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

Intra-class (intra-family) correlation coefficient

3.c Familial aggregation with dichotomous traits

Relative recurrence risk, IBD and kinship coefficient

3.d Quantifying genetics versus environment

Heritability

4 Segregation analysis

4.a What is segregation analysis?

Segregation ratios

4.b Genetic models

From easy to complex modes of inheritance

4.c Genetic heterogeneity

One locus, multiple loci

1 Basic epidemiology

Main references:

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

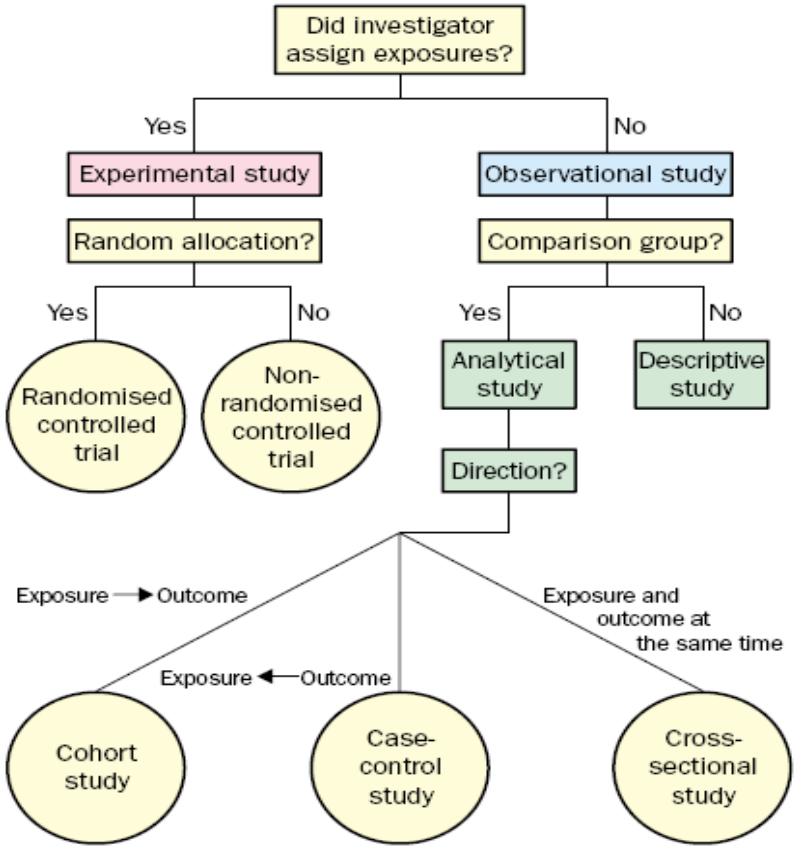
- Clayton D. Introduction to genetics (course slides Bristol 2003)
- Bonita R, Beaglehole R and Kjellström T. *Basic Epidemiology*. WHO 2nd edition
- URL:
  - http://www.dorak.info/

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.

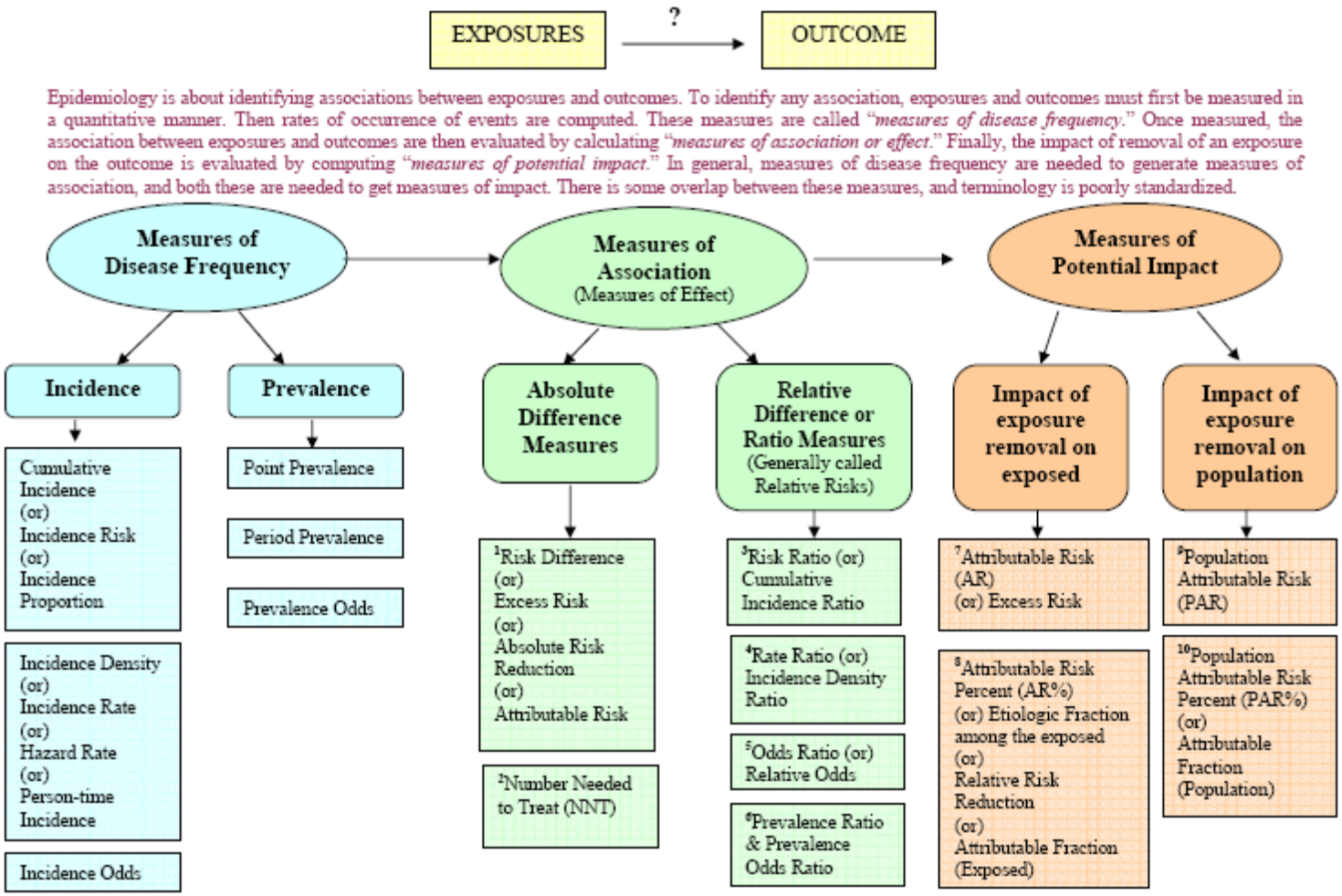
1.b Designs in epidemiology

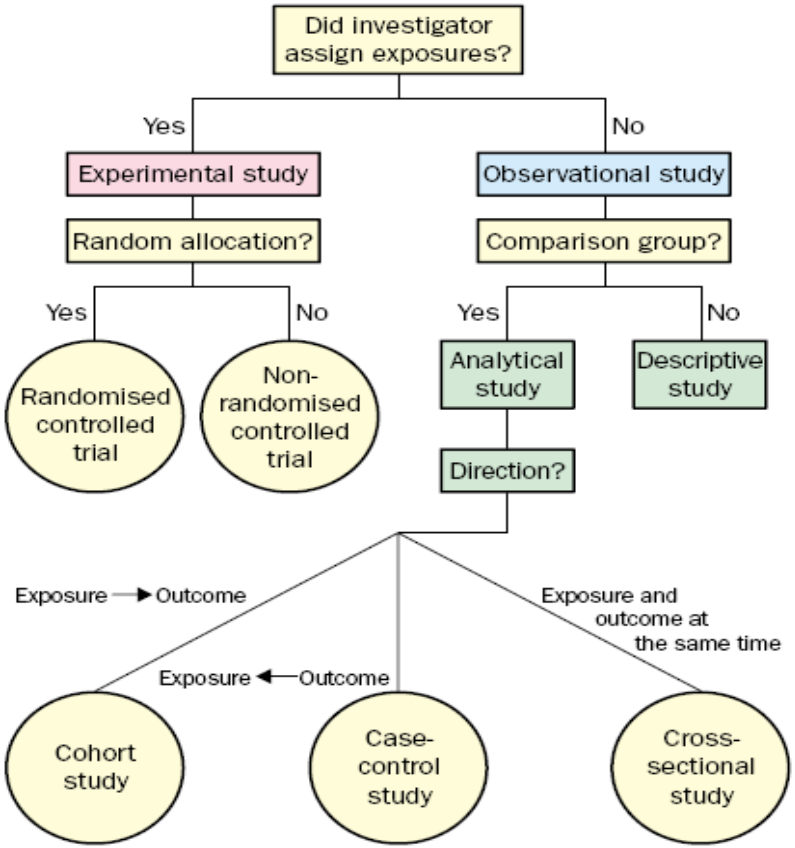
- A focus of an epidemiological study is the population defined in geographical or other terms



(Grimes & Schulz 2002)

1.c An overview of measurements in epidemiology





(Grimes and Schulz 2002)

