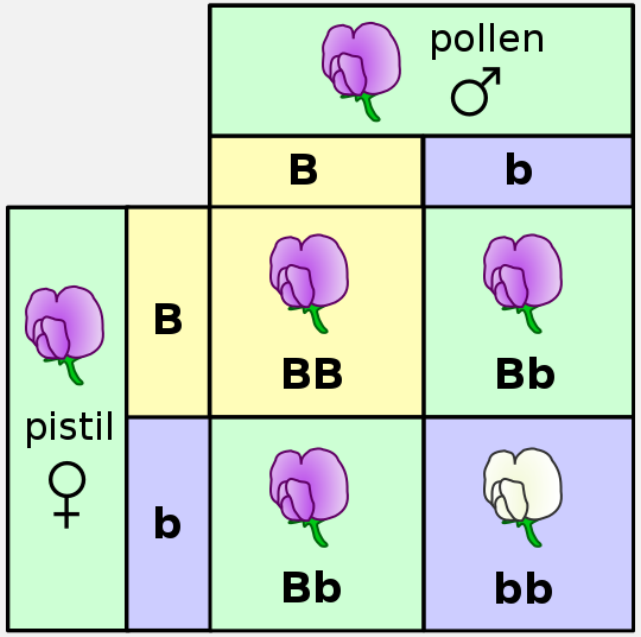
Elston et al. [2007][a]	Thomas [2004]	Ziegler and König [2010]
	1. Descriptive epidemiology	
1. Familiality	2. Familial aggregation	1. Familiality
		2. Heritability
2. Segregation analysis	3. Segregation analysis	3. Segregation analysis
3. Linkage analysis	4. Linkage analysis	4. Linkage analysis
	5. Fine mapping	
4. Association analysis	6. Association analysis	5. Association analysis
	7. Cloning	6. Risk estimation
	8. Characterization	7. Functional studies

SEGREGATION ANALYSIS

- Segregation analysis is a statistical technique that attempts to explain the causes of family aggregation of disease.
- It aims to determine the transmission pattern of the trait within families (often ascertained via probands as in aggregation studies) and to test this pattern against predictions from specific genetic models:
  - - Dominant? Recessive? Co-dominant? Additive?
- This information is useful in parametric linkage analysis, which assumes a defined model of inheritance

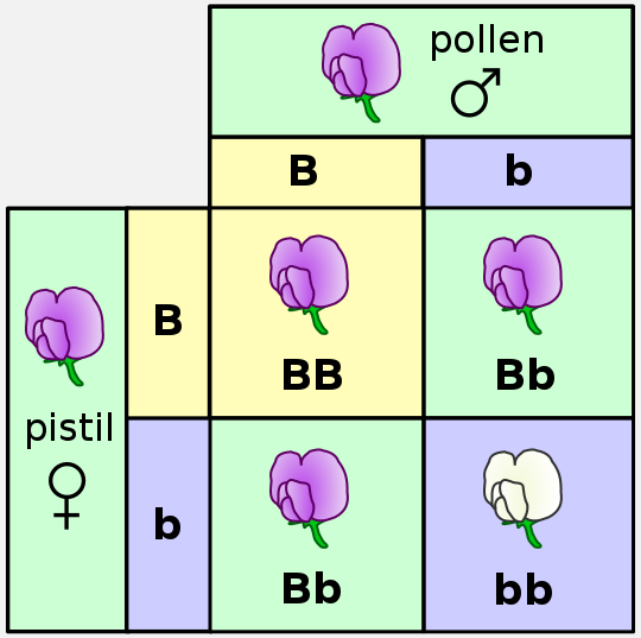
INTRODUCTION TO SEGREGATION ANALYSIS

- In the mid 1800's, Gregor Mendel demonstrated the existence of genes based on the regular occurrence of certain characteristic ratios (segregation ratios) of dichotomous characters (or traits) among the offspring of crosses between parents of various characteristics and lineages.
-



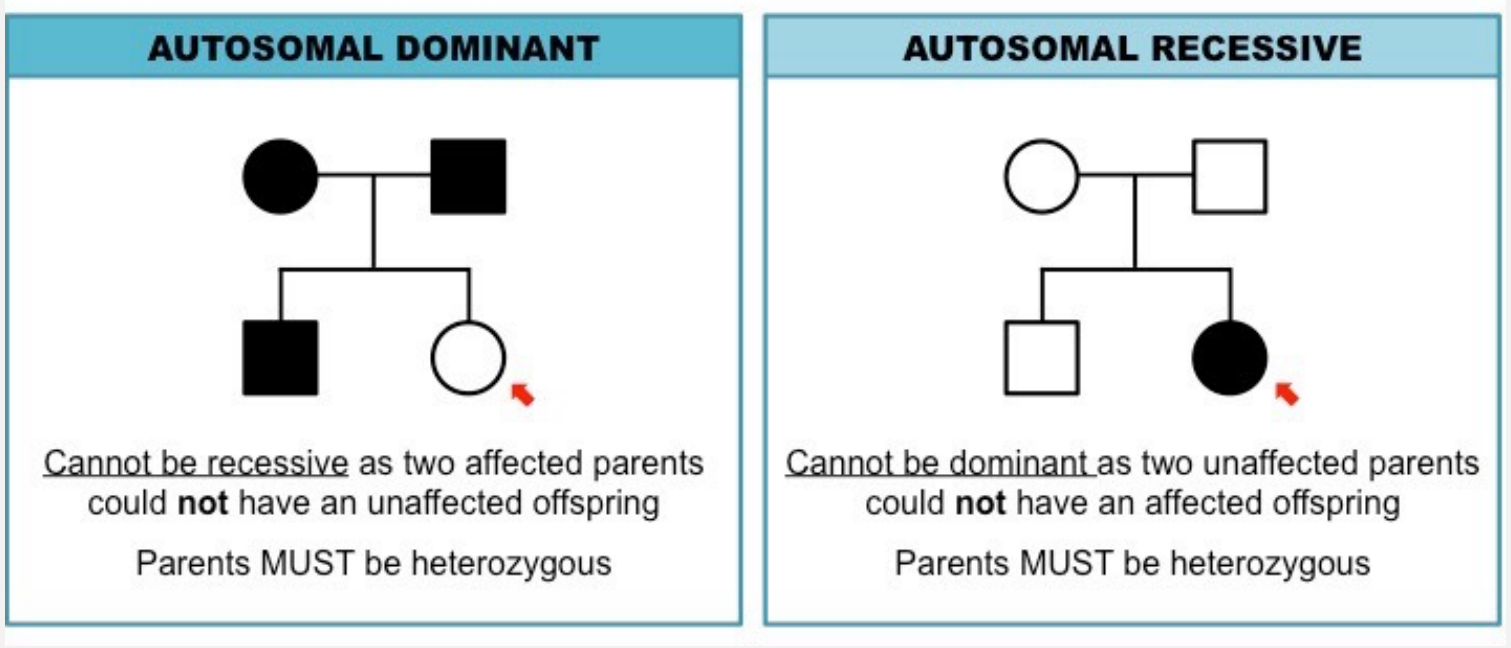
MENDELIAN INHERITANCE

- **Law of Segregation (The "First Law")**: Alleles at any given gene are transmitted randomly and with equal probability.
- **Law of Independent Assortment (The "Second Law"):** alleles of different genes are transmitted independently (we now know this does not apply when loci are located near each other on the same chromosome (linked).



MENDELIAN GENETICS

- Mode of Inheritance is the manner in which a particular genetic trait or disorder is passed from one generation to the next.



