INTRODUCTION TO GENETIC EPIDEMIOLOGY

(EPID0754)

Prof. Dr. Dr. K. Van Steen

DIFFERENT FACES OF GENETIC EPIDEMIOLOGY

1 Basic epidemiology

1.a Aims of epidemiology

1.b Designs in epidemiology

1.c An overview of measurements in epidemiology

2 Genetic epidemiology

2.a What is genetic epidemiology?

2.b Designs in genetic epidemiology

2.c Study types in genetic epidemiology

3 Phenotypic aggregation within families

3.a Introduction to familial aggregation?

3.b Familial aggregation with quantitative traits

IBD and kinship coefficient

3.c Familial aggregation with dichotomous traits

Relative recurrence risk

3.d Twin studies

4 Segregation analysis

4.a What is segregation analysis?

Modes of inheritance

4.b Classical method for sibships and one locus

Segregation ratios

4.c Likelihood method for pedigrees and one locus

Elston-Stewart algorithm

4.d Variance component modeling: a general framework

Decomposition of variability, major gene, polygenic and mixed models

4.e The ideas of variance component modeling adjusted for binary traits

Liability threshold models

4.f Quantifying the genetic importance of familial resemblance

Heritability

5 Genetic epidemiology and public health

1 Basic epidemiology

Main references:

- Burton P, Tobin M and Hopper J. Key concepts in genetic epidemiology. The Lancet, 2005

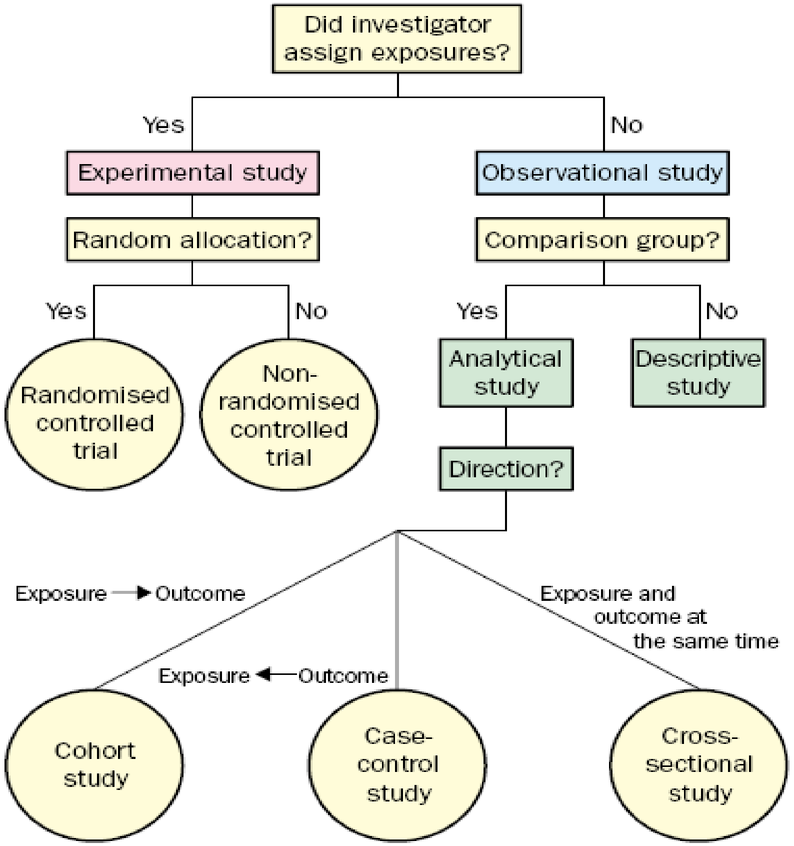
- Clayton D. Introduction to genetics (course slides Bristol 2003)
- Bonita R, Beaglehole R and Kjellström T. *Basic Epidemiology*. WHO 2nd edition
- URL:
  - http://www.dorak.info/

1.a Aims of epidemiology

- Epidemiology originates from Hippocrates' observation more than 2000 years ago that environmental factors influence the occurrence of disease. However, it was not until the nineteenth century that the distribution of disease in specific human population groups was measured to any large extent. This work marked not only the formal beginnings of epidemiology but also some of its most spectacular achievements.
- Epidemiology in its modern form is a relatively new discipline and uses quantitative methods to study diseases in human populations, to inform prevention and control efforts.

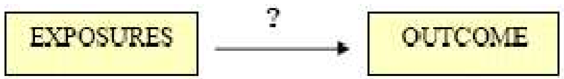
1.b Designs in epidemiology

- A focus of an epidemiological study is the population defined in geographical or other terms

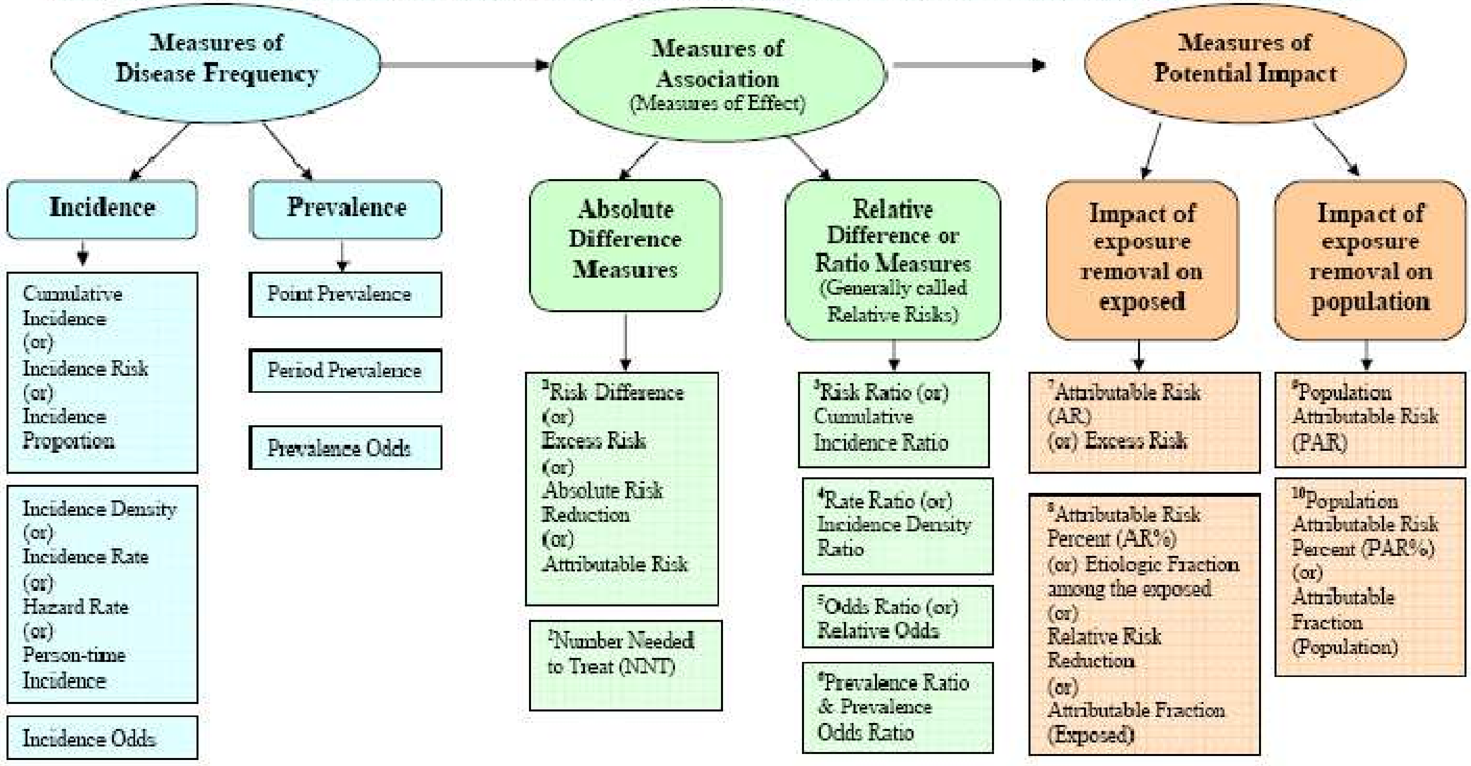


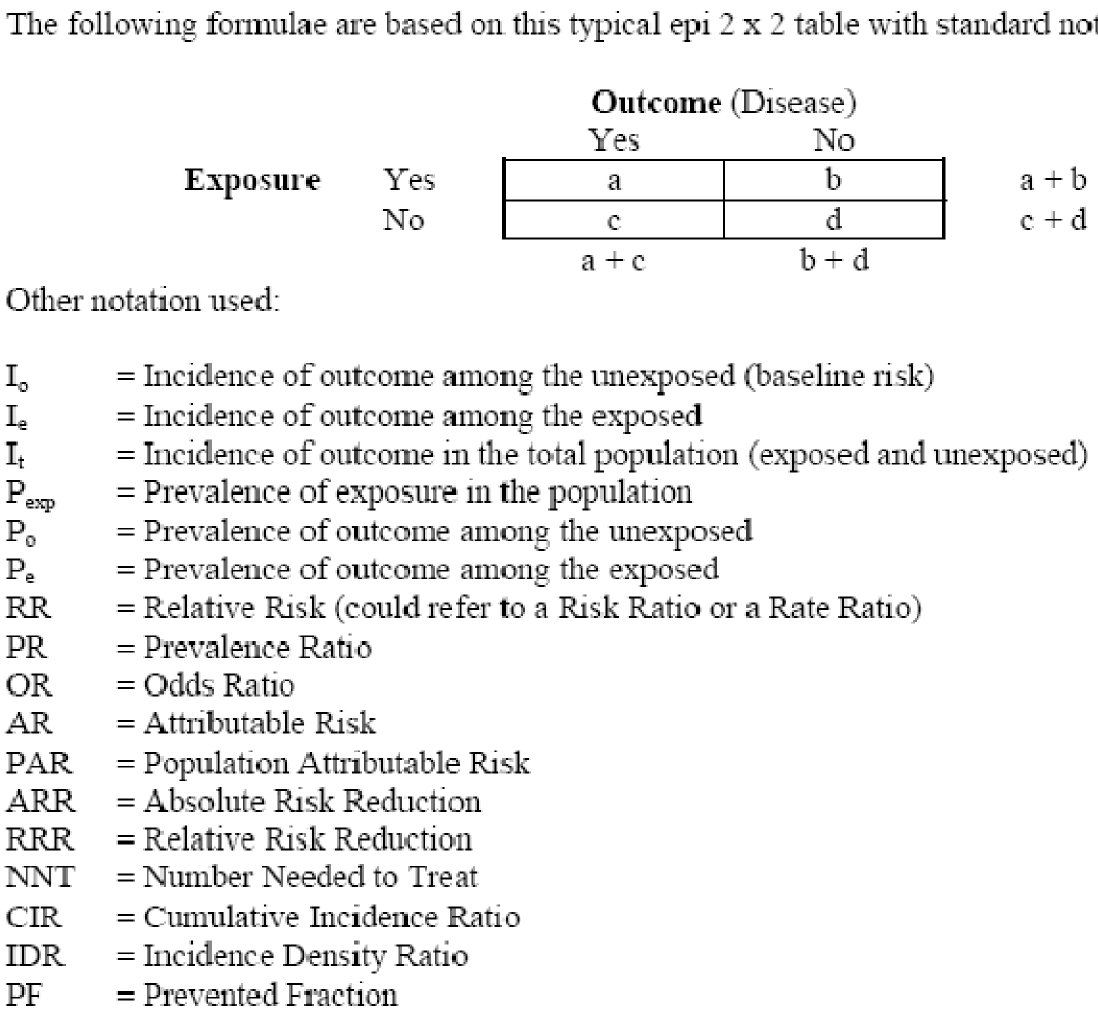
(Grimes & Schulz 2002)

1.c An overview of measurements in epidemiology



Epidemiology is about identifying associations between exposures and outcomes. To identify any association, exposures and outcomes must first be measured in a quantitative manner. Then rates of occurrence of events are computed. These measures are called "*measures of disease frequency*." Once measured, the association between exposures and outcomes are then evaluated by calculating "*measures of association or effect*." Finally, the impact of removal of an exposure on the outcome is evaluated by computing "*measures of potential impact*." In general, measures of disease frequency are needed to generate measures of association, and both these are needed to get measures of impact. There is some overlap between these measures, and terminology is poorly standardized.