Summary of most important features by design

	Cross-sectional study	Case-control study	Cohort study
Measure of disease frequency	Prevalence	Prevalence	Incidence
Direction of investigation	momentary/ Retrospective	Retrospective	Prospective
Samples (selections) involved	1 sample from the population	1 group of cases, 1 group of controls	1 cohort of exposed, 1 cohort of unexposed
Primary measure of association	Prevalence odds ratio	Odds ratio	Relative risk; attributable risk

Summary of major advantages (bold) and disadvantages

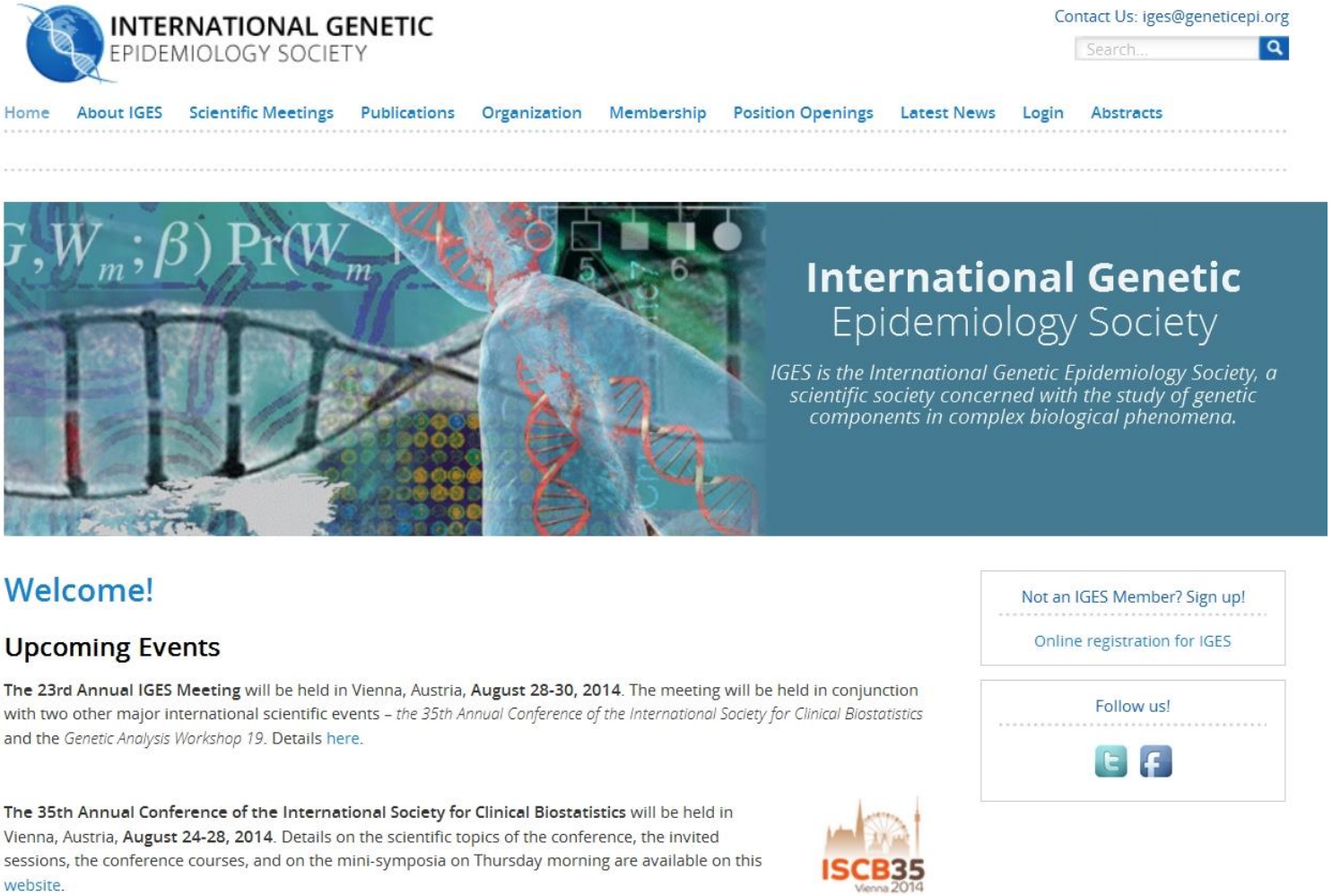
	Cross-sectional study	Case-control study	Cohort study
Marginal conditions	**quick** **relatively cheap**	**quick** **relatively cheap**	time-consuming relatively costly
Applicability	permanent risk factors quite common dis.	**more general** **rare diseases**	**more general**
Data quality	**as good as diagnosis**	errors in historic data	**as good as diagnosis**
Sample sizes	large (low prevalences)	**relatively small**	large (dropout, low inc.)
Inferences/ estimatability	no causal evidence no incidence **prev. of exposure** **prev. of disease**	limited causal evidence no incidence **prev. of exposure** no prev. of disease	**causal evidence** **incidence** no prev. of exposure **prev. of disease**

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/
  - http://www.answers.com/topic/
  - http://www.arbo-zoo.net/_data/ArboConFlu_StudyDesign.pdf

2.a What is genetic epidemiology?



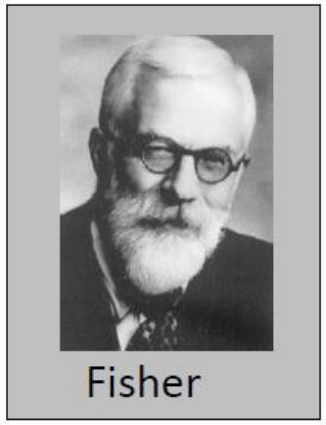
Statistical Genetics

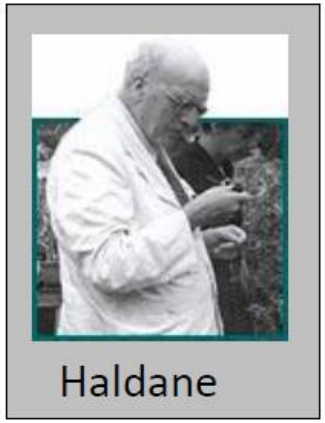
- Genetic epidemiology is closely allied to both molecular epidemiology and **statistical genetics**, but these overlapping fields each have distinct emphases, societies and journals.
- Statistical geneticists are highly trained scientific investigators who are specialists in both statistics and genetics: Statistical geneticists must be able to understand molecular and clinical genetics, as well as mathematics and statistics, to effectively communicate with scientists from these disciplines.
- Statistical genetics is a very exciting professional area because it is so new and there is so much demand. It is a rapidly changing field, and there are many fascinating scientific questions that need to be addressed. Additionally, given the interdisciplinary nature of statistical genetics, there are plenty of opportunities to interact with researchers and clinicians in other fields, such as epidemiology, biochemistry, physiology, pathology, evolutionary biology, and anthropology.

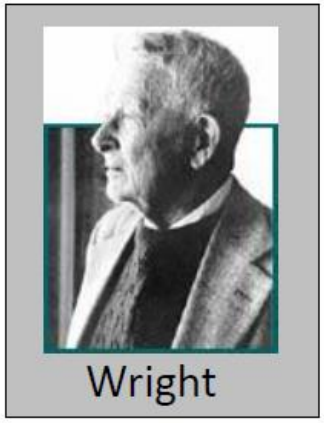
True or False?

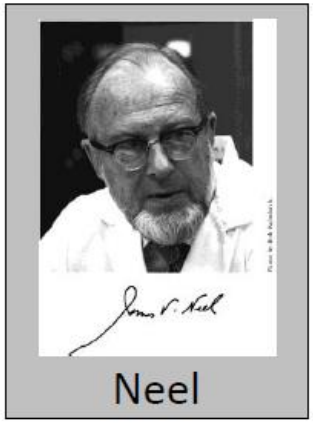
- A primary **difference between statistical genetics and genetic epidemiology** is that statistical geneticists are often more interested in the development and evaluation of new statistical methods, whereas genetic epidemiologists focus more on the application of statistical methods to biomedical research problems.

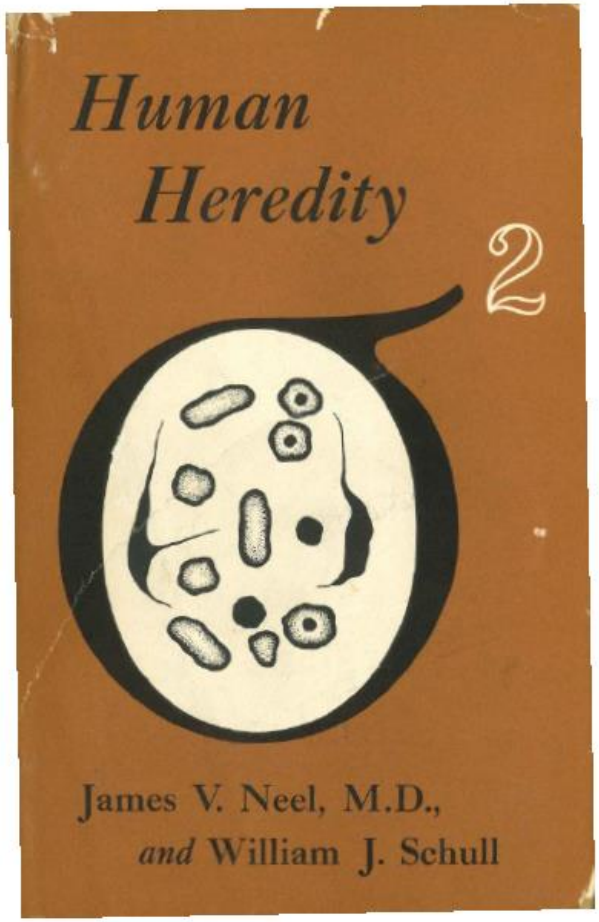
- A primary **difference between genetic and molecular epidemiology** is that the first is also concerned with the detection of inheritance patterns.