BioNinja

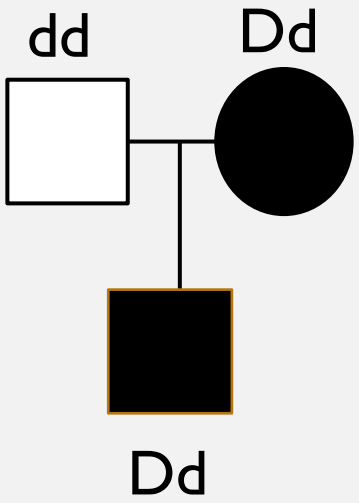
SEGREGATION ANALYSIS

- Segregation analysis entails fitting a variety of models (both genetic and non-genetic; major genes or multiple genes/polygenes) to the data obtained from families and evaluating the results to determine which model best fits the data.

SEGREGATION ANALYSIS FOR AUTOSOMAL DOMINANT DISEASE

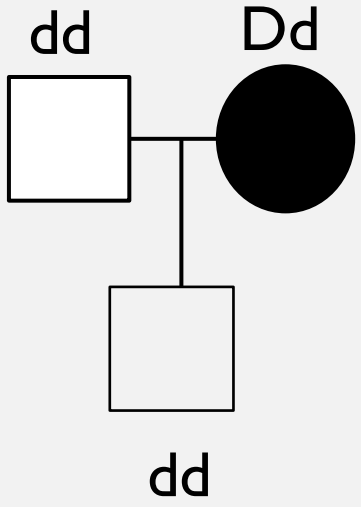
- Consider a disease that is believed to by the caused by a fully penetrant rare mutant allele at an autosomal locus.
- Let D be the allele causing the disorder and let d represent be the normal allele.
- There are 9 possible mating types (DD, Dd, or dd x DD, Dd, or dd)
- Each of these mating types will produce offspring with a characteristic distribution of genotypes and therefore a distribution of phenotypes.
- The proportions of the different genotypes and phenotypes in the offspring of the six mating types are known as the segregation ratios of the mating types.
- These specific values of the segregation ratios can be used to test whether a disease is caused by a single autosomal dominant gene.

K Van Steen slides



SEGREGATION ANALYSIS FOR AUTOSOMAL DOMINANT DISEASE

- Suppose that a random sample of matings between two parents where one is affected and one is unaffected is obtained
- Out of a total of n offspring, r are affected. Since autosomal dominant genes are usually rare, it is reasonable to assume that the frequency of allele D is quite low and that most affected individuals are expected to have genotype of Dd instead of DD.
- What are the matings in the sample under this assumption? How can we test if the observed segregation ratios in the offspring are what is expected if the disease were indeed caused by an autosomal dominant allele? The Binomial distribution can be used to model this data.



K Van Steen slides

AUTOSOMAL DOMINANT DISEASE EXAMPLE

- Marfan syndrome, a connective tissue disorder, is a rare disease that is believed to be autosomal dominant (and actually is!)
- 112 offspring of an affected parent and an unaffected parent are sampled
- 52 of the offspring are affected and 60 are unaffected
- Are these obervations consistent with an autosomal dominant disease?
  - P-value = $Pr(X \leq 52) + Pr(X > 60) = 0.5085$
- What if only 42 of the offspring are affected?
  - P-value = $Pr(X \leq 42) + Pr(X > 68) = 0.0104$

K Van Steen slides

BINOMIAL DISTRIBUTION

- The binomial distribution is a very common discrete probability distribution that arises in the following situation:
  - A fixed number, n, of trials
  - The n trials are independent of each other
  - Each trial has exactly two outcomes: "success" and "failure"
  - The probability of a success, p, is the same for each trial
- If X is the total number of successes in a binomial setting, then we say that the probability distribution of X is a binomial distribution with parameters n and p: $X \sim B(n, p)$, $P(X = x) = C_x^n \, p^x(1 - p)^{(n-x)}$

K Van Steen slides

BINOMIAL DISTRIBUTION

- Let X be the number of offspring that are affected.
- Under the null hypothesis, X will have a binomial distribution

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

  where p is the probability that an offspring is affected.
- We are interested in testing
  - H_0: $p = \frac{1}{2}$ vs. H_a: $p \neq \frac{1}{2}$
- Out of a total of n offspring, r are affected. The p-value is the probability of observing a value at least as extreme as r. If $r < \frac{n}{2}$, the p-value is

$$\sum_{x=0}^{r} \binom{n}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{(n-x)} + \sum_{x=n-r}^{n} \binom{n}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{(n-x)}$$

$$= \left(\frac{1}{2}\right)^{n-1} \sum_{x=0}^{r} \binom{n}{x}$$

K Van Steen slides

COMPLEX SEGREGATION ANALYSIS

- Elston-Stewart Algorithm
  - Likelihood based method with the elements, Pr(data, parameters):
  - $P(G_{founder})$: prior probabilities for founder genotypes
    - Based on allele frequencies assuming Hardy-Weinberg equilibrium
    - E.g., $P(A)=p$, $P(AA)=p^2$, $P(Aa)=2pq$, $P(aa)=(1-p)^2$ assuming HWE
  - $P(G_o \mid G_f, G_m)$: segregation probabilities for offspring genotypes given parents'
    - Mendel's first law
  - $P(X_i \mid G_i)$: penetrances for individual phenotypes given offspring genotypes
    - Complete or incomplete penetrance

COMPLEX SEGREGATION ANALYSIS

$$L = \sum_{G_1} \cdots \sum_{G_n} \prod_{f} P(G_f) \prod_{\{o,f,m\}} P(G_o \mid G_f, G_m) \prod_{i} P(X_i \mid G_i)$$

- Notice the three elements:
  - Probability of founder genotypes
  - Probability of children given parents
  - Probability of phenotypes given genotypes

PENETRANCE PARAMETERS DETERMINE MODEL TYPE

- Consider the following parameterization
  - f_{11}= Pr(Disease|11)= k
  - f_{12}= Pr(Disease|12)= k – c_{12}
  - f_{22}= Pr(Disease|22)= k – c_{22}

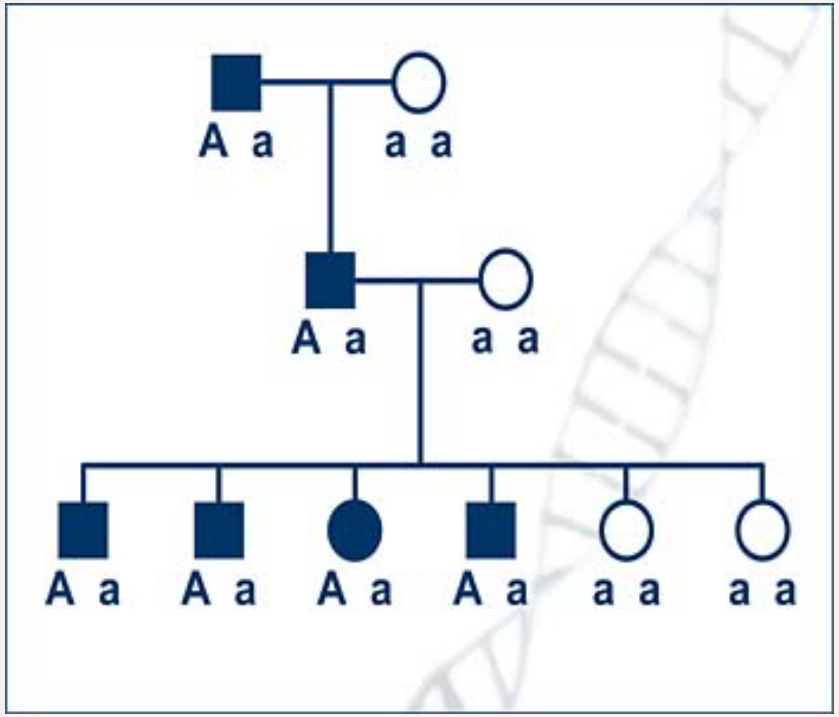
  where $0 \leq c_{12}, c_{22} \leq k$
- What is the relationship between c_{12} and c_{22} for an additive model?
- What are the parameter values for a fully penetrant dominant disease?
- Note that if $c_{12}=c_{22}=0$, then the locus is not involved with the phenotype, and k would be equal to K_p (population prevalence of the disease).