Measures of Disease Frequency
Measures of Association (Measures of Effect)
Measures of Potential Impact
Incidence
Prevalence
Absolute Difference Measures
Relative Difference or Ratio Measures (Generally called Relative Risks)
Impact of exposure removal on exposed
Impact of exposure removal on population
Cumulative Incidence (or) Incidence Risk (or) Incidence Proportion
Incidence Density (or) Incidence Rate (or) Hazard Rate (or) Person-time Incidence
Incidence Odds
Point Prevalence
Period Prevalence
Prevalence Odds
[1]Risk Difference (or) Excess Risk (or) Absolute Risk Reduction (or) Attributable Risk
[2]Number Needed to Treat (NNT)
[3]Risk Ratio (or) Cumulative Incidence Ratio
[4]Rate Ratio (or) Incidence Density Ratio
[5]Odds Ratio (or) Relative Odds
[6]Prevalence Ratio & Prevalence Odds Ratio
[7]Attributable Risk (AR) (or) Excess Risk
[8]Attributable Risk Percent (AR%) (or) Etiologic Fraction among the exposed (or) Relative Risk Reduction (or) Attributable Fraction (Exposed)
[9]Population Attributable Risk (PAR)
[10]Population Attributable Risk Percent (PAR%) (or) Attributable Fraction (Population)

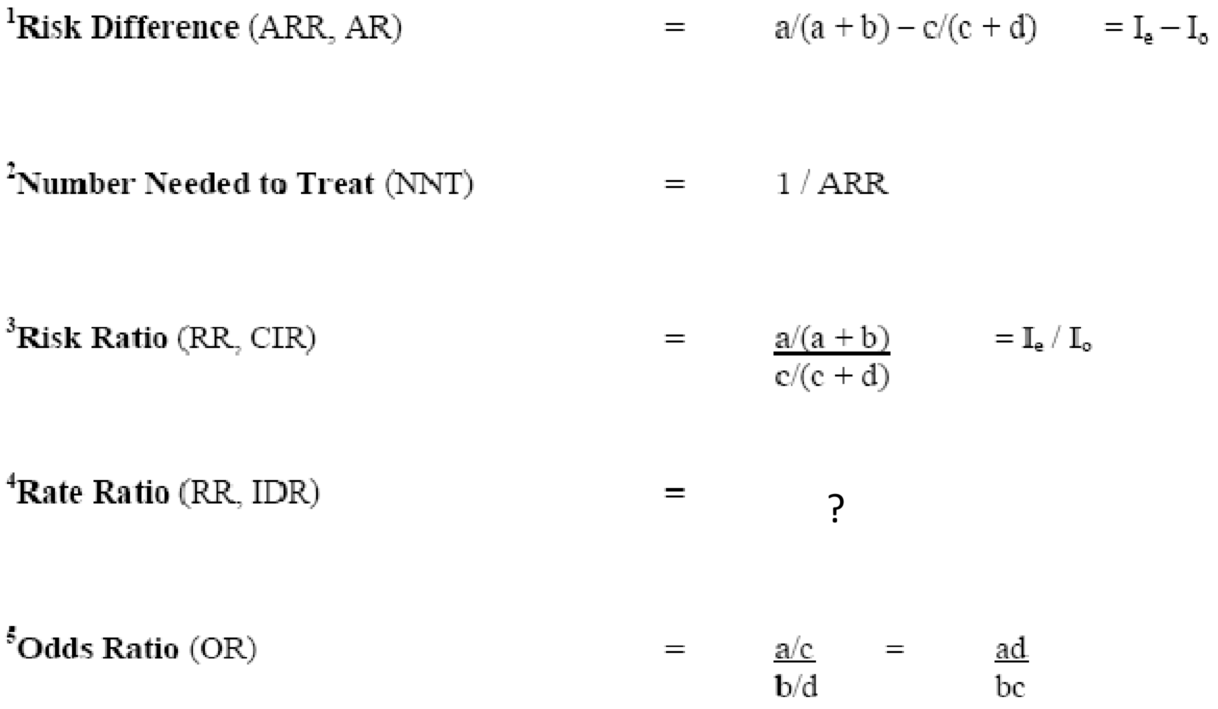


The following formulae are based on this typical epi 2 x 2 table with standard notation:

		Outcome (Disease)		
		Yes	No	
Exposure	Yes	a	b	a + b
	No	c	d	c + d
		a + c	b + d	

Other notation used:

I_o = Incidence of outcome among the unexposed (baseline risk)
I_e = Incidence of outcome among the exposed
I_t = Incidence of outcome in the total population (exposed and unexposed)
P_{exp} = Prevalence of exposure in the population
P_o = Prevalence of outcome among the unexposed
P_e = Prevalence of outcome among the exposed
RR = Relative Risk (could refer to a Risk Ratio or a Rate Ratio)
PR = Prevalence Ratio
OR = Odds Ratio
AR = Attributable Risk
PAR = Population Attributable Risk
ARR = Absolute Risk Reduction
RRR = Relative Risk Reduction
NNT = Number Needed to Treat
CIR = Cumulative Incidence Ratio
IDR = Incidence Density Ratio
PF = Prevented Fraction



[1]**Risk Difference** (ARR, AR) = $a/(a + b) - c/(c + d) = I_e - I_o$

[2]**Number Needed to Treat** (NNT) = $1 / ARR$

[3]**Risk Ratio** (RR, CIR) = $\dfrac{a/(a + b)}{c/(c + d)} = I_e / I_o$

[4]**Rate Ratio** (RR, IDR) = ?

[5]**Odds Ratio** (OR) = $\dfrac{a/c}{b/d} = \dfrac{ad}{bc}$

[6]**Prevalence Ratio** (PR) = P_e / P_o

[7]**Attributable Risk** (AR) = Same formula as Risk Difference

[8]**Attributable Risk Percent** (AR%)♣ = $\frac{I_e - I_o}{I_e} * 100 = \frac{AR}{I_e} * 100$

$= \frac{a/(a+b) - c/(c+d)}{a/(a+b)}$

Alternative formula for AR% = $\frac{(RR-1)}{RR} * 100$

AR% in a case-control study = $\frac{(OR-1)}{OR} * 100$

[9]**Population Attributable Risk** (PAR) = $I_t - I_o$

Alternative formula for PAR = $AR * P_{exp}$

[10]**Population Attributable Risk Percent** (PAR%) = $\frac{I_t - I_o}{I_t} * 100$

Alternative formula for PAR% = $\frac{P_{exp}(RR-1)}{P_{exp}(RR-1)+1} * 100$

[4]**Rate Ratio** (RR, IDR) $= \dfrac{a/N1}{b/N2}$

This formula for Rate Ratio is based on the following 2 x 2 table format:

	Cases (Outcome)	Person-time
Exposed	a	N1
Unexposed	b	N2

[♣] Note: In some situations (like in a clinical trial or a vaccine field study), the exposure is protective. In such situations, some of the above formulae will have to be computed and interpreted differently. Also, the names will change. For example, when the exposure is protective, the AR% is meaningless because the I_e is less than I_o, because exposure leads to a lower incidence. In this situation, the formula changes to:

$$\text{RRR (PF)} = \frac{I_o - I_e}{I_o} * 100$$

The name changes from AR% (also called *etiologic fraction)*, to Relative Risk Reduction [RRR] (also called *prevented fraction*). Similarly, when the exposure is protective, AR (also called Excess Risk) becomes Absolute Risk Reduction (ARR) and the formula becomes: ARR = $I_o - I_e$

2 Genetic epidemiology

Main references:

- Clayton D. Introduction to genetics (course slides Bristol 2003)
- Ziegler A. Genetic epidemiology present and future (presentation slides)
- URL:
    - http://www.dorak.info/

2.a What is genetic epidemiology?

Definitions

- Term firstly used by Morton & Chung (1978)

- Genetic epidemiology is a science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations . (Morton, 1982).

- Genetic epidemiology is the study of how and why diseases cluster in families and ethnic groups (King et al., 1984)

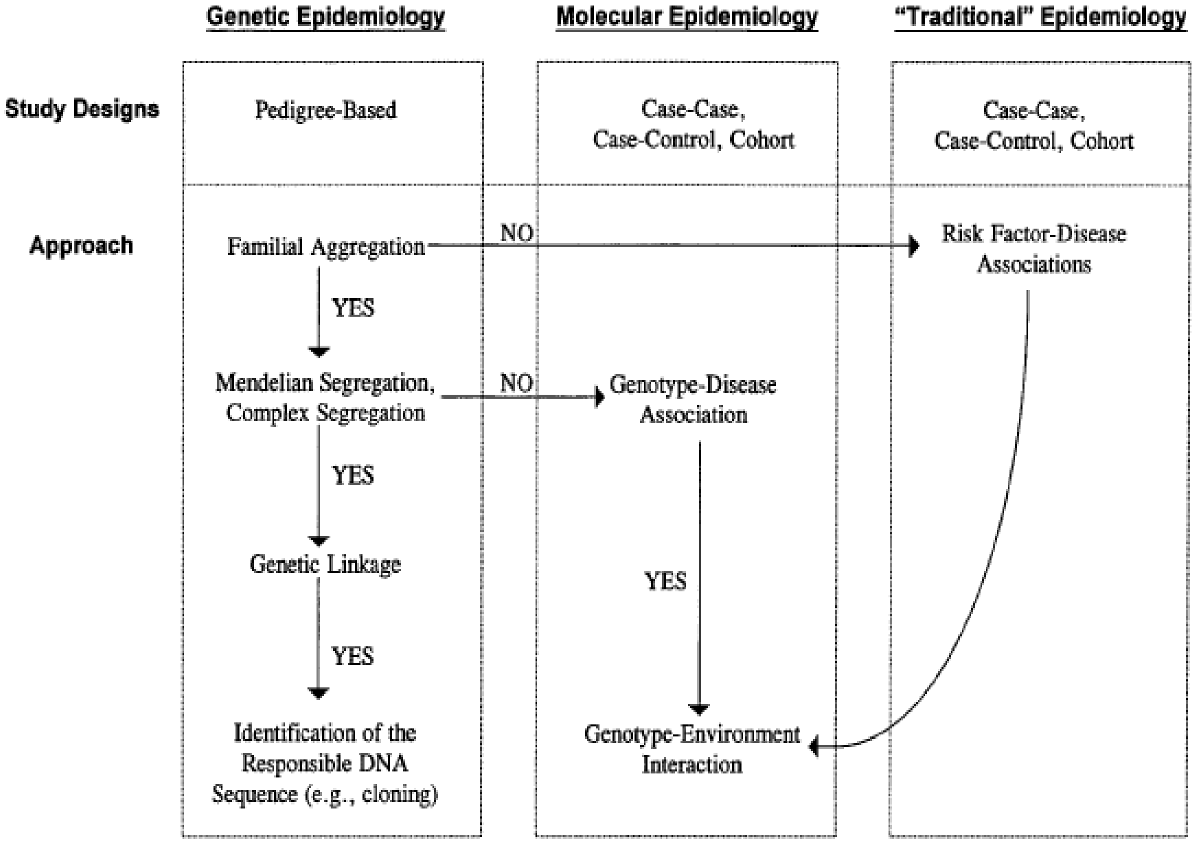
- Genetic epidemiology examines the role of genetic factors, along with the environmental contributors to disease, and at the same time giving equal attention to the differential impact of environmental agents, non-familial as well as familial, on different genetic backgrounds (Cohen, Am J Epidemiol, 1980)