Fisher


Haldane


Wright


Neel



Founders of Statistical Genetics

(IGES presidential address A Ziegler, Chicago 2013)

Towards a definition for genetic epidemiology …

No agreement

Khoury, Beaty, Cohen: Researchers have still not fully agreed on the definition and the scope of genetic epidemiology.



(IGES presidential address A Ziegler, Chicago 2013)

Towards a definition for genetic epidemiology …

interaction between “genetic” and “epi” (1984)?

Rao: Genetic epidemiology … represents an important interaction between the two parent disciplines: genetics and epidemiology. …

- [It] **differs from epidemiology** by its explicit consideration of genetic factors and family resemblance
- It **differs from population genetics** by its focus on disease
- It … **differs from medical genetics** by its emphasis on population aspects

(IGES presidential address A Ziegler, Chicago 2013)

Towards a definition for genetic epidemiology …

via the process of defining genetic basis (1086, 2004)?

Thomas: … The process of defining the genetic basis of a disease usually follows a progression such as the [following] …

- Descriptive epidemiology
- Familial aggregation
- Segregation analysis
- Linkage analysis
- Fine mapping
- Association
- Cloning
- Characterization

(IGES presidential address A Ziegler, Chicago 2013)

Towards a definition for genetic epidemiology …

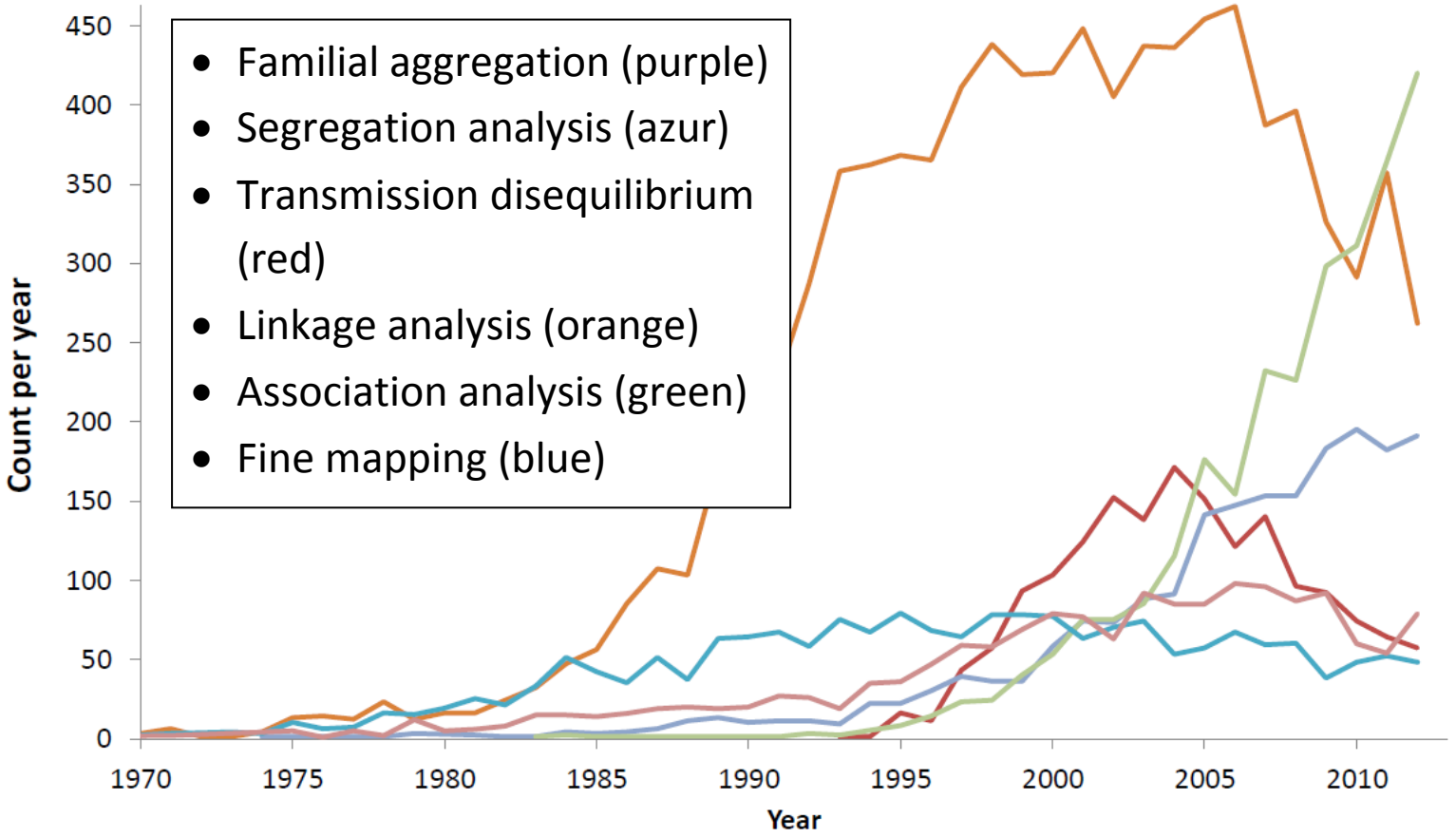
- Term firstly used by Morton & Chung (1978)
- Genetic epidemiology examines the role of genetic factors, along with the environmental contributors to disease, and at the same time giving equal attention to the differential impact of environmental agents, non-familial as well as familial, on different genetic backgrounds (Cohen, Am J Epidemiol, 1980)
- Genetic epidemiology is the study of how and why diseases cluster in families and ethnic groups (King et al., 1984)
- 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 & Chung, 1978 --> 1995).

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)

Use of genetic terms over time



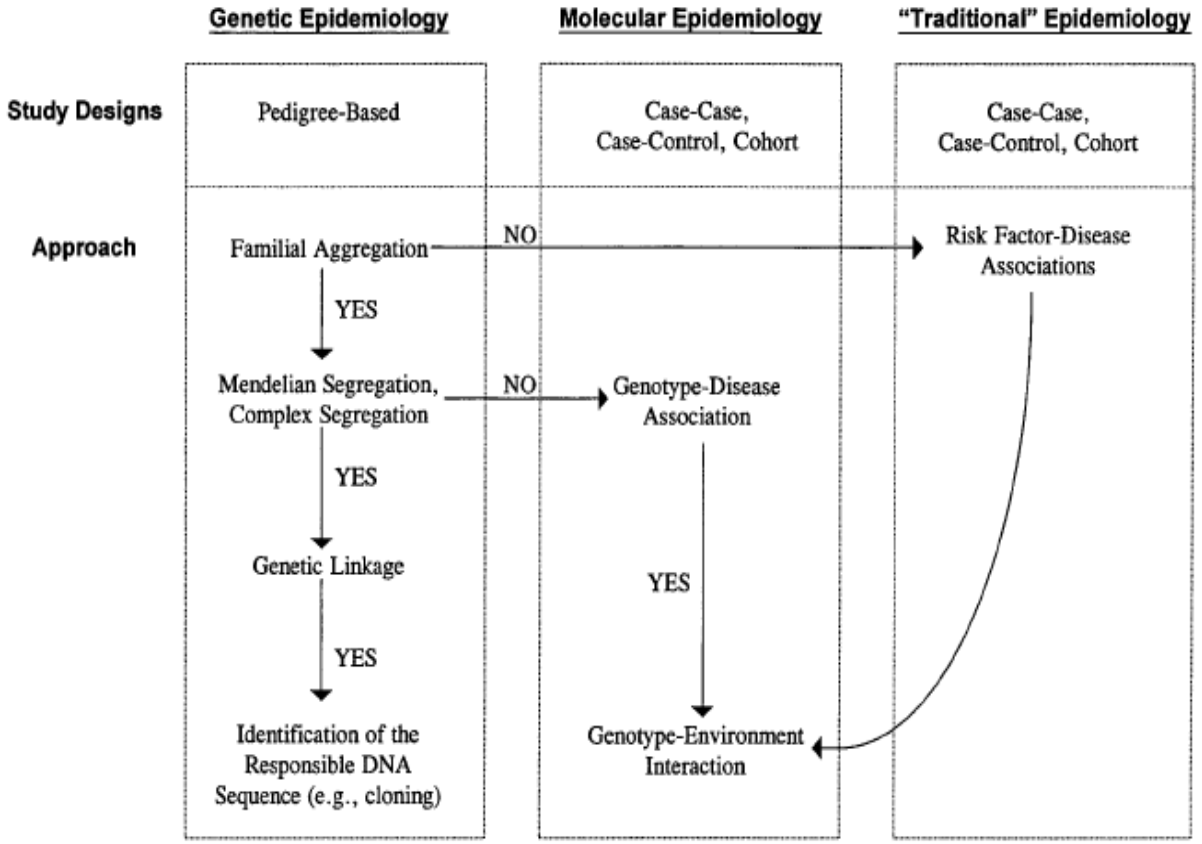
(adapted from IGES presidential address A Ziegler, Chicago 2013)

X-epidemiology

- The phrase "molecular epidemiology" was first coined in 1973 by Kilbourne in an article entitled "The molecular epidemiology of influenza".
- The term became more formalized with the formulation of the first book on "Molecular Epidemiology: Principles and Practice" by Schulte and Perera.