LINKAGE ANALYSIS

Chia-Ling Kuo, Department of Public Health Sciences,

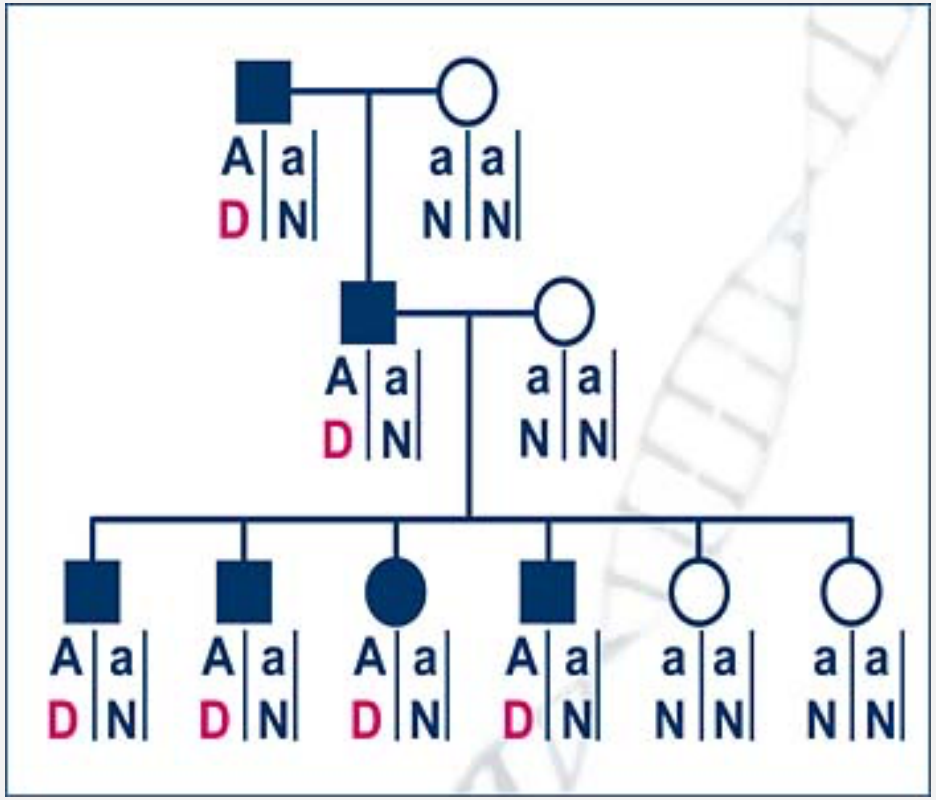
University of Connecticut Health

LINKAGE



- Segregation of a marker allele with disease phenotype within a family represents physical linkage between a marker and a disease locus.
- In this pedigree, the "A" allele segregates with the disease. It is shared identical-by-descent in all the affected individuals.

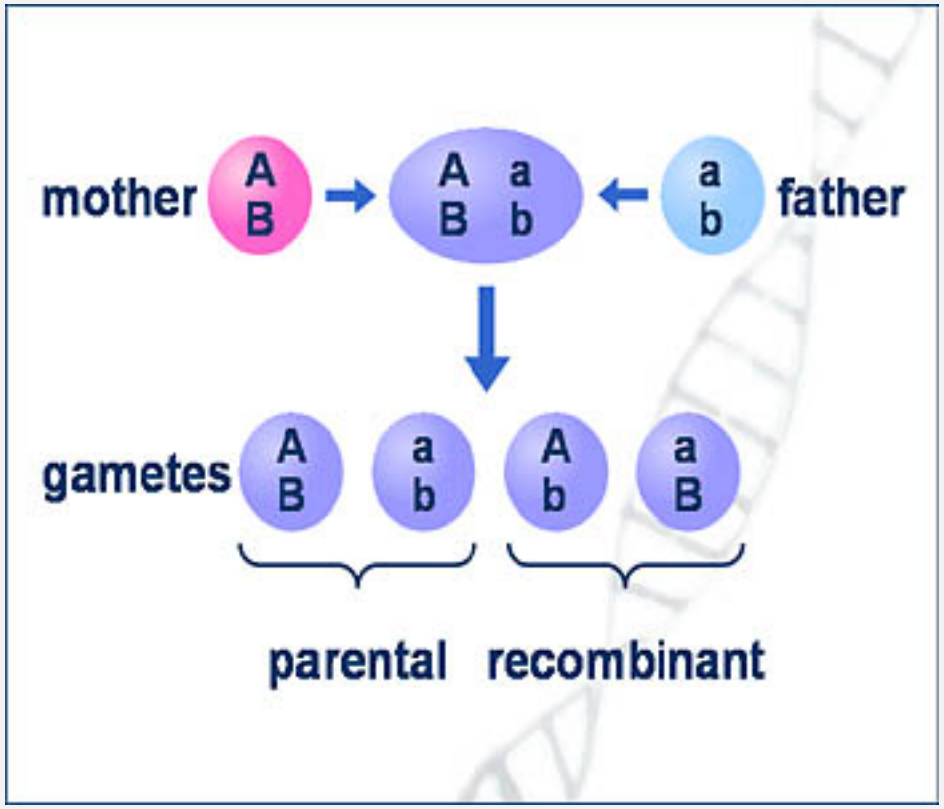
LINKAGE



John P. Hussman Institute for Human Genomics

- When the disease chromosomes are added, there is physical linkage between the marker locus and disease locus.
- "A" is segregating with the disease in this Autosomal Dominant family because it is near enough to the disease locus so that there has been no recombination.

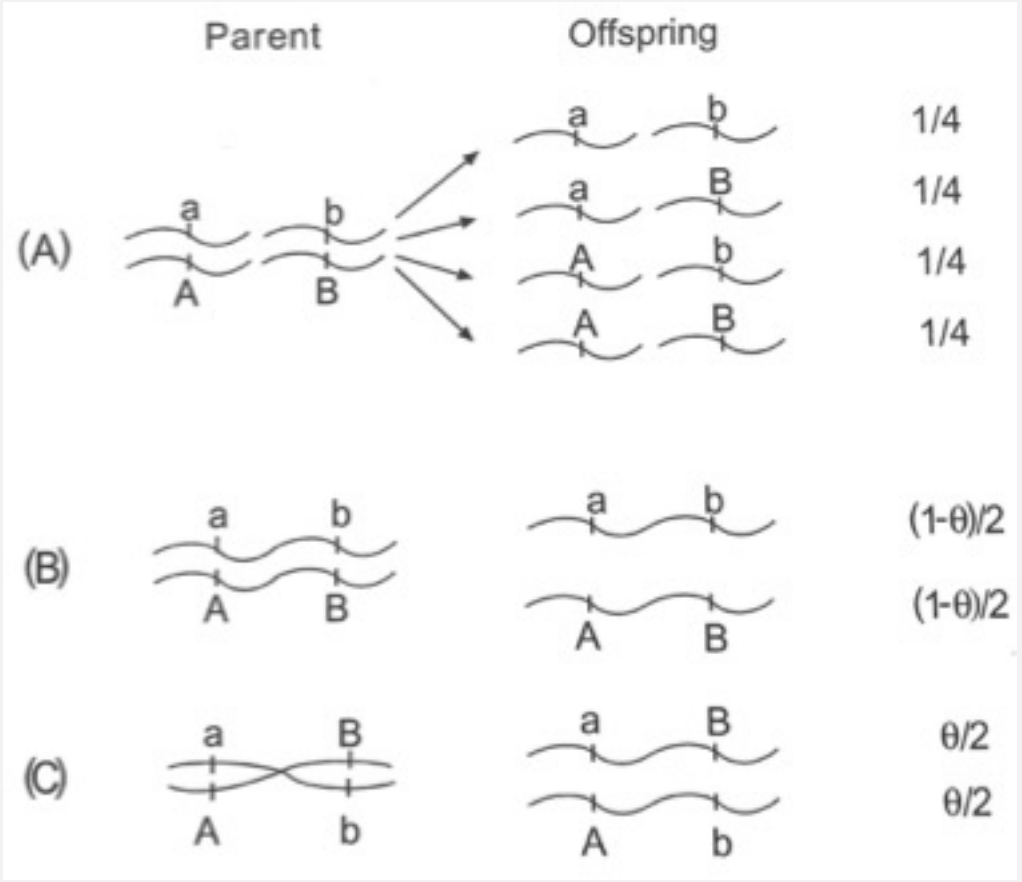
LINKAGE AND RECOMBINATION



- An individual carries two sets of chromosomes, one from the mother and one from the father. Imagine two loci, each with two alleles. When the individual forms gametes, there are 4 types of gametes that can be formed with respect to alleles at the two loci: 2 parental, 2 recombinant.
- Two loci are **linked** if the proportion of recombinant gametes is smaller than 1/2. When they are **unlinked** each type of gamete is formed with equal frequency.