Aim of genetic epidemiology

to detect the inheritance pattern of a particular disease,
to localize the gene and
to find a marker associated with disease susceptibility

(Photo: J. Murken via A Ziegler)

X – epidemiology

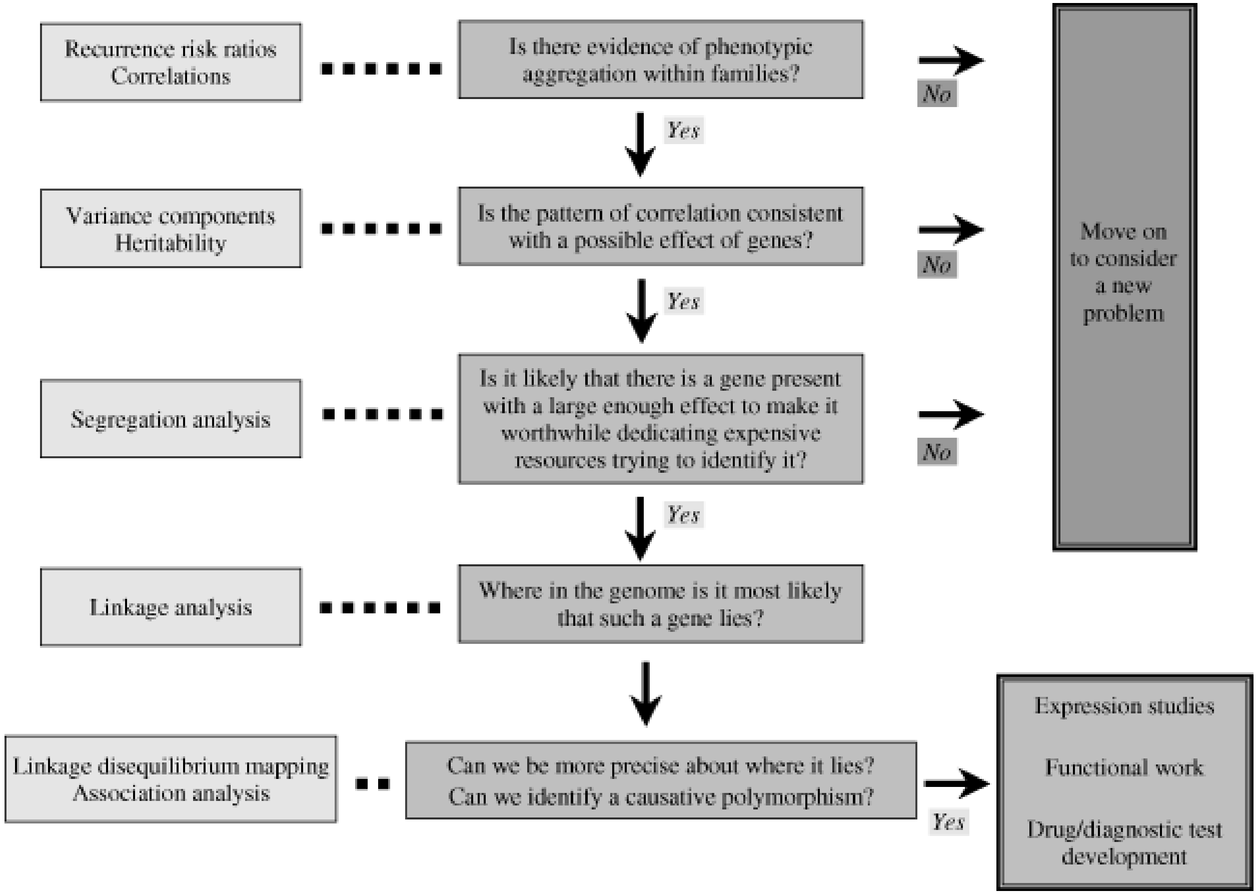

Genetic Epidemiology
Molecular Epidemiology
"Traditional" Epidemiology
Study Designs
Pedigree-Based
Case-Case, Case-Control, Cohort
Case-Case, Case-Control, Cohort
Approach
Familial Aggregation
NO
Risk Factor-Disease Associations
YES
Mendelian Segregation, Complex Segregation
NO
Genotype-Disease Association
YES
Genetic Linkage
YES
YES
Identification of the Responsible DNA Sequence (e.g., cloning)
Genotype-Environment Interaction

(Rebbeck TR, *Cancer*, 1999)

X-epidemiology

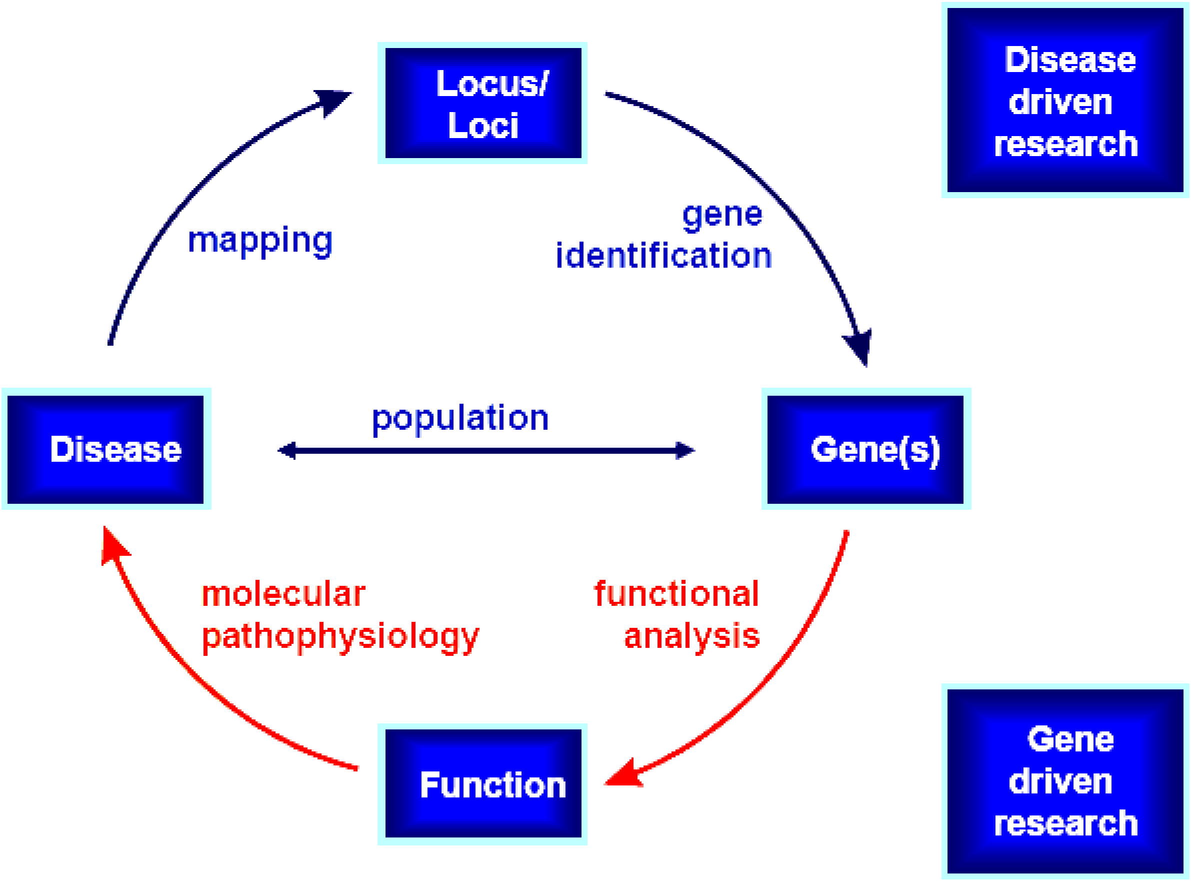
- In contrast to classic epidemiology, the three main complications in genetic epidemiology are
  - dependencies,
  - use of indirect evidence and
  - complex data sets
- Genetic epidemiology is highly dependent on the direct incorporation of family structure and biology. The structure of families and chromosomes leads to major dependencies between the data and thus to customized models and tests. In many studies only indirect evidence can be used, since the disease-related gene, or more precisely the functionally relevant DNA variant of a gene, is not directly observable. In addition, the data sets to be analyzed can be very complex.

Relevant questions in genetic epidemiology

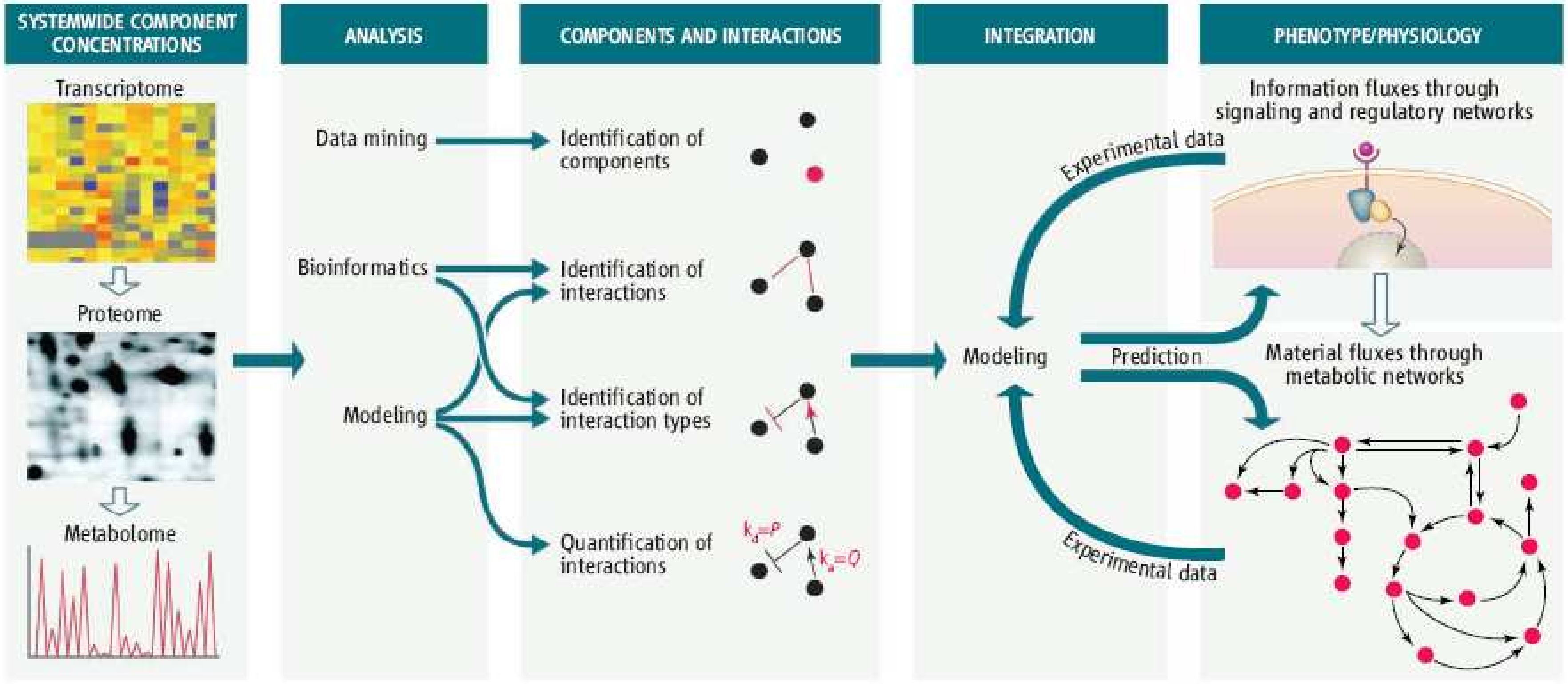

Recurrence risk ratios
Correlations
Is there evidence of phenotypic aggregation within families?
No
Yes
Variance components
Heritability
Is the pattern of correlation consistent with a possible effect of genes?
No
Yes
Segregation analysis
Is it likely that there is a gene present with a large enough effect to make it worthwhile dedicating expensive resources trying to identify it?
No
Yes
Move on to consider a new problem
Linkage analysis
Where in the genome is it most likely that such a gene lies?
Linkage disequilibrium mapping
Association analysis
Can we be more precise about where it lies?
Can we identify a causative polymorphism?
Yes
Expression studies
Functional work
Drug/diagnostic test development