- Nowadays, molecular epidemiologic studies measure exposure to specific substances (DNA adducts) and early biological response (somatic mutations), evaluate host characteristics (genotype and phenotype) mediating response to external agents, and use markers of a specific effect (like gene expression) to refine disease categories (such as heterogeneity, etiology and prognosis).

X – epidemiology



(Rebbeck TR, *Cancer*, **1999**)

New kids around the block

The field of public health genomics (Khoury 2010)

Khoury et al.: Public health genomics, a multidisciplinary field concerned with the **effective and responsible translation of genome-based knowledge and technologies to improve population health**. … Public health genomics uses population-based data on genetic variation and gene-environment interactions to develop, implement, and evaluate evidence-based tools for improving health and preventing disease

(IGES presidential address A Ziegler, Chicago 2013)

Towards a definition for genetic epidemiology …

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**

+

- Statistical methodology
- Genome-wide association studies
- Next generation sequencing
- Gene-environment interaction
- Family studies
- Risk score
- Predictive markers & pharmacogenetics
- Microbiome
- Epigenetics
- eQTL
- Other Omics

(IGES presidential address A Ziegler, Chicago 2013)

The genetic epidemiology context

- In contrast to classic epidemiology, the three main complications in **modern** 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.

Key concepts in genetic epidemiology

Genetic Epidemiology 1

Key concepts in genetic epidemiology

Paul R Burton, Martin D Tobin, John L Hopper

This article is the first in a series of seven that will provide an overview of central concepts and topical issues in modern genetic epidemiology. In this article, we provide an overall framework for investigating the role of familial factors, especially genetic determinants, in the causation of complex diseases such as diabetes. The discrete steps of the framework to be outlined integrate the biological science underlying modern genetics and the population science underpinning mainstream epidemiology. In keeping with the broad readership of *The Lancet* and the diverse background of today's genetic epidemiologists, we provide introductory sections to equip readers with basic concepts and vocabulary. We anticipate that, depending on their professional background and specialist knowledge, some readers will wish to skip some of this article.

Lancet 2005; 366: 941–51

See Comment page 880

This is the first in a Series of seven papers on genetic epidemiology.

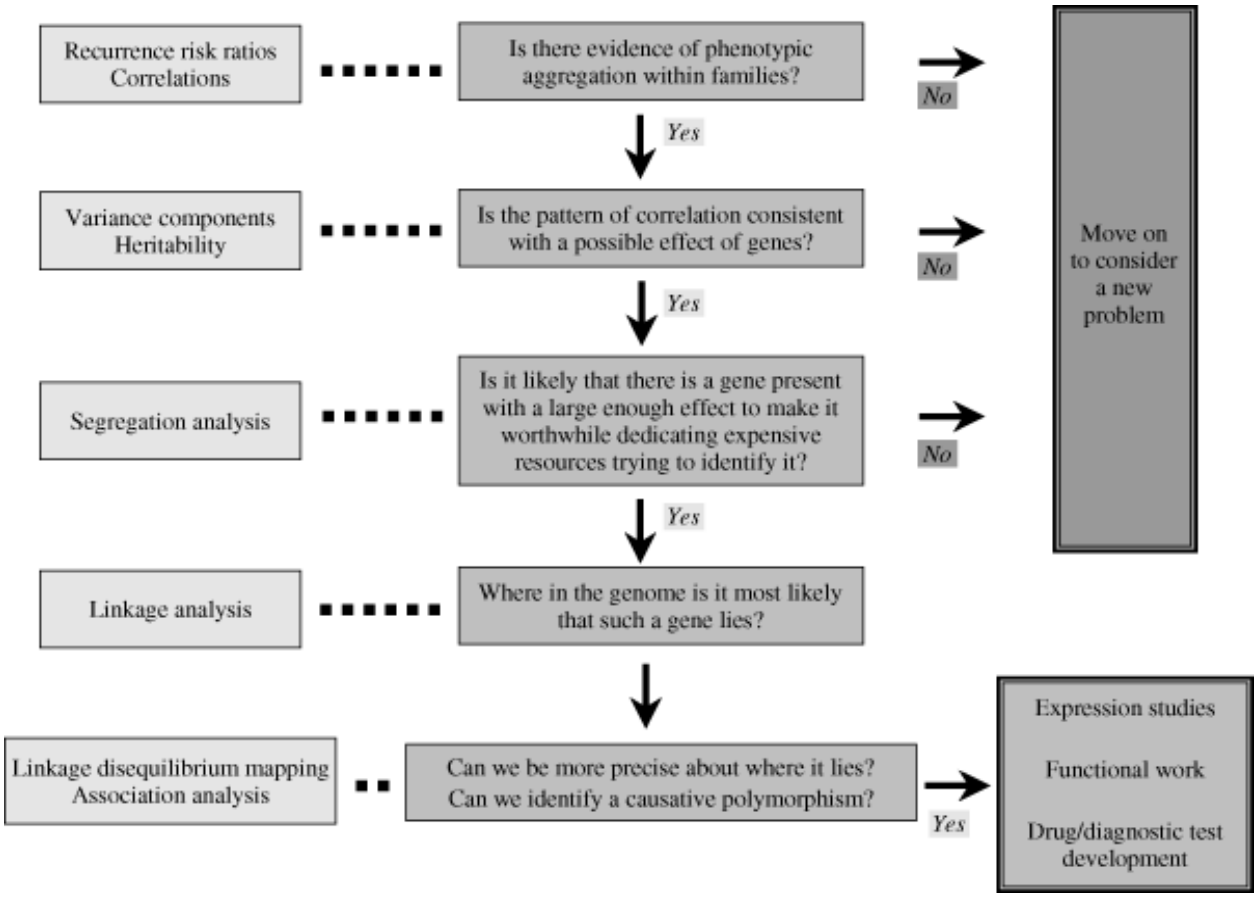
Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, UK (Prof P R Burton MD, M D Tobin PhD); and Centre for Genetic Epidemiology, University of Melbourne, Melbourne, Victoria, Australia (Prof J L Hopper PhD)

Correspondence to: Prof Paul R Burton, Department of Health Sciences, University of Leicester, 22–28 Princess Road West, Leicester LE1 6TP, UK
pb51@le.ac.uk

What is genetic epidemiology?