John P. Hussman Institute for Human Genomics

RECOMBINATION FRACTION



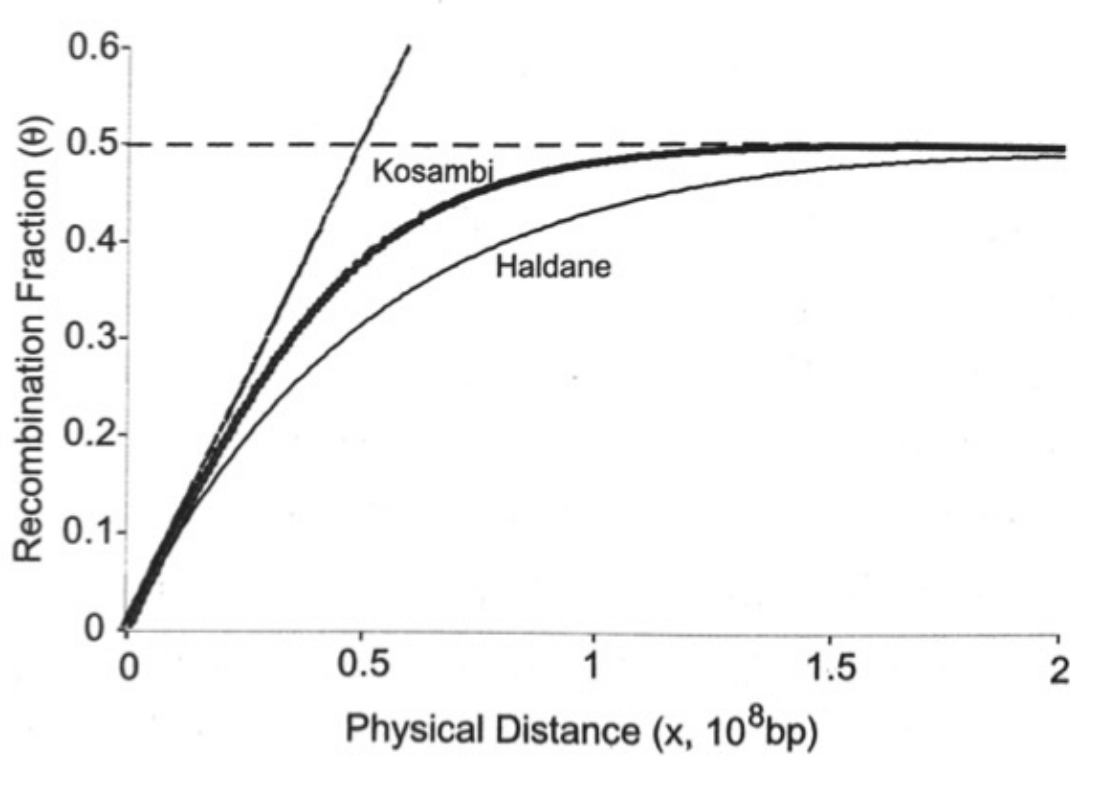
1-θ

θ

A. Two loci on different chromosomes segregate independently, each possibility having probability ¼
B. Two loci on the same chromosome segregating without recombination, with probability 1-θ or with recombination, with probability θ

Thomas 2004 Statistical Methods in Genetic Epidemiology

RECOMBINATION FRACTION TO PHYSICAL DISTANCE



- Map function showing the relationship between physical distance (x) in base pairs (bp) and recombination fraction (θ) in Morgans

Thomas 2004 Statistical Methods in Genetic Epidemiology

COMMON STUDY DESIGNS

Family data from

- Nuclear or extended families usually ascertained via an affected proband
- Relative pairs, e.g., affected sibpairs