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

Genetic research paradigm



Getting closer to the whole picture


SYSTEMWIDE COMPONENT CONCENTRATIONS
ANALYSIS
COMPONENTS AND INTERACTIONS
INTEGRATION
PHENOTYPE/PHYSIOLOGY
Transcriptome
Proteome
Metabolome
Data mining
Bioinformatics
Modeling
Identification of components
Identification of interactions
Identification of interaction types
Quantification of interactions
$k_d=P$
$k_a=Q$
Experimental data
Modeling
Prediction
Experimental data
Information fluxes through signaling and regulatory networks
Material fluxes through metabolic networks

(Sauer et al, *Science*, 2007)

Recent success stories of genetics and genetic epidemiology research

- Gene expression profiling to assess prognosis and guide therapy, e.g. breast cancer
- Genotyping for stratification of patients according to risk of disease, e.g. myocardial infarction
- Genotyping to elucidate drug response, e.g. antiepileptic agents
- Designing and implementing new drug therapies, e.g. imatinib for hypereosinophilic syndrome
- Functional understanding of disease causing genes, e.g. obesity

(Guttmacher & Collins, N Engl J Med, 2003)

Flow of research

Disease characteristics:	Descriptive epidemiology
Familial clustering:	Family aggregation studies
Genetic or environmental:	Twin/adoption/half-sibling/migrant studies
Mode of inheritance:	Segregation analysis
Disease susceptibility loci:	Linkage analysis
Disease susceptibility markers:	Association studies

2.b Designs in genetic epidemiology

The samples needed for genetic epidemiology studies may be

- nuclear families (index case and parents),
- affected relative pairs (sibs, cousins, any two members of the family),
- extended pedigrees,
- twins (monozygotic and dizygotic) or
- unrelated population samples.

2.c Study types in genetic epidemiology

Main methods in genetic epidemiology

- **Genetic risk studies**:
  - What is the contribution of genetics as opposed to environment to the trait? Requires family-based, twin/adoption or migrant studies.
- **Segregation analyses**:
  - What does the genetic component look like (*oligogenic* 'few genes each with a moderate effect', *polygenic* 'many genes each with a small effect', etc)?
  - What is the model of transmission of the genetic trait? Segregation analysis requires multigeneration family trees preferably with more than one affected member.

- **Linkage studies**:
  - What is the location of the disease gene(s)? Linkage studies screen the whole genome and use parametric or nonparametric methods such as allele sharing methods {affected sibling-pairs method} with no assumptions on the mode of inheritance, penetrance or disease allele frequency (the parameters). The underlying principle of linkage studies is the cosegregation of two genes (one of which is the disease locus).
- **Association studies**:
  - What is the allele associated with the disease susceptibility? The principle is the coexistence of the same marker on the same chromosome in affected individuals (due to linkage disequilibrium). Association studies may be family-based (TDT) or population-based. Alleles or haplotypes may be used. Genome-wide association studies (GWAS) are increasing in popularity.

3 Familial aggregation of a phenotype

Main references:

- Burton P, Tobin M and Hopper J. Key concepts in genetic epidemiology. *The Lancet*, 2005
- Thomas D. Statistical methods in genetic epidemiology. Oxford University Press 2004
- Laird N and Cuenco KT. Regression methods for assessing familial aggregation of disease. *Stats in Med* 2003

- Clayton D. Introduction to genetics (course slides Bristol 2003)
- URL:
  - http://www.dorak.info/

3.a Introduction to familial aggregation

What is familial aggregations?

- Consensus on a precise definition of familial aggregation is lacking
- The heuristic interpretation is that aggregation exists when cases of disease appear in families more often than one would expect if diseased cases were spread uniformly and randomly over individuals.
- The assessment of familial aggregation of disease is often regarded as the initial step in determining whether or not there is a genetic basis for disease.
- Absence of any evidence for familial aggregation casts strong doubt on a genetic component influencing disease, especially when environmental factors are included in the analysis.

What is familial aggregation? (continued)

- Actual approaches for detecting aggregation depend on the nature of the phenotype, but the common factor in existing approaches is that they are taken without any specific genetic model in mind.
- The basic design of familial aggregation studies typically involves sampling families
- In most places there is no natural sampling frame for families, so individuals are selected in some way and then their family members are identified. The individual who caused the family to be identified is called the *proband.*

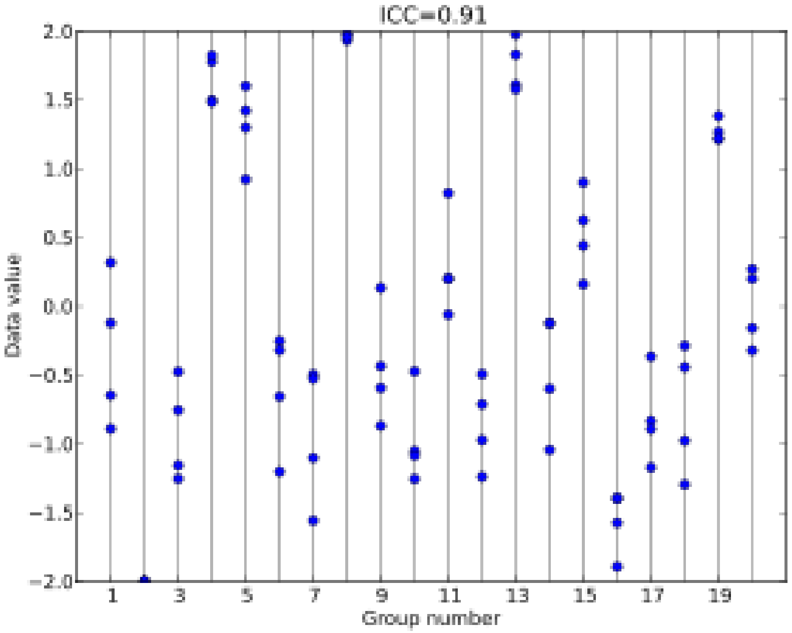
3.b Familial aggregation with quantitative traits

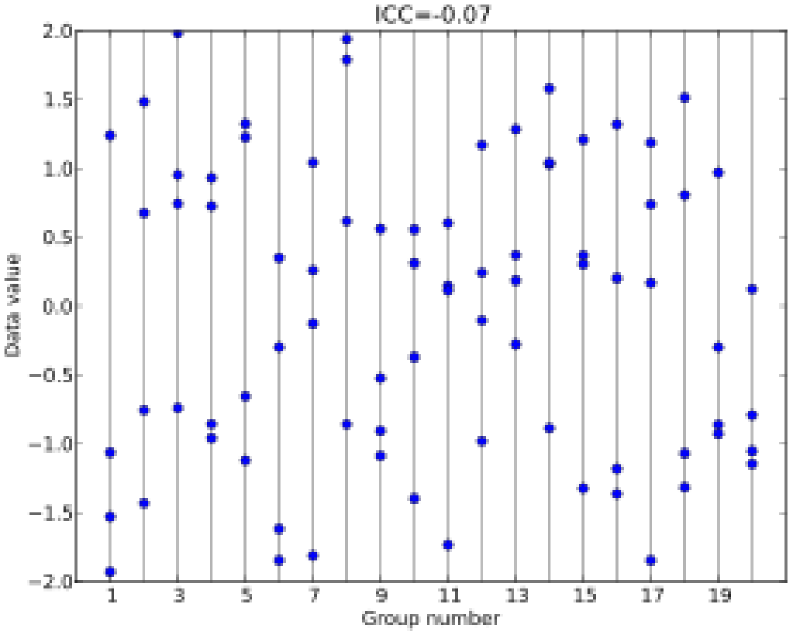
Proband selection

- For a continuous trait a random series of probands from the general population may be enrolled, together with their family members.

Correlations between trait values among family members

- For quantitative traits, such as blood pressure, familial aggregation can be assessed using a correlation or covariance-based measure
- For instance, the so-called *intra-family correlation coefficient* (ICC)
    - ICC can be interpreted as the proportion of the total variability in a phenotype that can reasonably be attributed to real variability between families
    - Techniques such as linear regression and mulitilevel modelling analysis of variance are useful to derive estimates
    - Non-random ascertainment can seriously bias an ICC.
- Alternatively, *familial correlation coefficients* are computed as in the programme FCOR within the Statistical Analysis for Genetic Epidemiology (SAGE) software package





(http://en.wikipedia.org/wiki/Intraclass_correlation)

3.c Familial aggregation with dichotomous traits

Proband selection

- It is a misconception that probands always need to have the disease of interest.
- In general, the sampling procedure based on proband selection closely resembles the case-control sampling design, for which exposure is assessed by obtaining data on disease status of relatives, usually first-degree relatives, of the probands. This selection procedure is particularly practical when disease is relatively rare.

Two main streams in analysis

- In a retrospective type of analysis, the outcome of interest is disease in the proband. Disease in the relatives serves to define the exposure.
- Recent literature focuses on a prospective type of analysis, in which disease status of the relatives is considered the outcome of interest and is conditioned on disease status in the proband.

Recurrence risks

- One parameter often used in the genetics literature to indicate the strength of a gene effect is the familial risk ratio λ_R, where

  $$\lambda_R = \lambda / K \,,$$

  K the disease prevalence in the population and λ the probability that an individual has disease, given that a relative also has the disease.
- The risk in relatives of type R of diseased *probands* is termed relative *recurrence risk* λ_R and is usually expressed versus the population risk as above.
- . We can use Fisher's (1918) results to predict the relationship between recurrence risk and relationship to affected probands, by considering a trait coded Y =0 for healthy and Y =1 for disease.

  Then,

  $$\text{Population mean}(Y) = \text{Prob}(Y = 1) = \text{Population risk}, \quad K$$

Recurrence risks (continued)

- An alternative algebraic expression for the covariance is

$$\text{Covariance}(Y_1, Y_2) = \text{Mean}(Y_1 Y_2) - \text{Mean}(Y_1)\text{Mean}(Y_2)$$

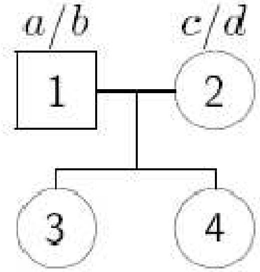
with Mean(Y_1Y_2) the probability that both relatives are affected. From this we derive for the familial risk ratio *λ*, defined before:

$$\frac{\text{Prob}(Y_2 = 1 | Y_1 = 1)}{K} = \frac{\text{Prob}(Y_1 = 1 \ \& \ Y_2 = 1)}{K^2} = 1 + \frac{\text{Covariance}(Y_1, Y_2)}{K^2}$$

- It is intuitively clear (and it can be shown formally) that the covariance between Y_1 and Y_2 depends on the type of relationship (the so-called *kinship coefficient* ϕ (see later)
  - Regression methods may be used for assessing familial aggregation of diseases, using logit link functions

Kinship coefficients

- Consider the familial configuration



and suppose that the first sib (3) inherits the a and c allele.

- Then if 2-IBD refers to the probability that the second sib inherits a and c, it is 1/4 = 1/2×1/2
- If 1-IBD refers to the probability that the second sib inherits a/d or b/c, it is 1/2=1/4 + 1/4
- If 0-IBD refers to the probability that the second sib inherits b and d, it is 1/4

Kinship coefficients (continued)

- We denote this by:

$$z_0 = \frac{1}{4}, \quad z_1 = \frac{1}{2}, \quad z_2 = \frac{1}{4}$$

- Consider a gene at a given locus picked at random, one from each of two relatives. Then the *kinship coefficient* ϕ is defined as the probability that these two genes are IBD.