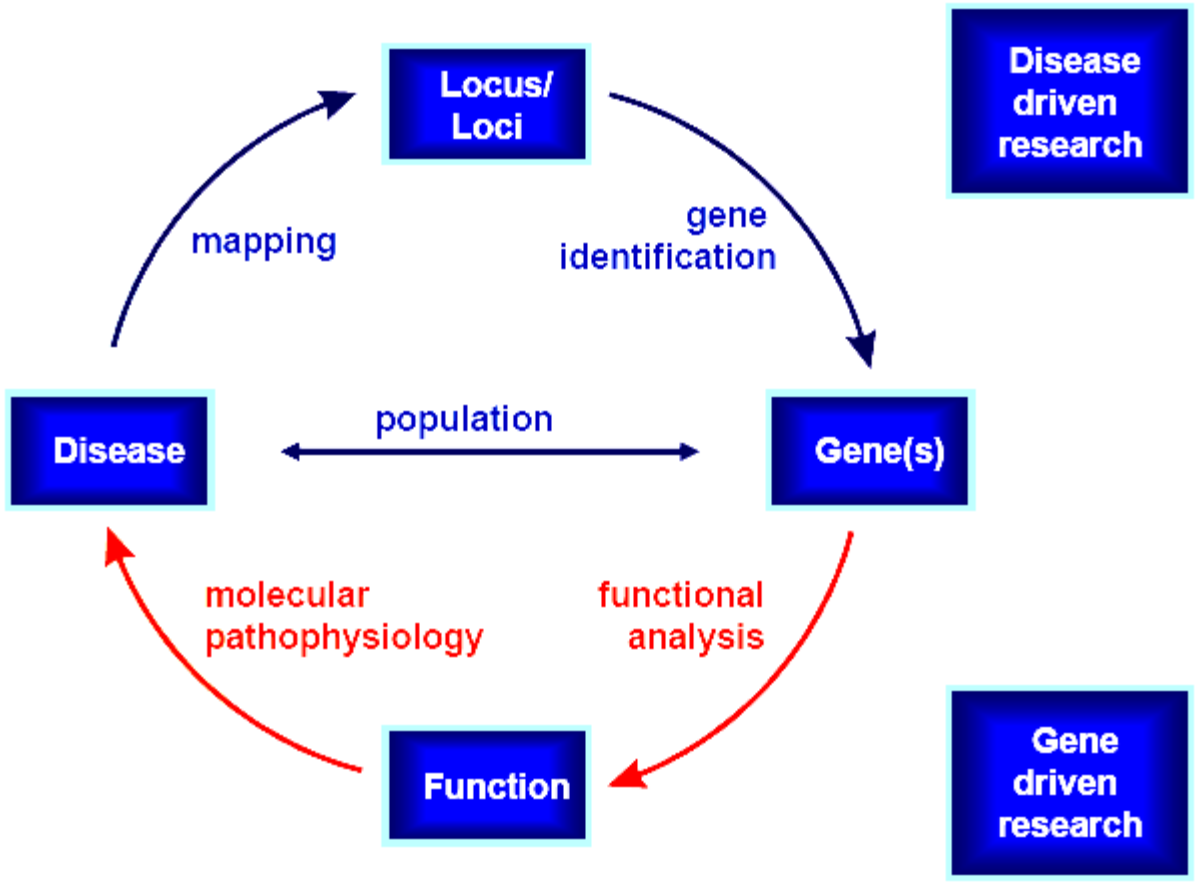
Epidemiology is usually defined as "the study of the distribution, determinants [and control] of health-related states and events in populations".[1] By contrast, genetic epidemiology means different things to different people.[2–7] We regard it as a discipline closely allied to traditional epidemiology that focuses on the familial, and in particular genetic, determinants of disease and the joint effects of genes and non-genetic determinants. Crucially, appropriate account is taken of the biology that underlies the action of genes and the close. The marker and the causative variant need not be within the same gene. This principle is the basis of genetic linkage analysis (see a later paper in this series[12]), which has achieved many of the breakthroughs in the genetics of disease causation. Many such breakthroughs involve conditions caused by variants in a single gene and have been achieved by geneticists and clinical geneticists who would not view themselves as genetic epidemiologists. Nevertheless, linkage analysis is one of the most important tools available to the genetic epidemiologist.

Relevant questions in genetic epidemiology



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

Genetic research paradigm



"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)

Genetic epidemiology and public health

Workshop paper (class 1) - 2003

European Journal of Epidemiology **18**: 607–616, 2003.

REVIEW

Prospects of genetic epidemiology in the 21st century

Marieke C.J. Dekker & Cornelia M. van Duijn
Department of Epidemiology and Biostatistics, Erasmus MC, Rotterdam, The Netherlands

Accepted in revised form 14 April 2003

Abstract. Genetic epidemiology is a young but rapidly developing discipline. Although its early years were largely dedicated to family-based research in monogenic disorders, now genetic–epidemiologic research increasingly focuses on complex, multifactorial disorders. Along with the development of the human-genome map and advances in molecular technology grows the importance of genetic–epidemiologic applications. Large-scale population-based studies, requiring close integration of genetic and epidemiologic research, determine future research in the field. In this paper, we review the basic principles underlying genetic–epidemiologic research, such as molecular genetics and familial aggregation of disease, as well as the typical study approaches of genome screening and candidate-gene studies.

Key words: Familial aggregation, Genetics, Genetic epidemiology, Polymorphisms, Study design

Abbreviations: APOE = apolipoprotein-E gene; CYP2D6 = cytochrome P450 debrisoquine-4-hydroxylase gene; DNA = deoxyribose nucleic acid; LOD score = logarithm-of-odds score; PSEN = presenilin gene; RNA = ribonucleic acid; SNP = single-nucleotide polymorphism; STR = short tandem-repeat; TDT = transmission-disequilibrium test; UCH-L1 = ubiquitin carboxy-terminal hydroxylase L1

Background reading - 2005

Genetic Epidemiology 7

Genetic epidemiology and public health: hope, hype, and future prospects

George Davey Smith, Shah Ebrahim, Sarah Lewis, Anna L Hansell, Lyle J Palmer, Paul R Burton

Lancet 2005; 366: 1484–98
This is the seventh, and final, paper in a Series on genetic epidemiology.

Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK (Prof G Davey Smith FRCP, Prof S Ebrahim FRCP, S Lewis PhD); Department of Epidemiology, Imperial College, London, UK (A L Hansell MB); Laboratory for Genetic Epidemiology, School of Population Health and Western Australian Institute for Medical Research and Centre for Medical Research, University of Western Australia, Perth, Australia (Prof L J Palmer PhD); and Department of Health Sciences and Genetics, University of Leicester, Leicester, UK (Prof P R Burton MD)

Genetic epidemiology is a rapidly expanding research field, but the implications of findings from such studies for individual or population health are unclear. The use of molecular genetic screening currently has some legitimacy in certain monogenic conditions, but no established value with respect to common complex diseases. Personalised medical care based on molecular genetic testing is also as yet undeveloped for common diseases. Genetic epidemiology can contribute to establishing the causal nature of environmentally modifiable risk factors, throught the application of mendelian randomisation approaches and thus contribute to appropriate preventive strategies. Technological and other advances will allow the potential of genetic epidemiology to be revealed over the next few years, and the establishment of large population-based resources for such studies (biobanks) should contribute to this endeavour.

The recent advances covered in this series have equipped genetic epidemiologists with powerful methods for studying the genetic architecture of complex diseases, but direct contributions to public health have been restricted so far. The major current focus is on attempts to use genetic variants to identify individuals who are at high risk of disease, coupled with appropriate management to reduce their risk.[1] The potential of pharmacogenomic studies to contribute to personalised medicine has also been widely heralded.[2–4] Major contributions to either health care or public health are only just beginning to be made. More encouragingly, findings from association thinking and appropriately designed studies are needed. In this article, we discuss the current and potential effects of the genomic revolution on public health science and mainstream epidemiology, especially in the context of the very large-scale population resources (Biobanks) that are being established internationally.

Genomic profiling in the prevention and treatment of common diseases

Since the launch of the human genome project the potential of increased genetic knowledge to improve human health has been widely championed.[15–17] In a

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

Q: How do you know which type of sample to collect?

Different flows of research in genetic epidemiology require specific designs

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

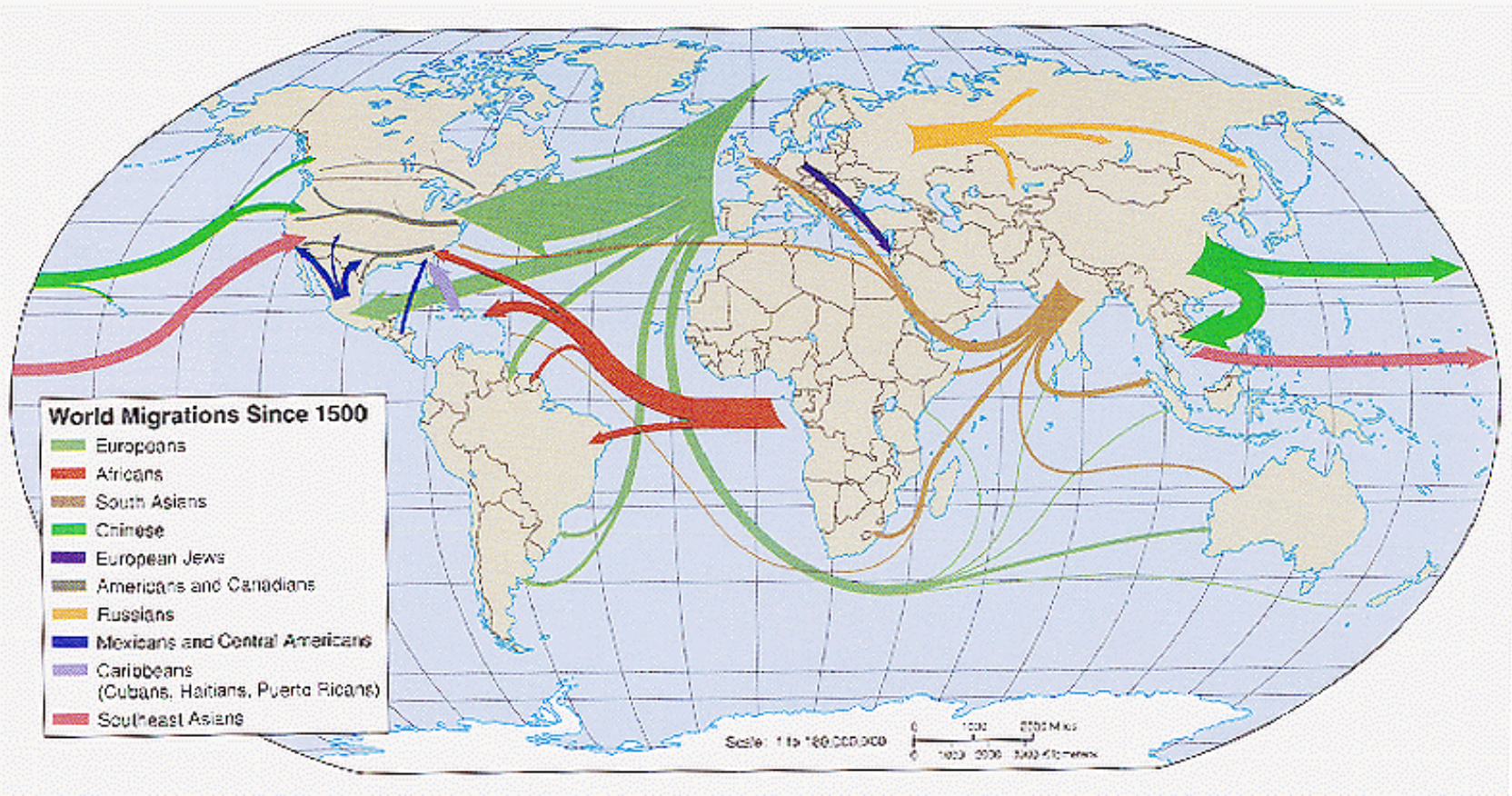
http://www.dorak.info/epi/genetepi.html

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?
  - Answering this question requires family-based, twin/adoption or migrant studies.