EXAMPLE METHODS

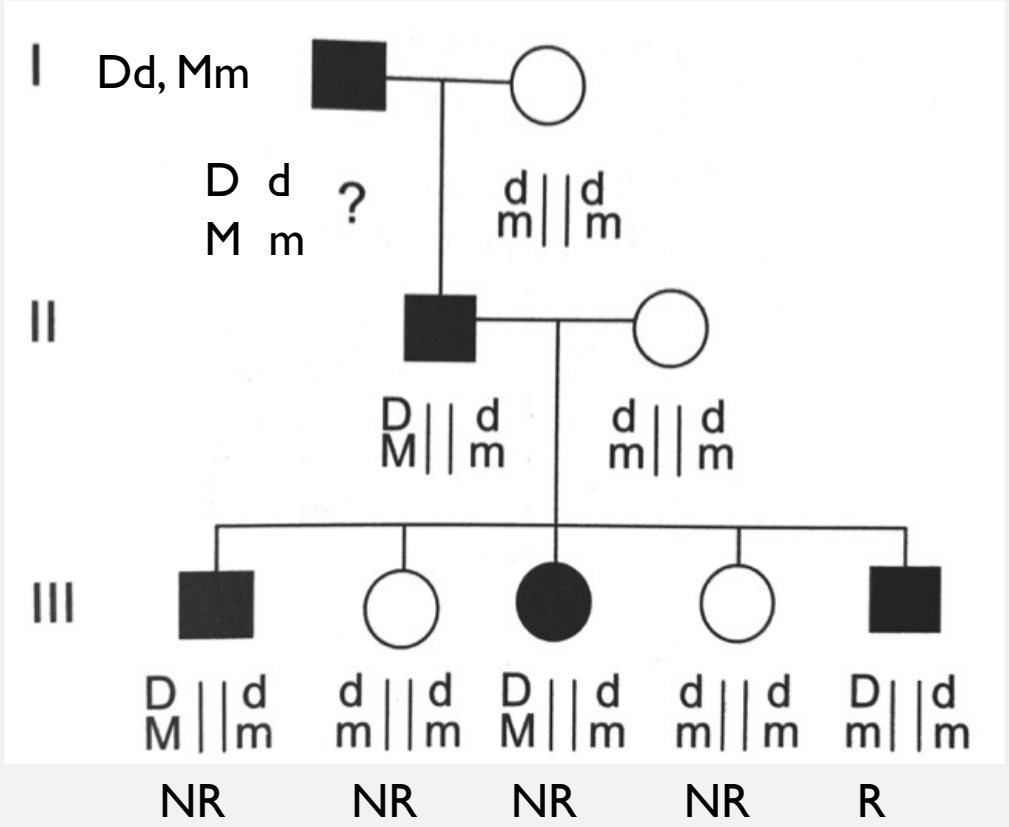
Binary traits

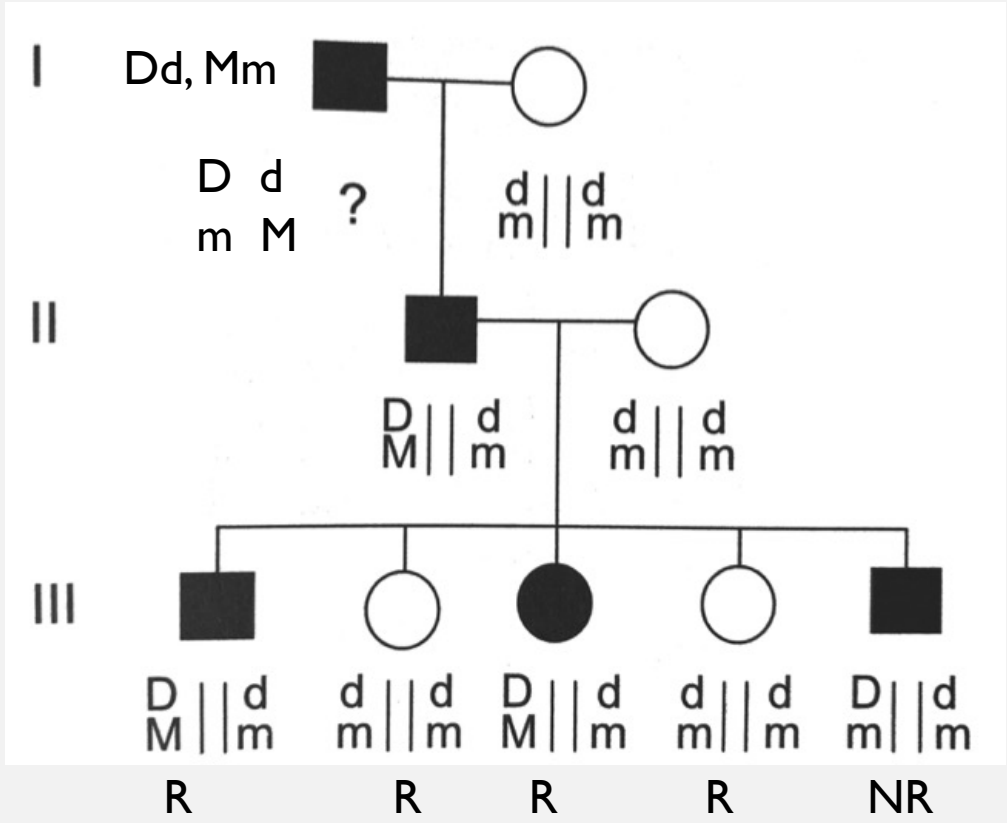
- LOD score method: likelihood based method
- Relative pair methods: observed shared genetic similarity vs. expected

Quantitative traits

- Haseman-Elston regression: phenotypic similarity linked to genetic similarity
- Variance components method