Kinship coefficients (continued)

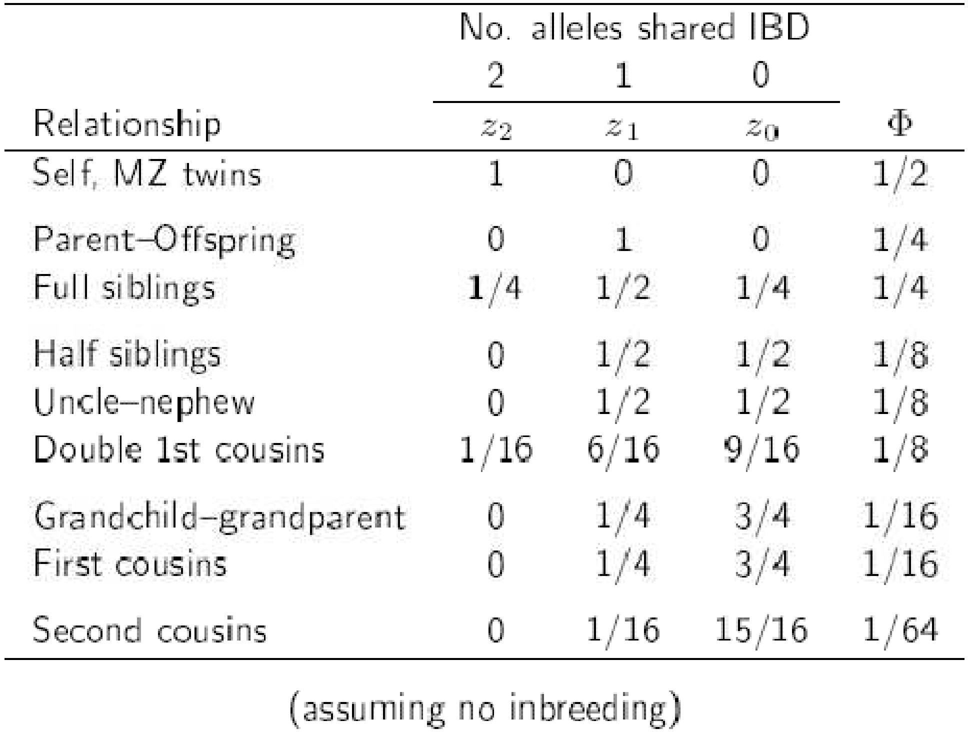
- Given there is no inbreeding (there are no loops in the pedigree graphical representation),
  - If they are 2-IBD, prob = ½
  - If they are 1-IBD, prob = ¼
  - If they are 0-IBD, prob= 0
- So the kinship coefficient

$$\Phi = \frac{1}{2} z_2 + \frac{1}{4} z_1,$$

which is exactly half the average proportion of alleles shared IBD.

- The average number of alleles shared IBD = $2 \times z_2 + 1 \times z_1$
- The average proportion of alleles shared IBD = $(2 \times z_2 + 1 \times z_1)/2$

IBD sharing and kinship by relationship

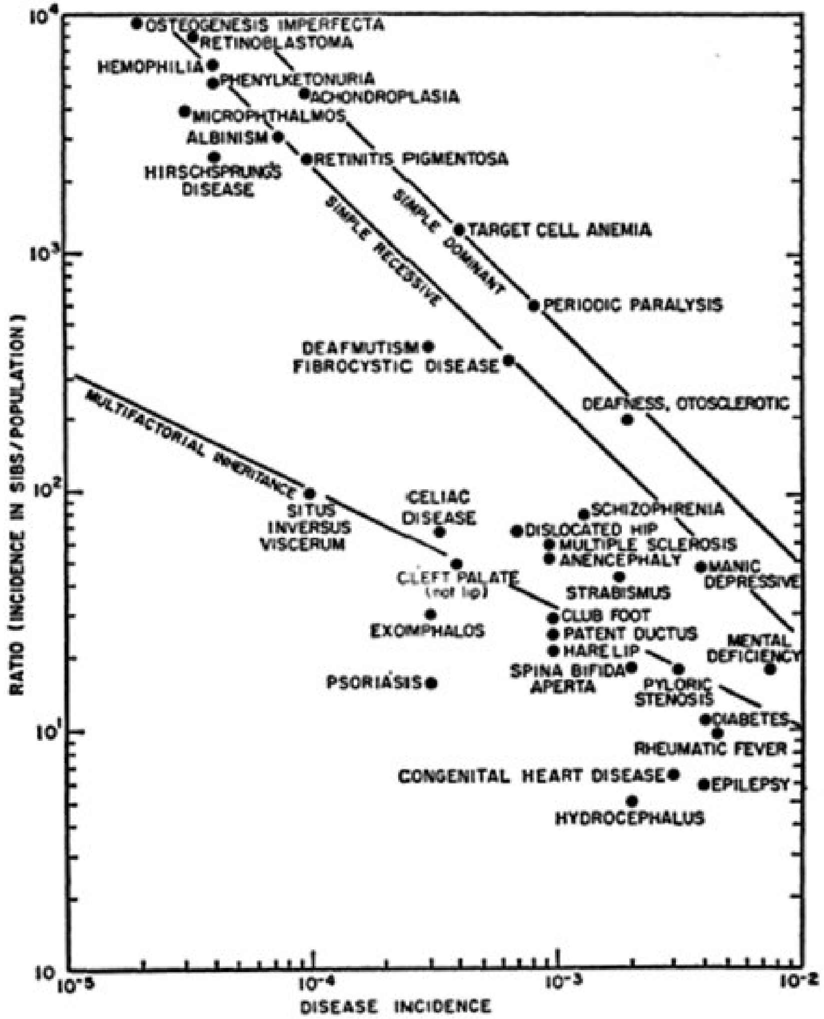


	No. alleles shared IBD			
	2	1	0	
Relationship	z_2	z_1	z_0	Φ
Self, MZ twins	1	0	0	1/2
Parent–Offspring	0	1	0	1/4
Full siblings	1/4	1/2	1/4	1/4
Half siblings	0	1/2	1/2	1/8
Uncle–nephew	0	1/2	1/2	1/8
Double 1st cousins	1/16	6/16	9/16	1/8
Grandchild–grandparent	0	1/4	3/4	1/16
First cousins	0	1/4	3/4	1/16
Second cousins	0	1/16	15/16	1/64

(assuming no inbreeding)

Interpretation of values of relative recurrence risk

- Examples for λ_S = ratio of risk in sibs compared with population risk.
  - cystic fibrosis: the risk in sibs = 0.25 and the risk in the population = 0.0004, and therefore λ_S =500
  - Huntington disease: the risk in sibs = 0.5 and the risk in the population = 0.0001, and therefore λ_S =5000
- Higher value indicates greater proportion of risk in family compared with population.
- The relative recurrence risk increases with
  - Increasing genetic contribution
  - Decreasing population prevalence



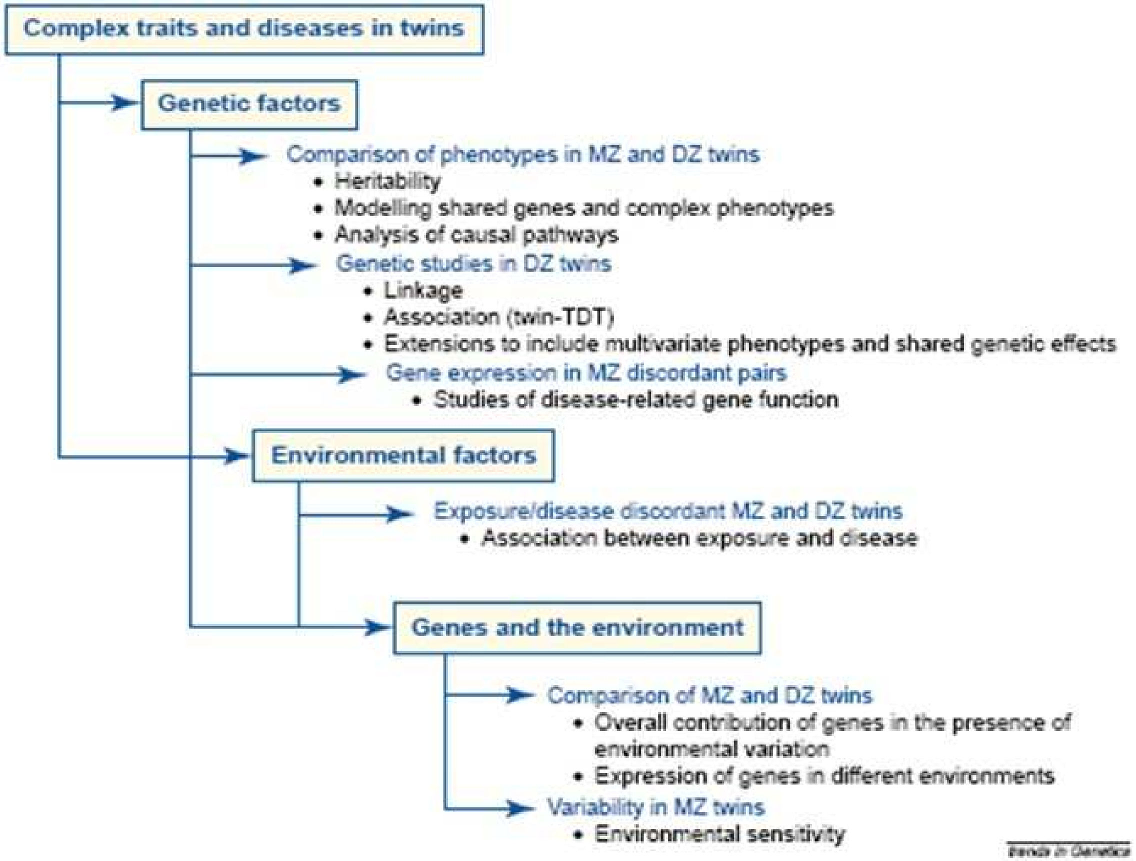
Relation between disease incidence and relative incidence in sibs of affected individuals for a number of diseases. The lines indicate the expected relationships for simple dominant, simple recessive and Edwards' (1963) approximation to multifactorial inheritance (from Newcombe, 1964).

Interpretation of values of relative recurrence risk (continued)

- The presence of familial aggregation can be due to many factors, including including shared family environment.
- Hence, familial aggregation is alone is not sufficient to demonstrate a genetic basis for the disease.
- Here, variance components modeling may come into play to explain the *pattern of familial aggregation* and to derive estimates of *heritability* (see next section: segregation analysis)
- When trying to decipher the importance of genetic versus environmental factors, twin designs are extremely useful:

3. e Twin studies

Environment versus genetics



Complex traits and diseases in twins

Genetic factors

Comparison of phenotypes in MZ and DZ twins

- Heritability
- Modelling shared genes and complex phenotypes
- Analysis of causal pathways

Genetic studies in DZ twins

- Linkage
- Association (twin-TDT)
- Extensions to include multivariate phenotypes and shared genetic effects

Gene expression in MZ discordant pairs

- Studies of disease-related gene function

Environmental factors

Exposure/disease discordant MZ and DZ twins

- Association between exposure and disease

Genes and the environment

Comparison of MZ and DZ twins

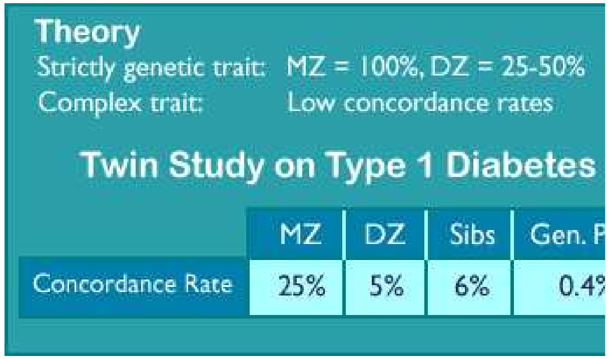
- Overall contribution of genes in the presence of environmental variation
- Expression of genes in different environments

Variability in MZ twins

- Environmental sensitivity

Trends in Genetics

Contribution of twins to the study of complex traits and diseases



Theory

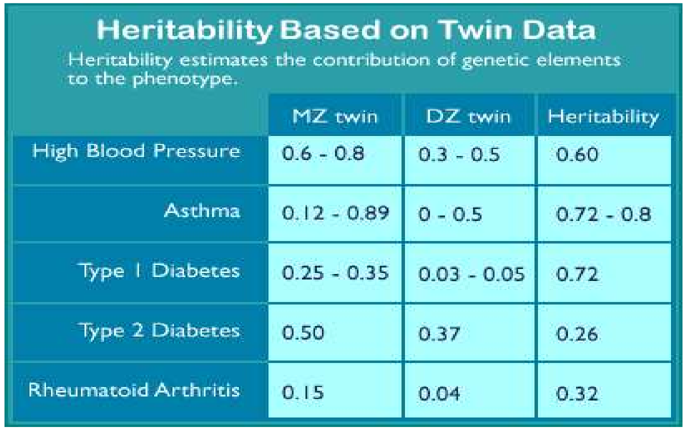
Strictly genetic trait: MZ = 100%, DZ = 25-50%

Complex trait: Low concordance rates

Twin Study on Type 1 Diabetes

	MZ	DZ	Sibs	Gen. P
Concordance Rate	25%	5%	6%	0.4

(Roche Genetics)



Heritability Based on Twin Data

Heritability estimates the contribution of genetic elements to the phenotype.