Migration studies: an unexpected role in genetic epidemiology?



(Weeks, Population. 1999)

Migration studies

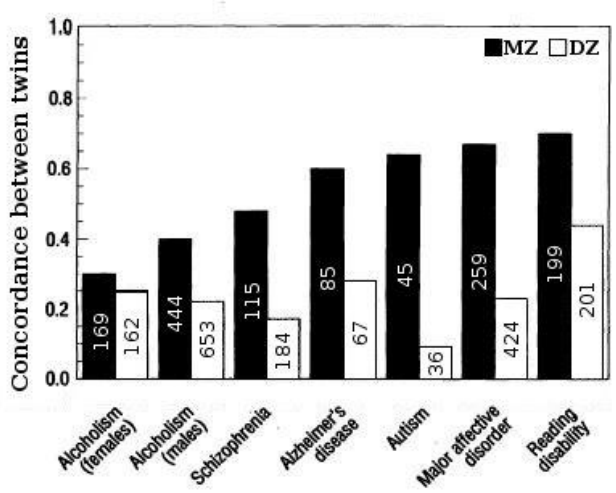
- As one of the initial steps in the process of genetic epidemiology, one could use information on populations who migrate to countries with different genetic and environmental backgrounds - as well as rates of the disease of interest - than the country they came from.
- Here, one compares people who migrate from one country to another with people in the two countries.
- If the migrants' disease frequency does not change –i.e., remains similar to that of their original country, not their new country—then the disease might have genetic components.
- If the migrants' disease frequency does change—i.e., is no longer similar to that of their original country, but now is similar to their new country—then the disease might have environmental components

Contribution of twins to the study of complex traits and diseases

- *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
- 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

- *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.

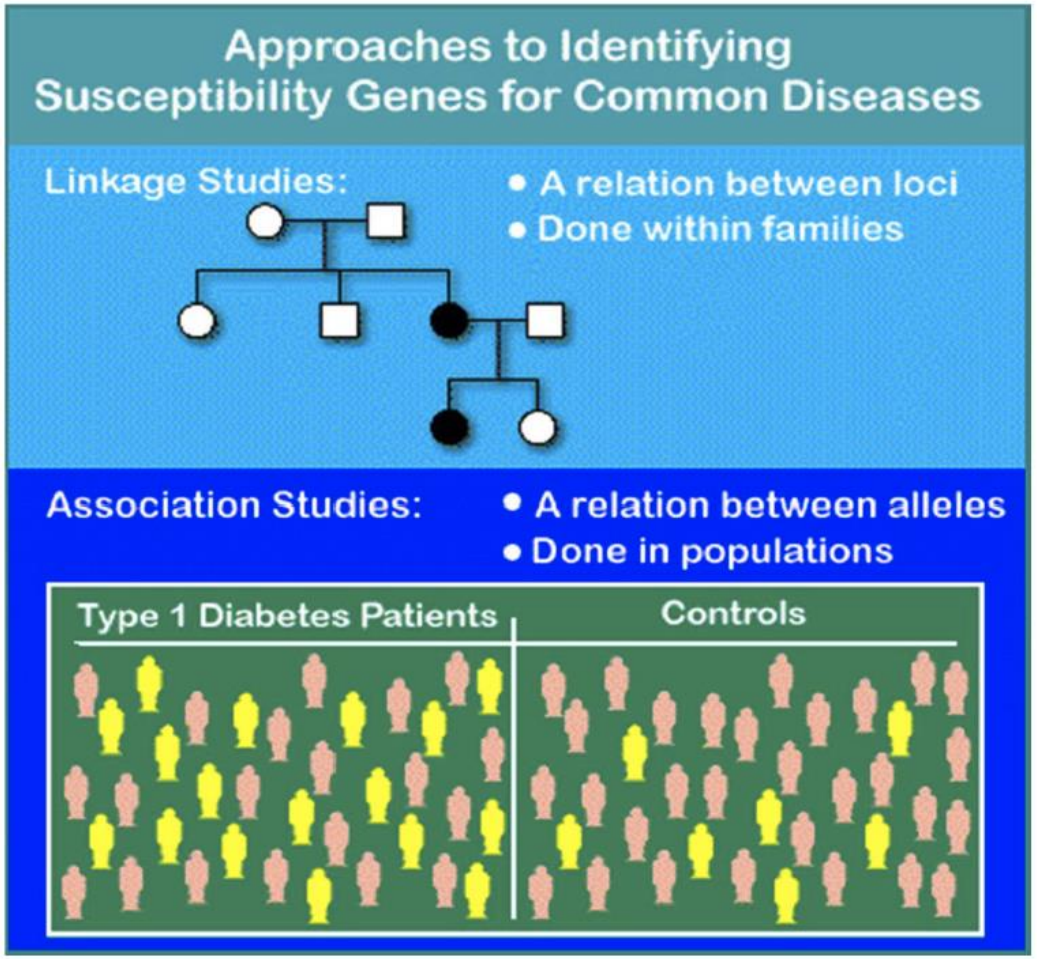


(http://en.wikipedia.org/wiki/File:Twin-concordances.jpg)

- **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).

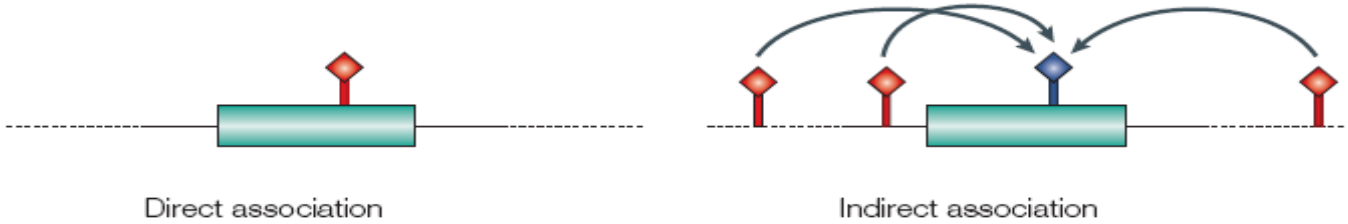
Linkage and Association

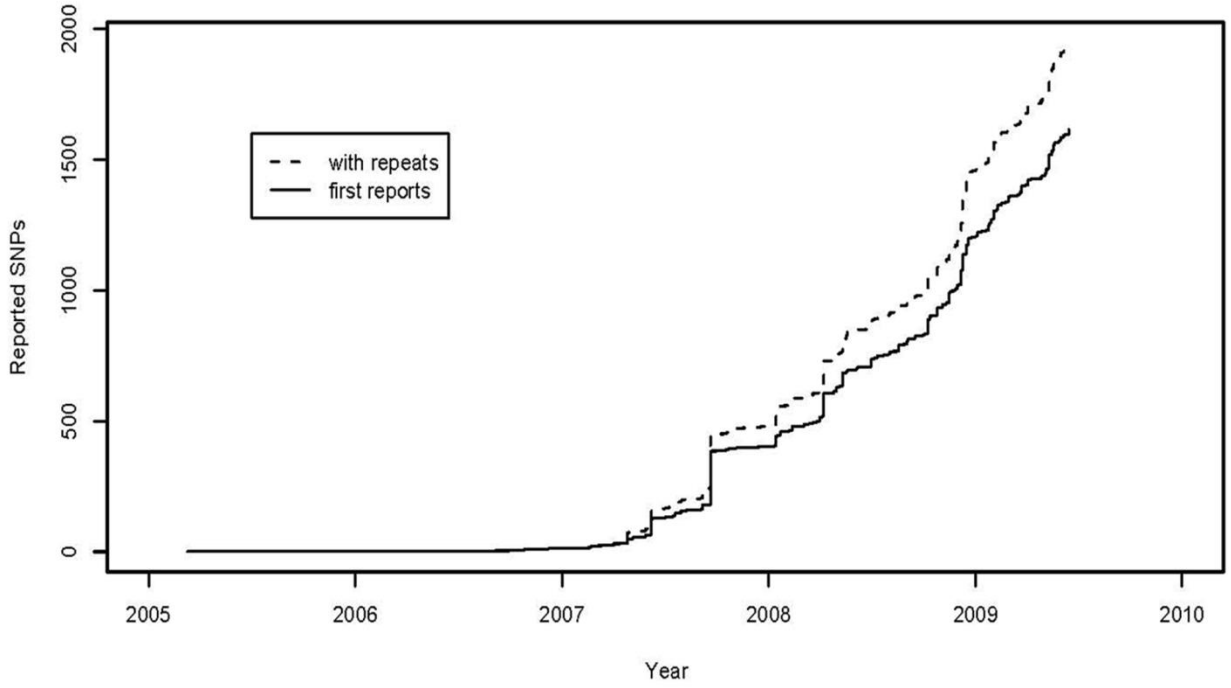


(Roche Genetics Education)

- **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.