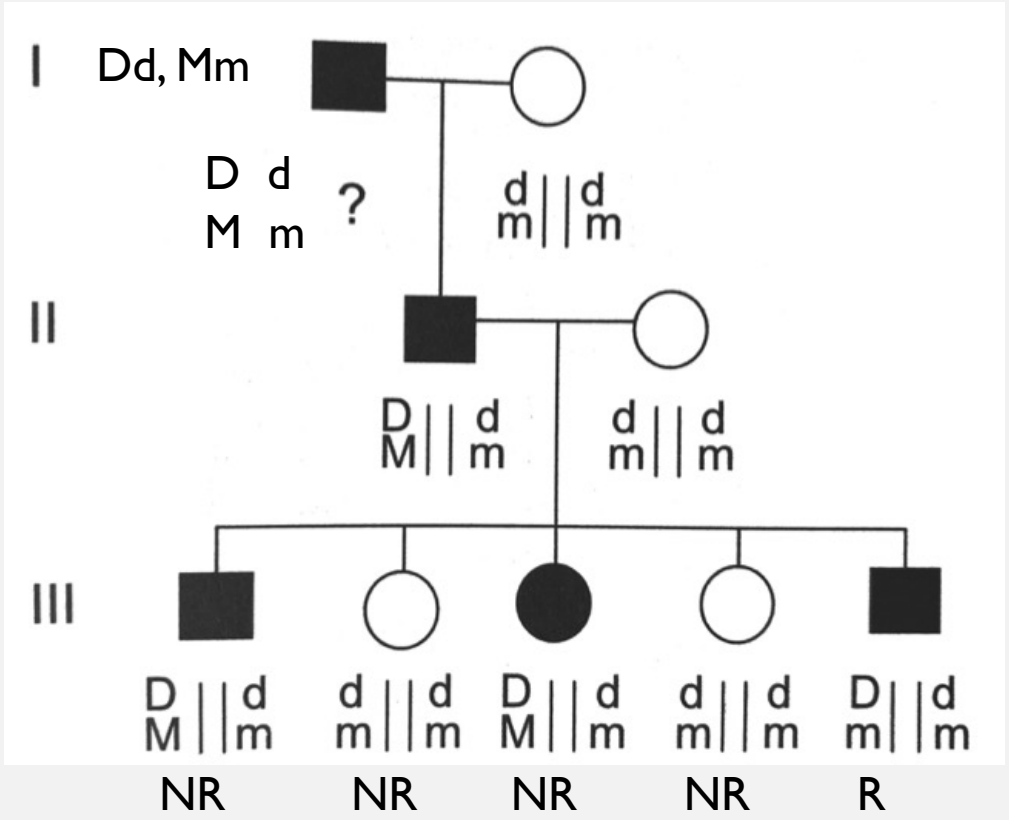
COUNTING RECOMBINANTS

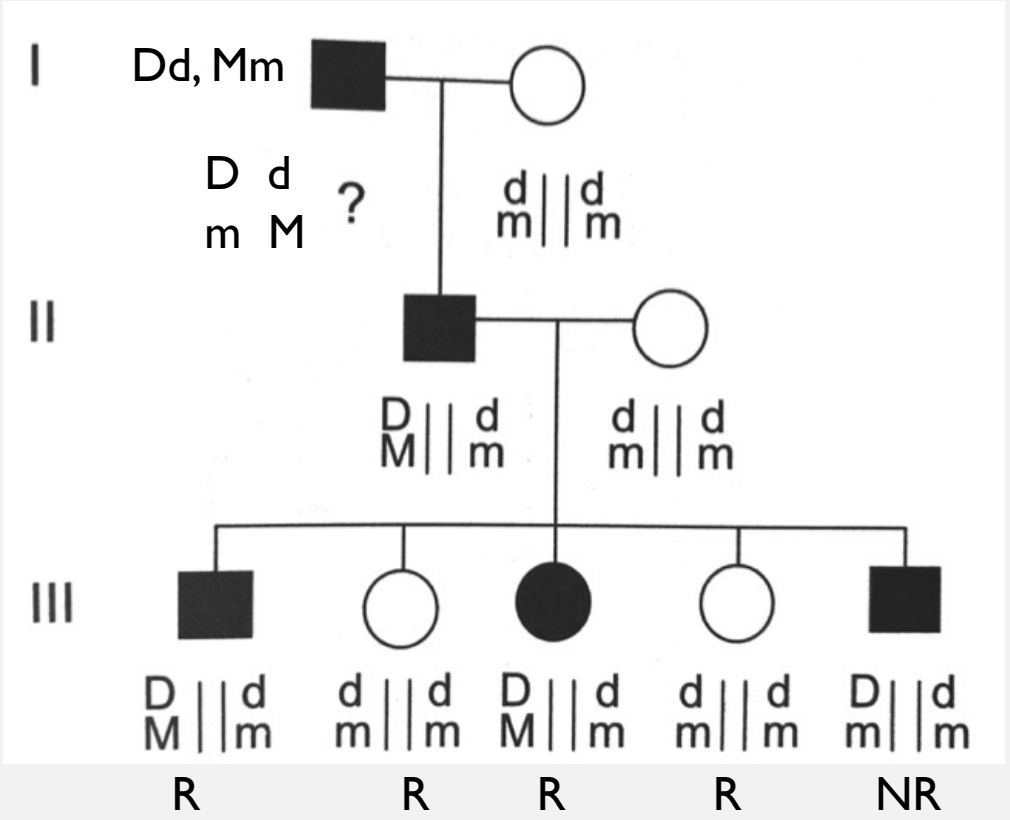




Thomas 2004 Statistical Methods in Genetic Epidemiology

COUNTING RECOMBINANTS

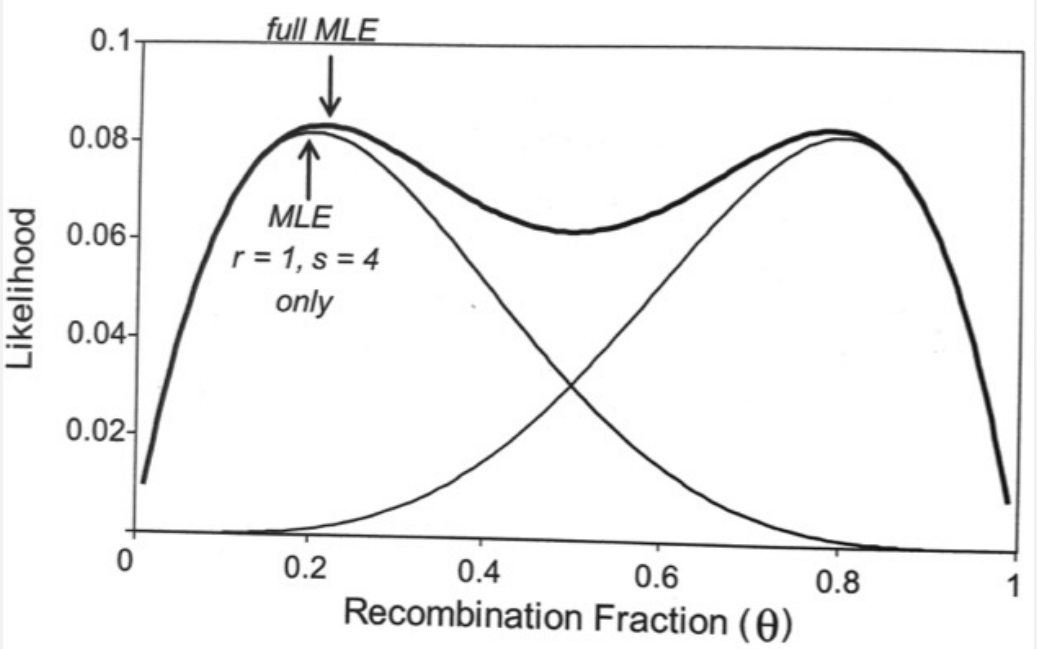




$$L(\theta) = 0.5\ L1 + 0.5\ L2 = (1-\theta)^4\theta + 0.5\theta^4(1-\theta)^1$$

LOGARITHM OF THE ODDS (LOD) SCORE

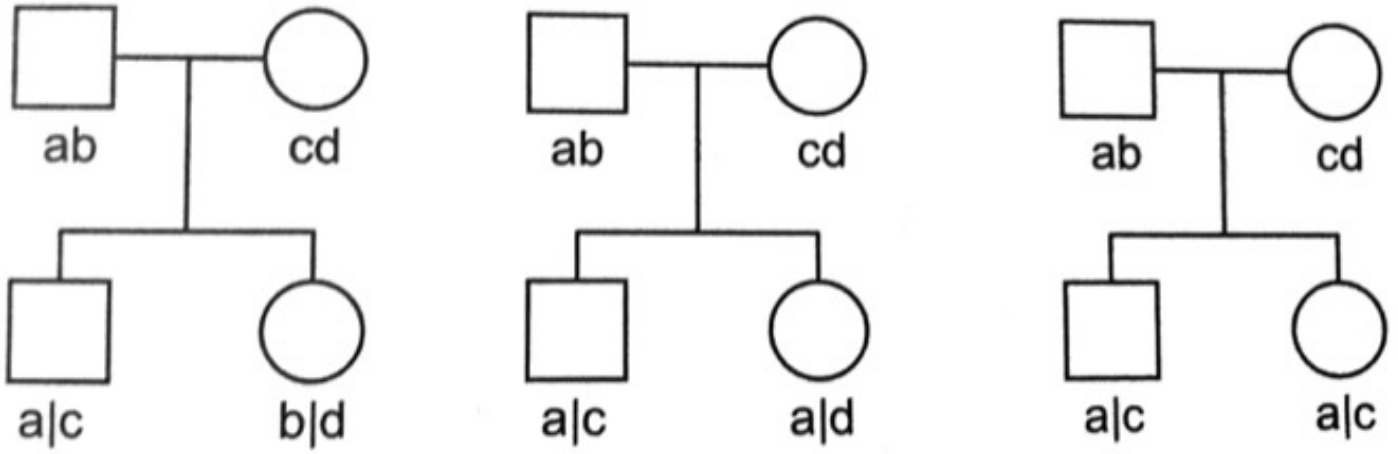
- LOD($\boldsymbol{\theta}$) = $\log_{10}$ [L($\hat{\theta}$)/L(0.5)], where $\boldsymbol{\theta}$=0.5 under the null hypothesis of no linkage between this marker and the disease locus



- By convention, a LOD score greater than 3 is considered evidence of linkage.
- On the other hand, a LOD score less than -2 is considered evidence to exclude linkage.

Thomas 2004 Statistical Methods in Genetic Epidemiology

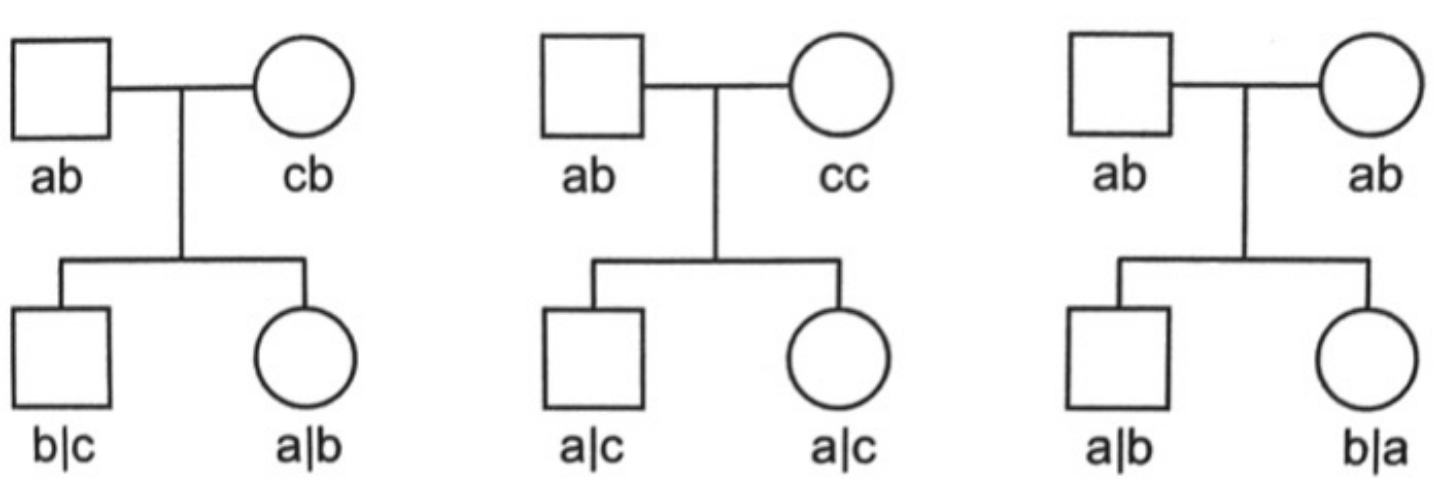
GENETIC SIMILARITY: IDENTICAL BY DESCENT



IBD=0 IBD=1 IBD=2

Thomas 2004 Statistical Methods in Genetic Epidemiology

GENETIC SIMILARITY: IDENTICAL BY DESCENT



IBD=0

IBD=1 or 2

IBD=0 or 2

Thomas 2004 Statistical Methods in Genetic Epidemiology

IBD DISTRIBUTION UNDER THE NULL: THETA = 0.5

Type of Relative Pair	*Probability of Sharing IBD Alleles*		
	π_0	π_1	π_2
Monozygotic twins	0	0	1
Full sibs	1/4	1/2	1/4
Parent-offspring	0	1	0
First cousins	3/4	1/4	0
Double first cousins	13/16	1/8	1/16
Grandparent-grandchild, half-sibs, avuncular	1/2	1/2	0

Full sibs:

Mean IBD = 0x0.25 + 1x(1/2) + 2x(1/4) = 1

Mean IBD proportion = ½ = 0.5

Thomas 2004 Statistical Methods in Genetic Epidemiology

EXAMPLE OF AFFECTED SIB PAIR TEST

	Number of Alleles Shared IBD			
	0	1	2	Total
Observed	20	45	35	100
Expected	25	50	25	100

IBD, identical by descent.