	MZ twin	DZ twin	Heritability
High Blood Pressure	0.6 - 0.8	0.3 - 0.5	0.60
Asthma	0.12 - 0.89	0 - 0.5	0.72 - 0.8
Type I Diabetes	0.25 - 0.35	0.03 - 0.05	0.72
Type 2 Diabetes	0.50	0.37	0.26
Rheumatoid Arthritis	0.15	0.04	0.32

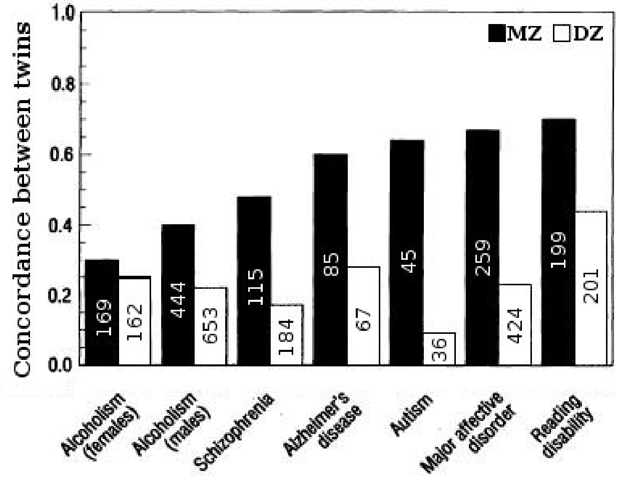
- *Concordance* is defined as is the probability that a pair of individuals will both have a certain characteristic, given that one of the pair has the characteristic.
- For example, twins are concordant when both have or both lack a given trait

Contribution of twins to the study of complex traits and diseases (continued)

- One can distinguish between pairwise concordance and proband wise concordance:
    - *Pairwise concordance* is defined as C/(C+D), where C is the number of concordant pairs and D is the number of discordant pairs
    - For example, a group of 10 twins have been pre-selected to have one affected member (of the pair). During the course of the study four other previously non-affected members become affected, giving a pairwise concordance of 4/(4+6) or 4/10 or 40%.

Contribution of twins to the study of complex traits and diseases (continued)

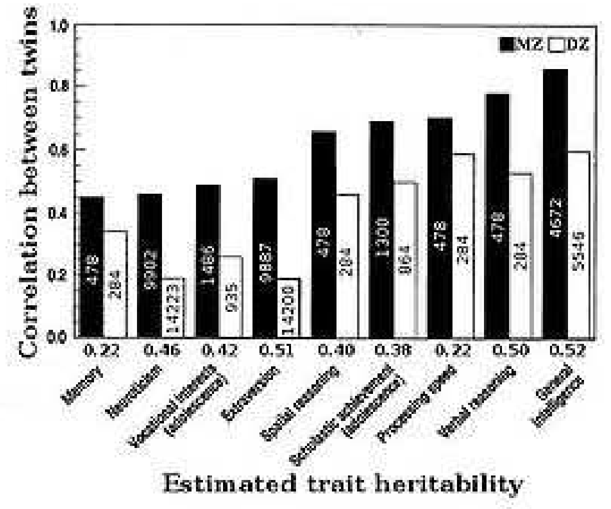
- *Proband wise concordance* is the proportion $(2C_1+C_2)/(2C_1+C_2+D)$, in which $C = C_1+C_2$ and C is the number of concordant pairs, C_2 is the number of concordant pairs in which one and only one member was ascertained and D is the number of discordant pairs.



(http://en.wikipedia.org/wiki/File:Twin-concordances.jpg)

Some details about twin studies

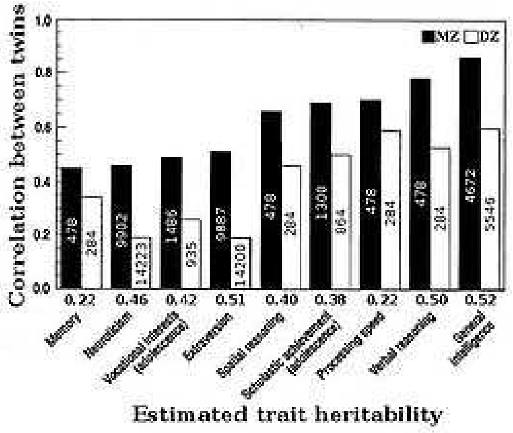
- The basic logic of the twin study can be understood with very little mathematics beyond an understanding of correlation and the concept of variance.
- Classic twin studies begin from assessing the variance of trait in a large group / attempt to estimate how much of this is due to genetic effects (*heritability*), how much appears to be due to shared environmental effects, and how much is due to unique environm. effects (i.e., events occurring to one twin but not another).



(http://en.wikipedia.org/wiki/File:Heritability-from-twin-correlations1.jpg)

Some details about twin studies (continued)

- Identical twins (MZ twins) are twice as genetically similar as DZ twins. So heritability (h^2) is approximately twice the difference in correlation between MZ and DZ twins.
- Unique environmental variance (e^2 or E) is reflected by the degree to which identical twins raised together are dissimilar, and is approximated by 1-MZ correlation.



- The effect of shared environment (c^2 or C) contributes to similarity in all cases and is approximated by the DZ correlation minus the difference between MZ and DZ correlations

Some details about twin studies (continued)

- The three components A (additive genetics), C (common environment) and E (unique environment) give rise to the so-called *ACE Model*.
- It is also possible to examine non-additive genetics effects (often denoted D for dominance (*ADE model*).
- Given the ACE model, researchers can determine what proportion of variance in a trait is heritable, versus the proportions which are due to shared environment or unshared environment, for instance using programs that implement structural equation models (SEM) - available in the freeware Mx software .
    - How does this work in practice? …

Some details about twin studies (continued)

- *Monozygous (MZ) twins* raised in a family share both 100% of their genes, and all of the shared environment (actually, this is often just an assumption). Any differences arising between them in these circumstances are random (unique).
  - The correlation we observe between MZ twins therefore provides an estimate of A + C .
- *Dizygous (DZ) twins* have a common shared environment, and share on average 50% of their genes.
  - *S*o the correlation between DZ twins is a direct estimate of ½A + C .

4 Segregation analysis

Main references:

- Burton P, Tobin M and Hopper J. Key concepts in genetic epidemiology. *The Lancet*, 2005
- Thomas D. Statistical methods in genetic epidemiology. Oxford University Press 2004

- Clayton D. Introduction to genetics (course slides Bristol 2003)
- URL:
    - http://www.dorak.info/

Additional reading:

- Ginsburg E and Livshits G. Segregation analysis of quantitative traits, *Annals of human biology*, 1999

4.a What is a segregation analysis?

Definition of 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 and to test this pattern against predictions from specific genetic models.
- 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.
- As in aggregation studies, families are often ascertained through probands

Definition of segregation analysis (continued)

- Segregation models are fitted using the method of maximum likelihood. In particular, the parameters of the model are fitted by finding the values that maximize the probability (*likelihood*) of the observed data.
- The essential elements of (this often complex likelihood) are
  - the penetrance function
  - the population genotype
  - the transmission probabilities within families
  - the method of ascertainment

Two terms frequently used in a segregation analysis

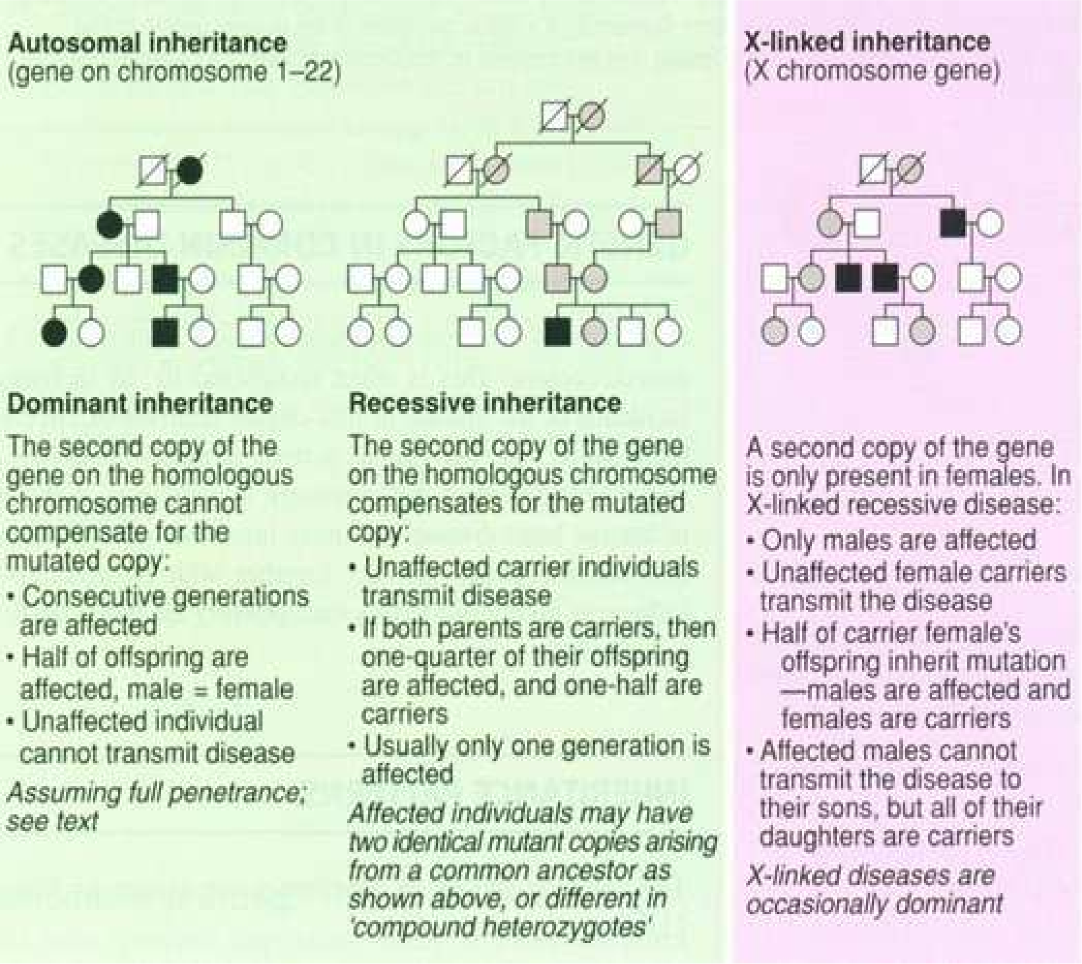
- So the aim of segregation analysis is to find evidence for the existence of a major gene for the phenotype under investigation and to estimate the corresponding mode of inheritance
- The *segregation ratios* are the predictable proportions of genotypes and phenotypes in the offspring of particular parental crosses. e.g. 1 AA : 2 AB : 1 BB following a cross of AB X AB
- *Segregation ratio distortion* is a departure from expected segregation ratios. The purpose of segregation analysis is to detect significant segregation ratio distortion. A significant departure would suggest one of our assumptions about the model wrong.