Scaling up to “genome-wide” levels …





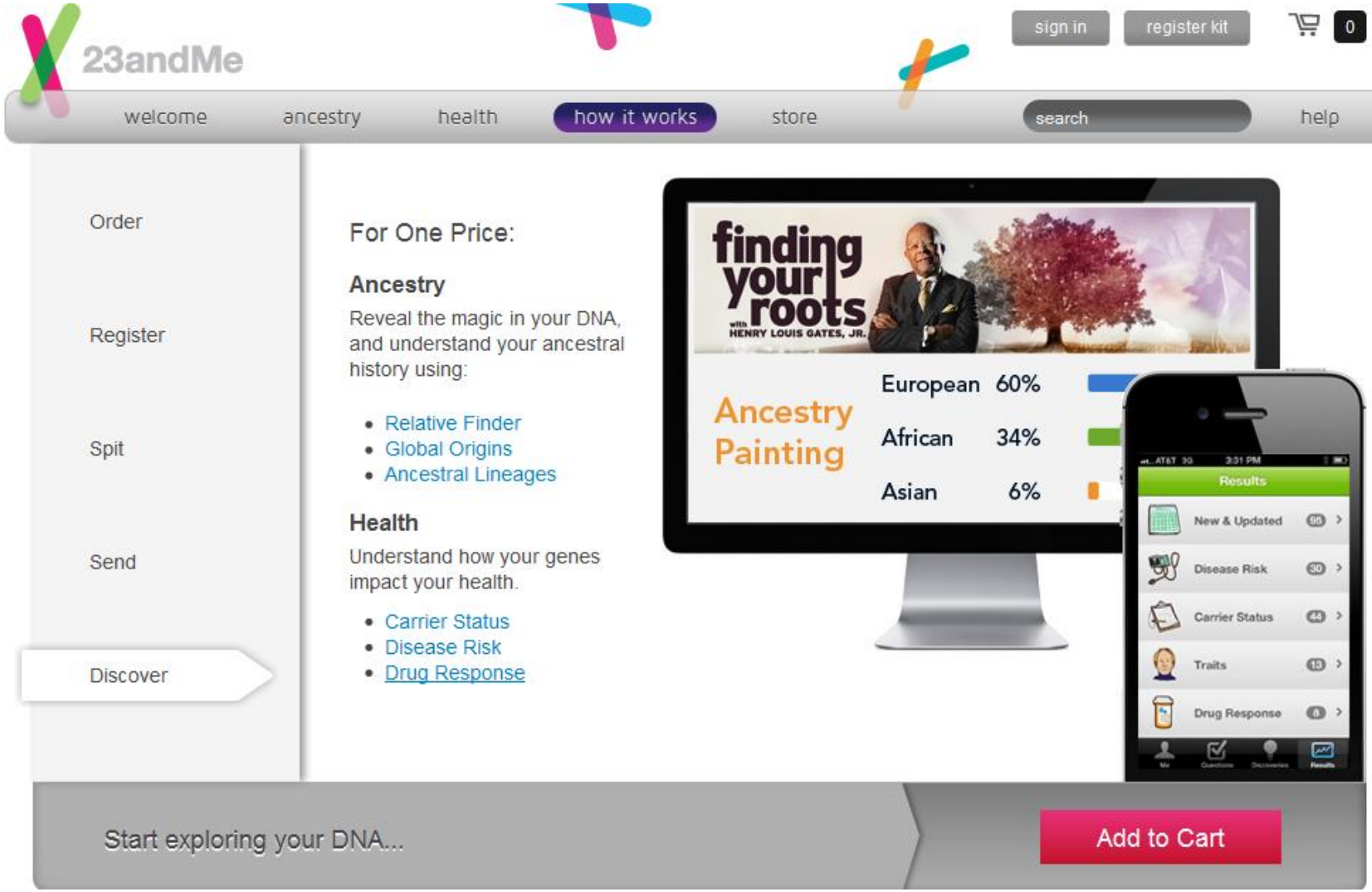
Top: Hirschhorn & Daly, Nat Rev Genet 2005; Bottom: Witte An Rev Pub Health 2009

Genetic testing based on GWA studies

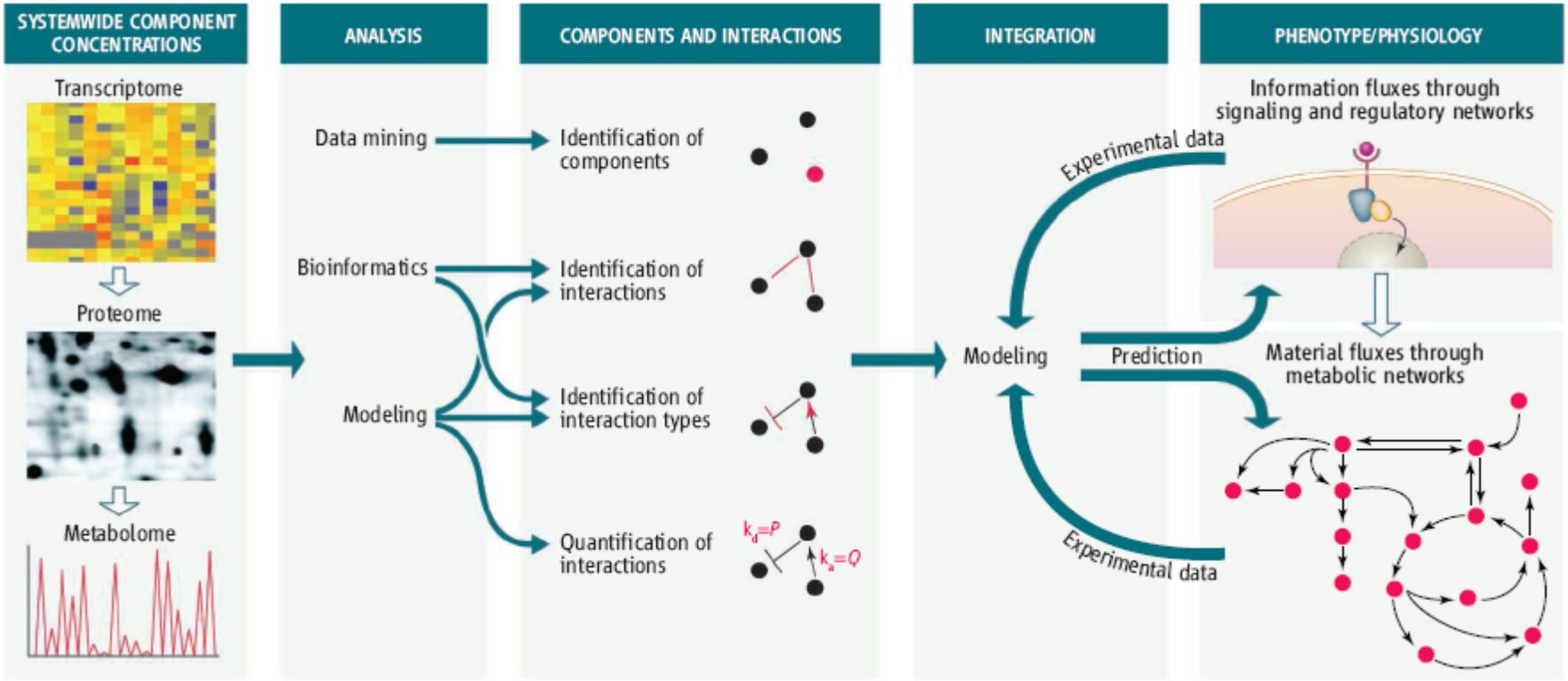
- Multiple companies marketing direct to consumer genetic ‘test’ kits.
- Send in spit.
- Array technology (Illumina / Affymetrix).
- Many results based on GWAS.
- Companies:
  - 23andMe 

  - deCODEme 

  - Navigenics


23andMe
sign in
register kit
0
welcome
ancestry
health
how it works
store
search
help
Order
Register
Spit
Send
Discover
For One Price:
Ancestry
Reveal the magic in your DNA, and understand your ancestral history using:
Relative Finder
Global Origins
Ancestral Lineages
Health
Understand how your genes impact your health.
Carrier Status
Disease Risk
Drug Response
finding your roots
with HENRY LOUIS GATES, JR.
Ancestry Painting
European 60%
African 34%
Asian 6%
Results
New & Updated
Disease Risk
Carrier Status
Traits
Drug Response
Start exploring your DNA...
Add to Cart