Full sibs: P(IBD=0) = 0.25, P(IBD=1) = 0.5, P(IBD=2) = 0.25

Thomas 2004 Statistical Methods in Genetic Epidemiology

EXAMPLE OF AFFECTED SIB PAIR TEST

$$H_0 : \bar{\pi} = 1/2.$$ (mean IBD proportion = ½)

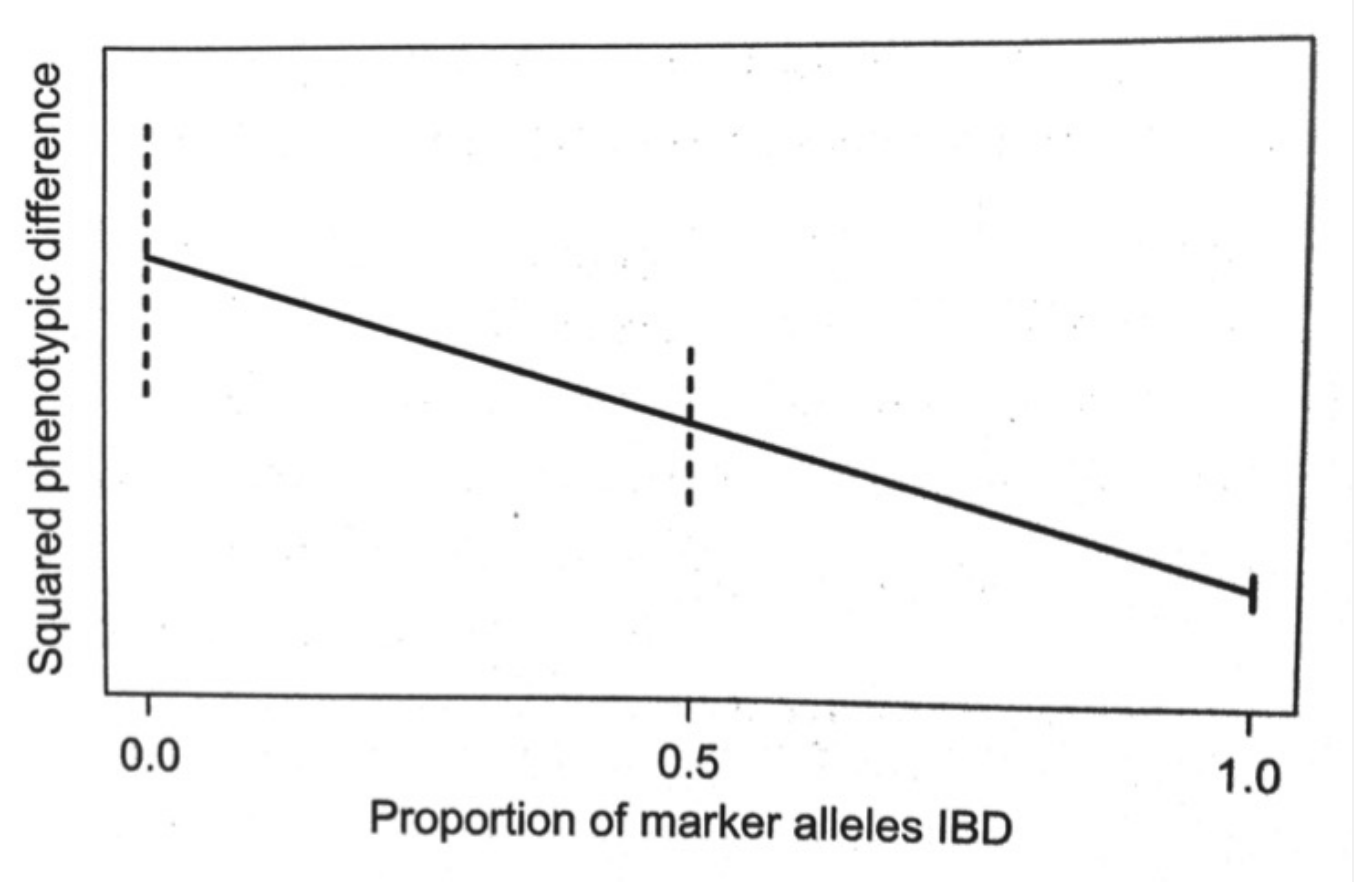
$$\frac{(20-25)^2}{25} + \frac{(45-50)^2}{50} + \frac{(35-25)^2}{25} = 5.50 \sim \chi_2^2$$

P-value = P(χ_2^2 > 5.50) > 0.05

Thomas 2004 Statistical Methods in Genetic Epidemiology

Ziegler and König 2006 A Statistical Approach to Genetic Epidemiology

HASEMAN-ELSTON REGRESSION



HASEMAN-ELSTON REGRESSION

$$E(y_i|IBD_t) = \alpha + \beta\tau_{t,i} + \gamma z_{t,i1} \qquad (8.4)$$

where IBD_t denotes the IBD information at the trait locus t, $\tau_{t,i}$ is the proportion of alleles shared IBD at the trait locus t by family i, and $z_{t;i1}$ denotes the probability that sib-pair i shares one allele IBD at the trait locus t. Furthermore, $\alpha = \sigma_\epsilon^2 + 2\sigma_g^2$, $\beta = -2\sigma_g^2$, and $\gamma = \sigma_d^2$.

Equation (8.4) shows that the squared trait difference at the trait locus is a linear function of τ_t and $z_{t,1}$. If there is no linkage between the genetic locus and the phenotype, $\sigma_g^2 = \sigma_d^2 = 0$. In this case, the regression of y_i on the τ is parallel to the x-axis with intercept σ_ϵ^2. However, if the marker and the trait locus are linked, the linear regression of y_i on $\tau_{t,i}$ has negative slope $\beta = -2\sigma_g^2$ (see Figure 8.1).

VARIANCE COMPONENTS METHODS

Phenotype of j-th individual in the i-th family

$$x_{ij} = \mu + g_{it} + G_{it} + \boldsymbol{\beta}'\boldsymbol{u_{it}} + e_{it}$$

assuming
x_i follows a multivariate normal distribution,
g (trait locus), G (random polygenic effect), and e (error term) are pairwise uncorrelated.

Ziegler and König 2006 A Statistical Approach to Genetic Epidemiology

VARIANCE COMPONENTS METHOD

$$\mathcal{L}(\sigma_a^2, \sigma_d^2, \sigma_G^2, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\ln \det(\boldsymbol{\Sigma}_i) + (\boldsymbol{x}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i (\boldsymbol{x}_i - \boldsymbol{\mu}_i) \right).$$

Maximization of the kernel of the log likelihood can be done with standard software, and linkage can be tested by LRTs. For example, the hypothesis that the variance components are different from 0 can be tested by comparing the estimate of the unrestricted $-2\,\mathcal{L}(\sigma_a^2, \sigma_d^2, \sigma_G^2, \theta)$ with a model where $\sigma_a^2 = 0$, i.e., with the estimate of $-2\,\mathcal{L}(\sigma_a^2|_{\sigma_a^2=0}, \sigma_d^2, \sigma_G^2, \theta)$, and the test statistic is asymptotically χ_1^2 distributed.