4.b Classical method for sibships and one locus

Steps of a simple segregation analysis

- Identify mating type(s) where the trait is expected to segregate in the offspring.
- Sample families with the given mating type from the population.
- Sample and score the children of sampled families.
- Estimate segregation ratio or test H_0: “expected segregation ratio” (e.g., hypothesizing a particular mode of inheritance) .

Modes of inheritance



Autosomal inheritance
(gene on chromosome 1–22)

Dominant inheritance

The second copy of the gene on the homologous chromosome cannot compensate for the mutated copy:

- Consecutive generations are affected
- Half of offspring are affected, male = female
- Unaffected individual cannot transmit disease

Assuming full penetrance; see text

Recessive inheritance

The second copy of the gene on the homologous chromosome compensates for the mutated copy:

- Unaffected carrier individuals transmit disease
- If both parents are carriers, then one-quarter of their offspring are affected, and one-half are carriers
- Usually only one generation is affected

Affected individuals may have two identical mutant copies arising from a common ancestor as shown above, or different in 'compound heterozygotes'

X-linked inheritance
(X chromosome gene)

A second copy of the gene is only present in females. In X-linked recessive disease:

- Only males are affected
- Unaffected female carriers transmit the disease
- Half of carrier female's offspring inherit mutation —males are affected and females are carriers
- Affected males cannot transmit the disease to their sons, but all of their daughters are carriers

X-linked diseases are occasionally dominant

Left: single gene and Mendelian inheritance

Also (more complicated):

- Single gene and non-Mendelian (e.g., mitochondrial DNA)
- Multiple genes (e.g., polygenic, oligogenic)

See also Roche Genetics

Example: Autosomal dominant

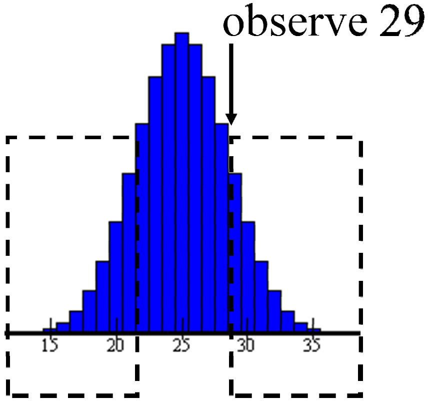
Data and hypothesis:

- Obtain a random sample of matings between affected (*Dd*) and unaffected (*dd*) individuals.
- Sample *n* of their offspring and find that *r* are affected with the disease (i.e. *Dd*).
- H_0: proportion of affected offspring is 0.5

Example: Autosomal dominant (continued)

Binomial test:

- H_0: $p = 0.5$
- If $r \leq n/2$
  - p-value = $2P(X \leq r)$
- If $r > n/2$
  - p-value = $2P(X \leq n\text{-}r)$
- $P(X \leq c) = \sum_{x=0}^{c} \binom{n}{x} \left(\frac{1}{2}\right)^{n}$



p-value = 0.32

4.c Likelihood method for pedigrees and one locus

Segregation analysis involves computing (often very complicated!) probabilities

- For extended pedigrees with many individuals and several generations a numerical procedure is needed for all probability calculations.
- Let *L* denote the likelihood for the observed phenotypes *Y*, given a genetic model *M* and the pedigree structure. *L* can be calculated by summing over all possible genotypic constellations g_i, $i = 1,\dots,N$, where *N* denotes the number of individuals in the pedigree:

$$L(Y) = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_N} P(Y|g_1 g_2 \cdots g_N) P(g_1 g_2 \cdots g_N) \ .$$

- It is assumed that the phenotype of an individual is independent of the other pedigree members given its genotype.
- Widely used in segregation analysis is the **Elston–Stuart** algorithm (Elston and Stuart 1971), a recursive formula for the computation of the likelihood *L* given as

$$L = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_N} \prod_{j=1}^{N} f(g_j) \prod_{k=1}^{N_1} P(g_k) \prod_{m=1}^{N_2} \tau(g_m | g_{m1} g_{m2}) \; .$$

(Bickeböller – Genetic Epidemiology)

- The Elston-Stewart peeling algorithm involves starting at the bottom of a pedigree and computing the probability of the parent’s genotypes, given their phenotypes and the offspring’s phenotypes, and working up from there, at each stage using the genotype probabilities that have been computed at lower levels of the pedigree

The notation for the formula is as follows: N denotes the number of individuals in the pedigree. N_1 denotes the number of *founder* individuals in the pedigree. Founders are individuals without specified parents in the pedigree. In general, these are the members of the oldest generation and married-in spouses.N_2 denotes the number of *non-founder* individuals in the pedigree, such that $N = N_1 + N_2$. g_i, $i = 1,\ldots,N$, denote the genotype of the ith individual of the pedigree.

The parameters of the genetic model M fall into three groups: (1) The genotype distribution $P(g_k)$, $k = 1,\ldots,N_1$, for the founders is determined by population parameters and often Hardy–Weinberg equilibrium is assumed. (2) The transmission probabilities for the transmission from parents to offspring $\tau(g_m|g_{m1}, g_{m2})$, where $m1$ and $m2$ are the parents of m, are needed for all non-founders in the pedigree. It is assumed that transmissions to different offspring are independent given the parental genotypes and that transmissions of one parent to an offspring are independent of the transmission of the other parent. Thus, transmission probabilities can be parametrized by the product of the individual transmissions. Under Mendelian segregation the transmission probabilities for parental transmission are $\tau(S_1|\ S_1\ S_1) = 1$; $\tau(S_1|\ S_1\ S_2) = 0.5$ and $\tau(S_1|\ S_2\ S_2) = 0$. (3) The penetrances $f\ (g_i)$, $i = 1,\ldots,N$, parametrise the genotype-phenotype correlation for each individual i.

4.d Variance component modeling; a general framework

Introduction

- The extent to which any familial aggregation identified is caused by genes, can be estimated by a biologically rational model that specifies how a trait is modulated by the effect of one or more genes.
- One of the most common such models is the additive model:
  - a given allele at a given locus adds a constant to, or subtracts a constant from, the expected value of the trait
- Note that no information about genotypes or measured environmental determinants is required! Hence, no blood needs to be taken for DNA analysis

(Burton et al, The Lancet, 2005)

It is all about variances and covariance of the trait

Consider a trait *Y* which is *quantitative*

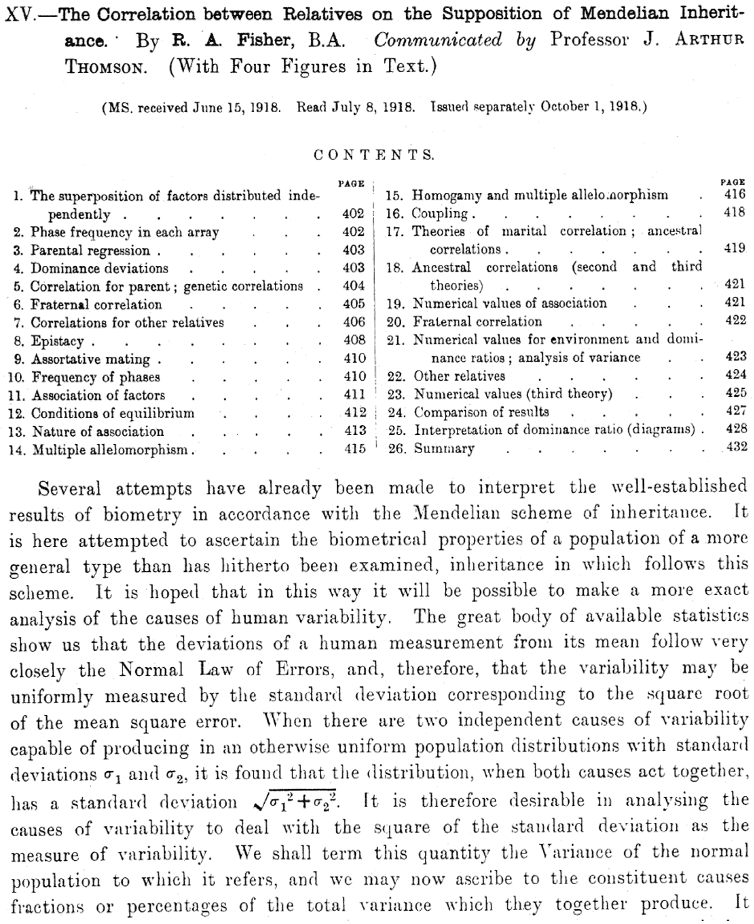
- The *variance* of the trait is the mean squared deviation of *Y* from the population mean

$$\text{Variance}(Y) = \text{Mean}\left\{(Y - \overline{Y})^2\right\}$$

- The *covariance* of the trait between two subjects is the mean of the products of their deviations from the population mean:

$$\text{Covariance}(Y) = \text{Mean}\left\{(Y_1 - \overline{Y}) \times (Y_2 - \overline{Y})\right\}$$

- The *correlation coefficient* is the covariance scaled to lie between 1 and -1

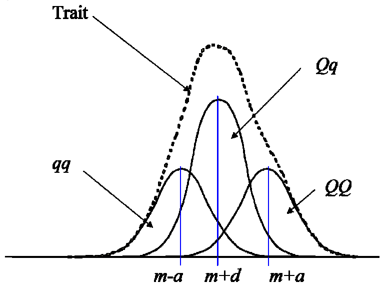


XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

CONTENTS.

Several attempts have already been made to interpret the well-established results of biometry in accordance with the Mendelian scheme of inheritance. It is here attempted to ascertain the biometrical properties of a population of a more general type than has hitherto been examined, inheritance in which follows this scheme. It is hoped that in this way it will be possible to make a more exact analysis of the causes of human variability. The great body of available statistics show us that the deviations of a human measurement from its mean follow very closely the Normal Law of Errors, and, therefore, that the variability may be uniformly measured by the standard deviation corresponding to the square root of the mean square error. When there are two independent causes of variability capable of producing in an otherwise uniform population distributions with standard deviations σ_1 and σ_2, it is found that the distribution, when both causes act together, has a standard deviation $\sqrt{\sigma_1^2+\sigma_2^2}$. It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance of the normal population to which it refers, and we may now ascribe to the constituent causes fractions or percentages of the total variance which they together produce. It



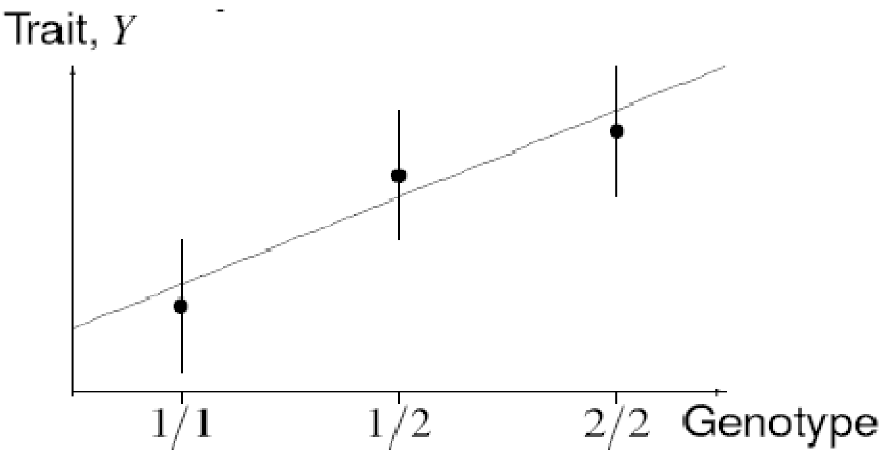
RA Fisher (1918). *Transactions of the Royal Society of Edinburgh* **52**: 399-433.