Ziegler and König 2006 A Statistical Approach to Genetic Epidemiology

GENETIC ASSOCIATION ANALYSIS

Chia-Ling Kuo

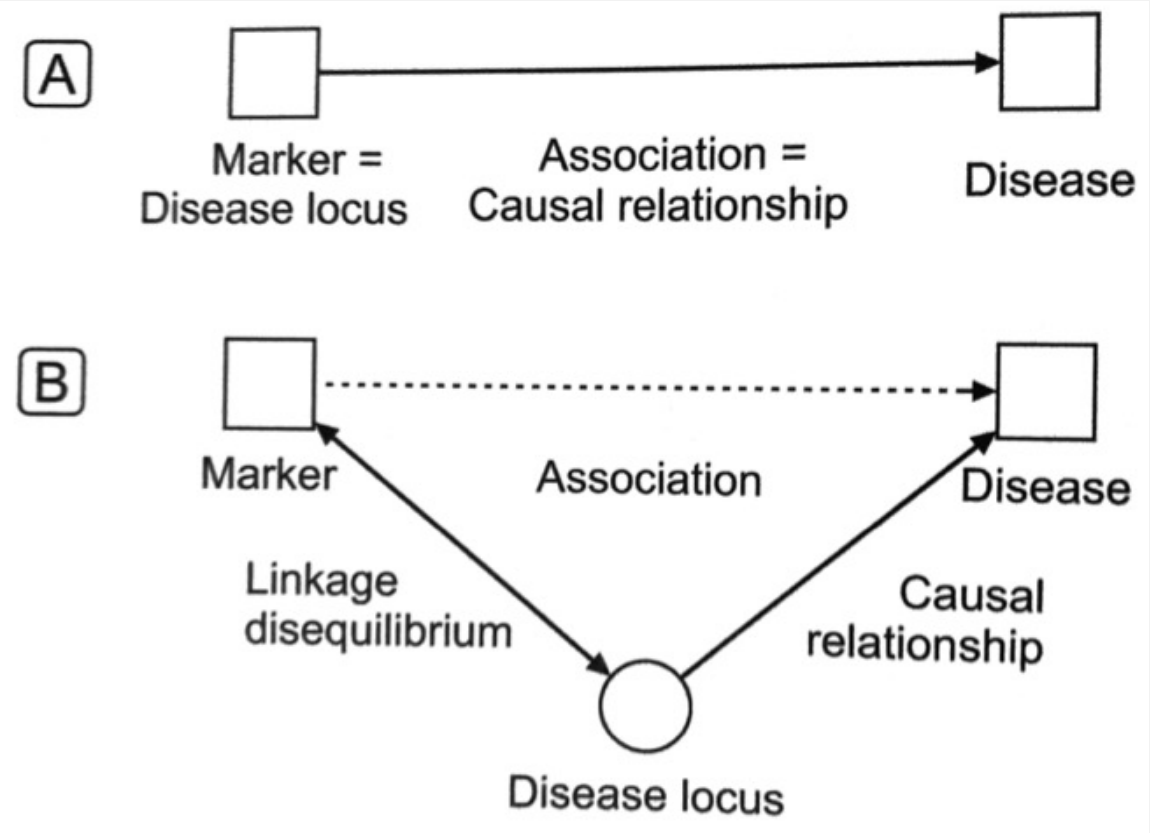
Associate Professor, Department of Public Health Sciences

University of Connecticut Health

LINKAGE AND ASSOCIATION

- Linkage looks at the transmission of a locus with a disease, whereas association focuses on the relation of an allele with a disease.
- Whereas linkage is based on transmission withing families, association is within populations.

CASUAL MODELS FOR GENETIC ASSOCIATION



KEY TO THE SUCCESS: LINKAGE DISEQUILIBRIUM

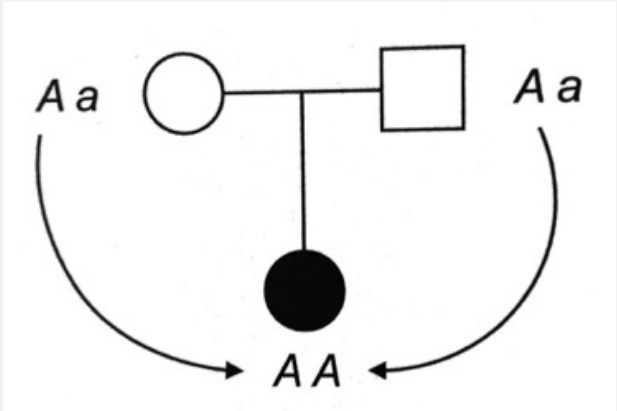
Marker	Disease Locus B		Disease Locus b		Total
A	p_{AB}	$= p_A p_B + \mathcal{D}$	p_{Ab}	$= p_A(1 - p_B) - \mathcal{D}$	p_A
a	p_{aB}	$(1-p_A)p_B - \mathcal{D}$	p_{ab}	$(1-p_A)(1-p_B) + \mathcal{D}$	$1-p_A$
Total		p_B		$1 - p_B$	1

$D = p_{AB} - p_A p_B$

AN ASSOCIATION TEST FOR TRIOS

Transmission Disequilibrium Test (TDT)

- Trio data: one affected child and his or her parents, regardless of their affection status
- Based on the genotypes of a trio, four parental alleles can be classified into transmitted alleles and non-transmitted alleles.
- Example: Aa (father) x Aa (mother) → AA (child)
  - Father: A (transmitted) a (non-transmitted)
  - Mother: A (transmitted) a (non-transmitted)



AN ASSOCIATION TEST FOR TRIOS

Transmitted allele	Non-transmitted allele A	a	Total
A	$n_{A,A}$	$n_{A,a}$	n_{TA}
a	$n_{a,A}$	$n_{a,a}$	n_{Ta}
Total	n_{NA}	n_{Na}	$2n$

1 from the father and 1 from the mother in that trio

AN ASSOCIATION TEST FOR TRIOS

$$H_0: \frac{p_{A,a}}{p_{A,a}+p_{a,A}} = \frac{1}{2} \quad \text{vs.} \quad H_1: \frac{p_{A,a}}{p_{A,a}+p_{a,A}} \neq \frac{1}{2}.$$

$$T_{\text{TDT}} = \frac{(n_{A,a}-n_{a,A})^2}{n_{A,a}+n_{a,A}}.$$

Essentially McNemar test statistic following the chi-square test with 1 degree of freedom under the null hypothesis

ASSOCIATION TESTS FOR CASE-CONTROL DATA

- Unrelated individual data mostly case-control data
- Genetic association analysis to associate genotypes and the disease status
- May or may not assume the mode of inheritance
- Statistically, models to consider
  - Case-control status ~ two indicators for AA and aA
  - Case-control status ~ genetic score (additive 0, 1, 2 for aa, Aa, and AA)
  - Case-control status ~ genetic score (dominant 0, 1, 1 for aa, Aa, and AA)
  - Case-control status ~ genetic score (recessive 0, 0, 1 for aa, Aa, and AA)

Data

Subject.	Diseased	Marker
1	Yes	AA
2	No	aa
3	No	Aa
4	Yes	Aa
5	Yes	AA
.	.	.
.	.	.
.	.	.

IN REALITY….

- Case-control status ~ genetic score **+ age + sex + genetic principal components + technical variables + …**
- Genotyped and imputed data
- Mixed family and unrelated individual data
- Significance evaluated at the genome-wide significance level ($P < 5 \times 10^{-8}$) regardless of the number of test
- ……covered by following lectures?!

THANK YOU FOR YOUR ATTENTION!

Chia-Ling Kuo, PhD

Associate Professor

Department of Public Health Sciences

University of Connecticut Health

My email: kuo@uchc.edu

Research key words: telomere length, APOE, UK Biobank, COVID-19, Mendelian randomization, biological age, aging