Components of the genetic variance

- In 1918, Fisher established the relationship between the covariance in trait values between two relatives and their relatedness
- The resulting correlation matrix can be analyzed by variance components or path analysis techniques to estimate the proportion of variance due to shared environmental and genetic influences.
- In an "analysis of variance" framework,

- the additive component of variance is the variance explained by a model in which maternal and paternal alleles have simple additive effects on the mean trait value.
- The dominance component represents residual genetic variance not explained by a simple sum of effects

Example: a bi-allelic locus



- Environment variance is represented by the vertical bars
- Total genetic variance is variance between genotype means (●)
    - Additive component is that *due to* the regression line,
    - Dominance component is that *about* the regression line

Trait covariances and IBD (no shared environmental influences)

- Two individuals who share **2 alleles IBD** at the trait locus are genetically identical in so far as that trait is concerned. The covariance between their trait values is the total genetic variance

$$\sigma^2_{\text{Gen}} = \sigma^2_{\text{Add}} + \sigma^2_{\text{Dom}}$$

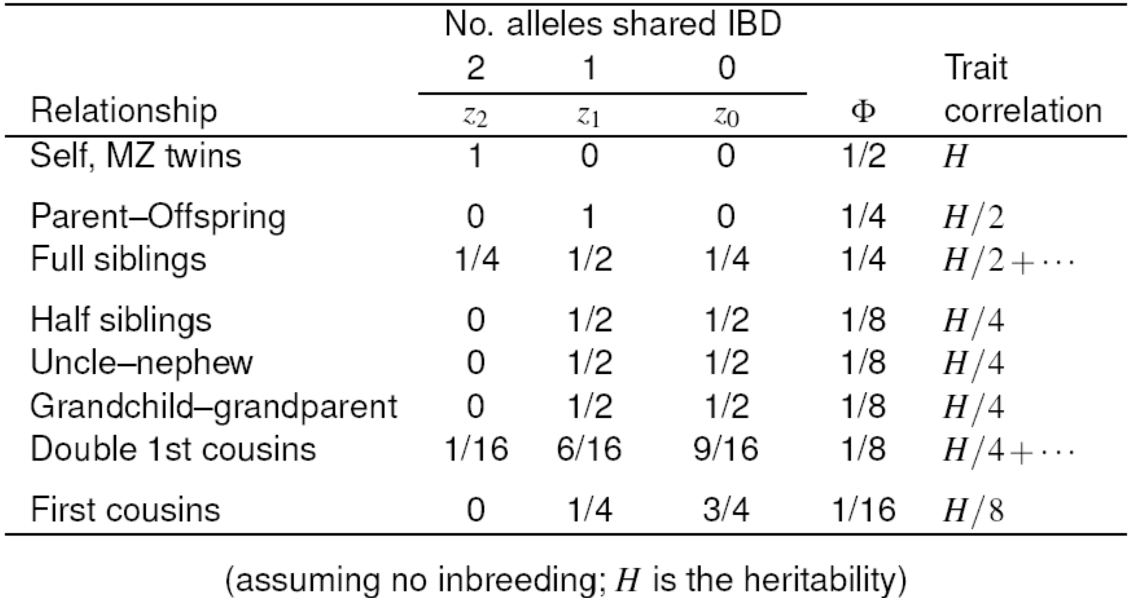
- Two individuals who share **1 allele IBD** at the trait locus share the genetic effect of that allele. The covariance between their trait values is half the additive component of variance,

$$\sigma^2_{\text{Add}}/2$$

- Two individuals who share **0 alleles IBD** at the trait locus are effectively unrelated. The covariance between their trait values is zero
  Therefore, the covariance between trait values in two relatives is

$$z_1\frac{\sigma^2_{\text{Add}}}{2} + z_2(\sigma^2_{\text{Add}} + \sigma^2_{\text{Dom}}) = 2\Phi\sigma^2_{\text{Add}} + z_2\sigma^2_{\text{Dom}}$$

IBD sharing, kinship and trait correlation



	No. alleles shared IBD				
	2	1	0		Trait
Relationship	z_2	z_1	z_0	Φ	correlation
Self, MZ twins	1	0	0	1/2	H
Parent–Offspring	0	1	0	1/4	$H/2$
Full siblings	1/4	1/2	1/4	1/4	$H/2+\cdots$
Half siblings	0	1/2	1/2	1/8	$H/4$
Uncle–nephew	0	1/2	1/2	1/8	$H/4$
Grandchild–grandparent	0	1/2	1/2	1/8	$H/4$
Double 1st cousins	1/16	6/16	9/16	1/8	$H/4+\cdots$
First cousins	0	1/4	3/4	1/16	$H/8$

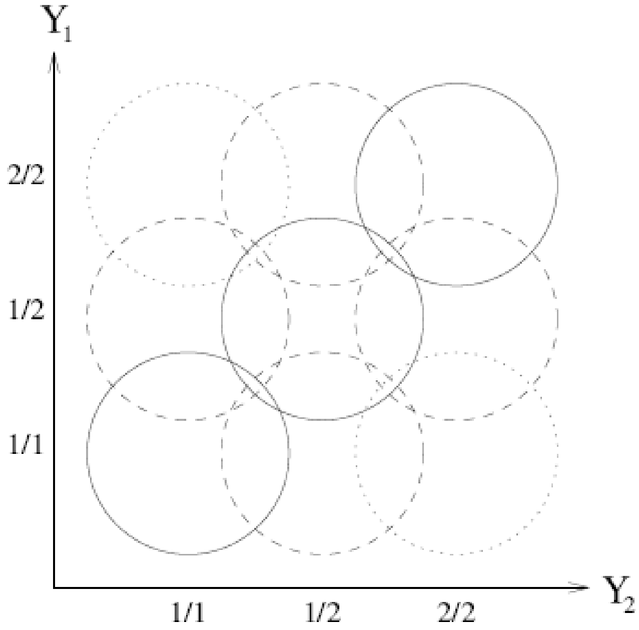
(assuming no inbreeding; H is the heritability)

- The dominance component is frequently small so that covariance (and hence correlation) is proportional to the kinship coefficient

Single major locus

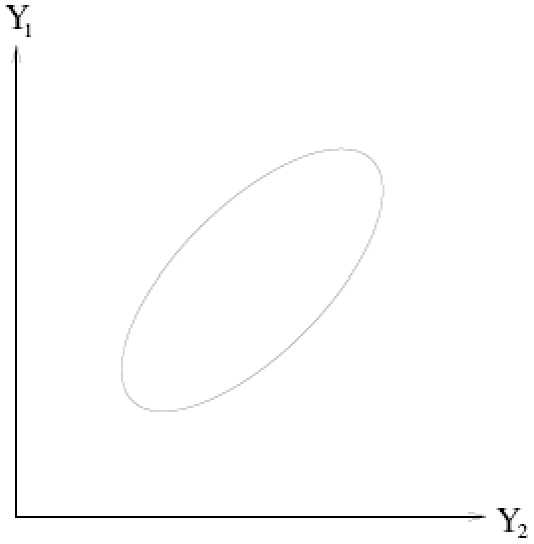
- If inheritance of the trait were due to a single major locus, the bivariate distribution for two relatives would be a mixture of circular clouds of points

  - Spacing of cloud centres depends on additive and dominance effects
  - Marginal distributions depend on allele frequency
  - Tendency to fall along diagonals depends on IBD status (hence on relationship



Polygenic model

- In the model for polygenic inheritance, the trait is determined by the sum of very many small effects of different genes
- The distribution of the trait in two relatives, *Y*1 and *Y*2, is *bivariate normal* . an elliptical cloud of points

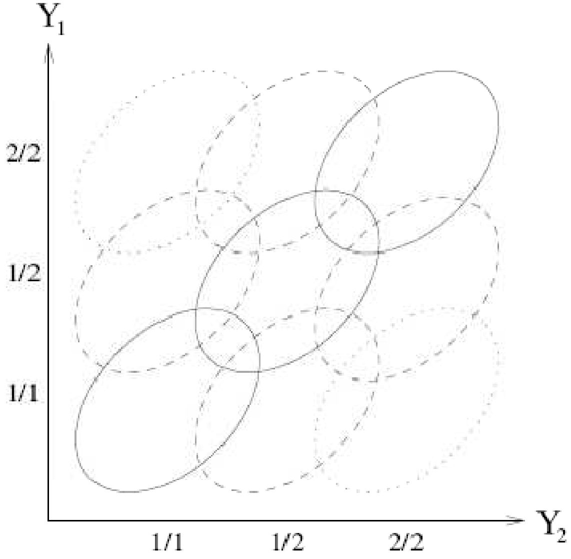


Correlation is determined by

- Degree of relationship (IBD probabilities)
- Heritability

The Morton-Maclean model (the "mixed model")

- In this model, the trait is determined by additive effects of a single major locus plus a polygenic component. The bivariate distribution for two relatives is now a mixture of elliptical clouds:



- The regressive model provides a convenient approximation to the mixed model.

The Morton-Maclean model (continued)

- This model can be fitted to trait values for individuals in pedigrees, using the method of maximum likelihood
- It is necessary to allow for the manner in which pedigrees have been recruited into the study, or ascertained pedigrees in the study may be skewed, either deliberately or inadvertently, towards those with extreme trait values for one or more family members

- Segregation analyses were often over-interpreted . the results depend on very strong model assumptions:
  - additivity of effects (major gene, polygenes, and environment)
  - bivariate normality of distribution of trait given genotype at the major locus

Types of variance component modeling

- Variance components analysis can be undertaken with conventional techniques such as maximum likelihood and generalised least squares, or Markov chain Monte Carlo based approaches.
- Genetic epidemiologists use various approaches to aid the specification of such models, including path analysis, which was invented by Sewall
- Wright nearly 100 years ago and the fitting is achieved by various programs.
- If information is available about characterised genotypes, measured environmental determinants, and known demographics, it can enter the analysis.
- Equivalent approaches can also be used for binary phenotypes (liability threshold models) and for traits that can best be expressed as a survival time such as age at onset or age at death.

(Burton et al, The Lancet, 2005)

4.e The ideas of variance component modeling adjusted for binary traits

- Aggregation of discrete traits, such as diseases in families have been studied by an extension of the Morton.Maclean model
- Assume a latent liability to disease behaves as a quantitative trait, with a mixture of major gene and polygene effects. When liability exceeds a threshold, disease occurs
- This model may be fitted by maximum likelihood, although ascertainment corrections can be troublesome
- As in the quantitative trait case, this approach relies upon strong modeling assumptions

4.f Quantifying the genetic importance in familial resemblance

Heritability

- Recall: One of the principal reasons for fitting a variance components model is to estimate the variance attributable to additive genetic effects
- This quantity represents that component of the total phenotypic variance, usually after adjustment for measured genetic and non-genetic determinants that can be attributed to unmeasured additive genetic effects:
    - Leads to a *narrow-sense* definition of *genetic heritability*.
- *Broad sense heritability* is defined as the proportion of the total phenotypic variance that is attributable to all genetic effects, including non-additive effects at individual loci and between loci.

(Burton et al, The Lancet, 2005)

5 Genetic epidemiology and public health

See first class

In-class discussion document

- Visscher et al. Heritability in the genomics era. *Nature Genetics*, 2008.
- Janssens et al. An epidemiological perspective on the future of
- direct-to-consumer personal genome testing. *Investigative Genetic,;* 2010

Background Reading:

- WHO document by Bonita et al: Basic epidemiology
- Burton et al 2005. Genetic Epidemiology 1: Key concepts in genetic epidemiology. The Lancet, 366: 941–51
- MCJ Dekker and CM van Duijn. Prospects of genetic epidemiology in the 21st century. European Journal *of Epidemiology*, 2003.