Getting closer to the whole picture



(Sauer et al, *Science*, 2007)

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

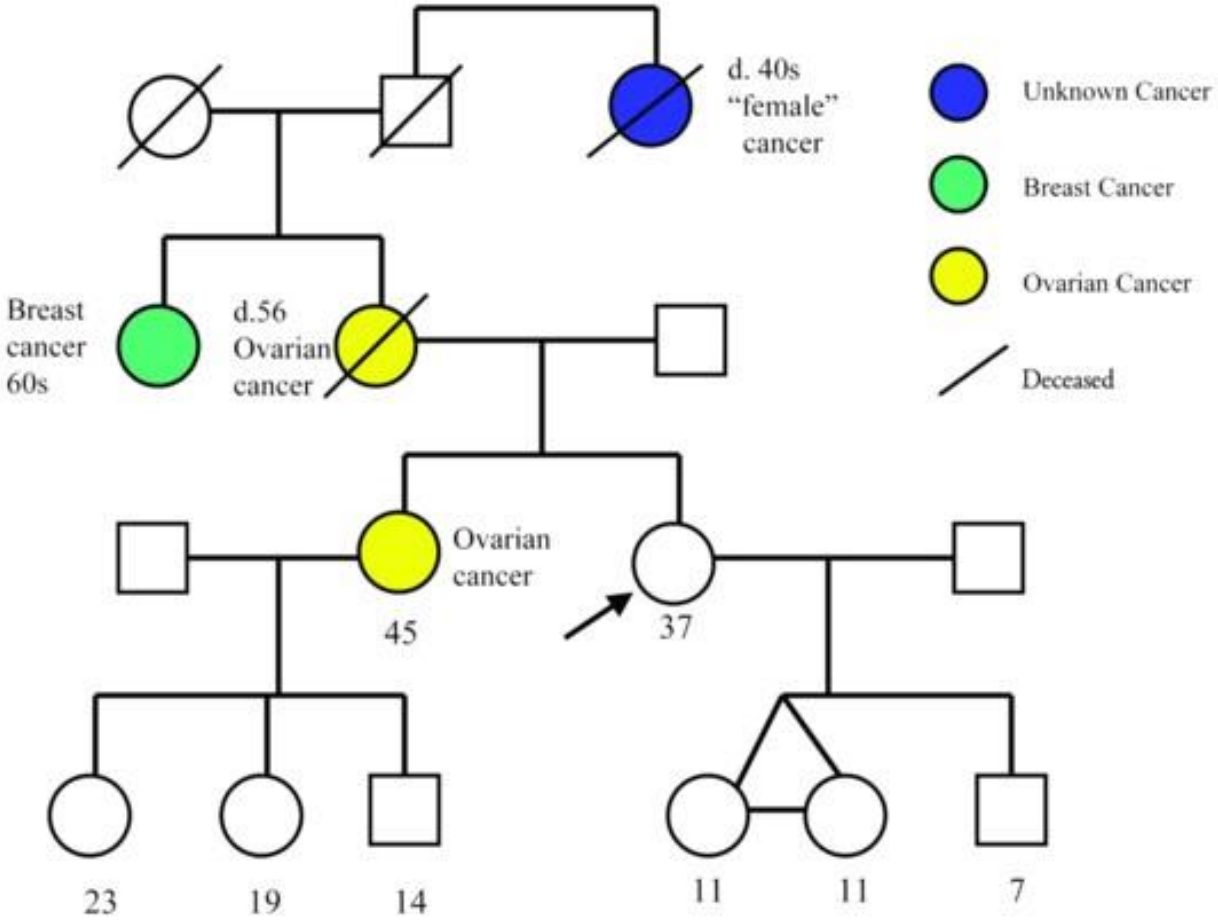
Aggregation and segregation studies in human genetics

- Aggregation and segregation studies are generally the first step when studying the genetics of a human trait.
- Aggregation studies evaluate the evidence for whether there is a genetic component to a study.
- They do this by examining whether there is familial aggregation of the trait.
- Questions of interest include:
  - Are relatives of diseased individuals more likely to be diseased than the general population?
  - Is the clustering of disease in families different from what you would expect based on the prevalence in the general population?

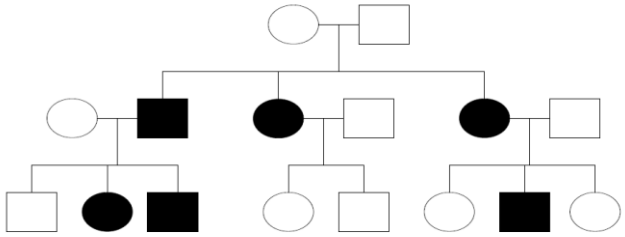
Definition of familial aggregation

- 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: “it runs in the family”
- 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.*

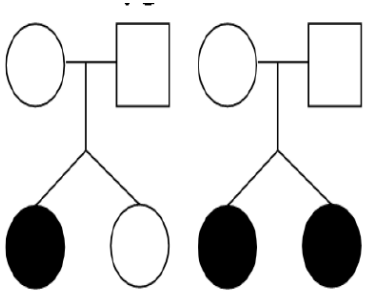
Example 1: does the phenotype run in the family?



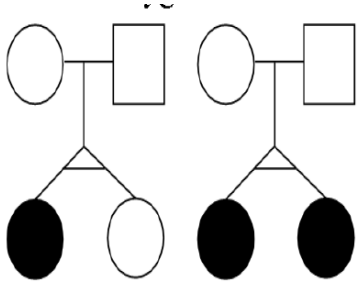
- **Pedigree** - A diagram of the genetic relationships and medical history of a family using standardized symbols and terminology
- **Founder** - Individuals in a pedigree whose parents are not part of the pedigree
- **Extended pedigrees**



Dizygotic twins

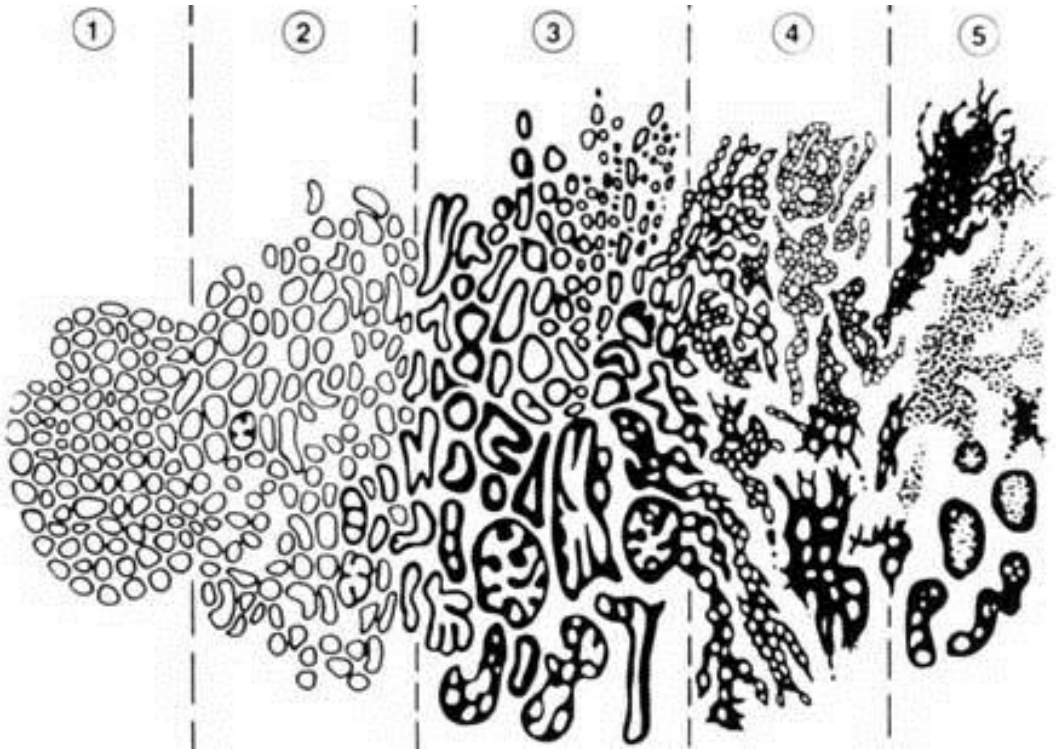


Monozygotic twins



Working with phenotypes

- Define the phenotype accurately. This is not always an easy task !!!



Gleason DF. In Urologic Pathology: The Prostate. 1977; 171-198

Example: Alzheimer's disease

- Studies based on twins have found differences in concordance rates between monozygotic and dizygotic twins. In particular, 80% of monozygotic twin pairs were concordant whereas only 35% of dizygotic twins were concordant. In a separate study, first-degree relatives of individuals (parents, offspring, siblings) with Alzheimer's disease were studied. First degree relatives of patients had a 3.5 fold increase in risk for developing Alzheimer's disease as compared to the general population. This was age-dependent with the risk decreasing with age-of-onset.

Reference: Bishop T, Sham P (2000) Analysis of multifactorial disease. Academic Press, San Diego

3.b Familial aggregation with quantitative traits

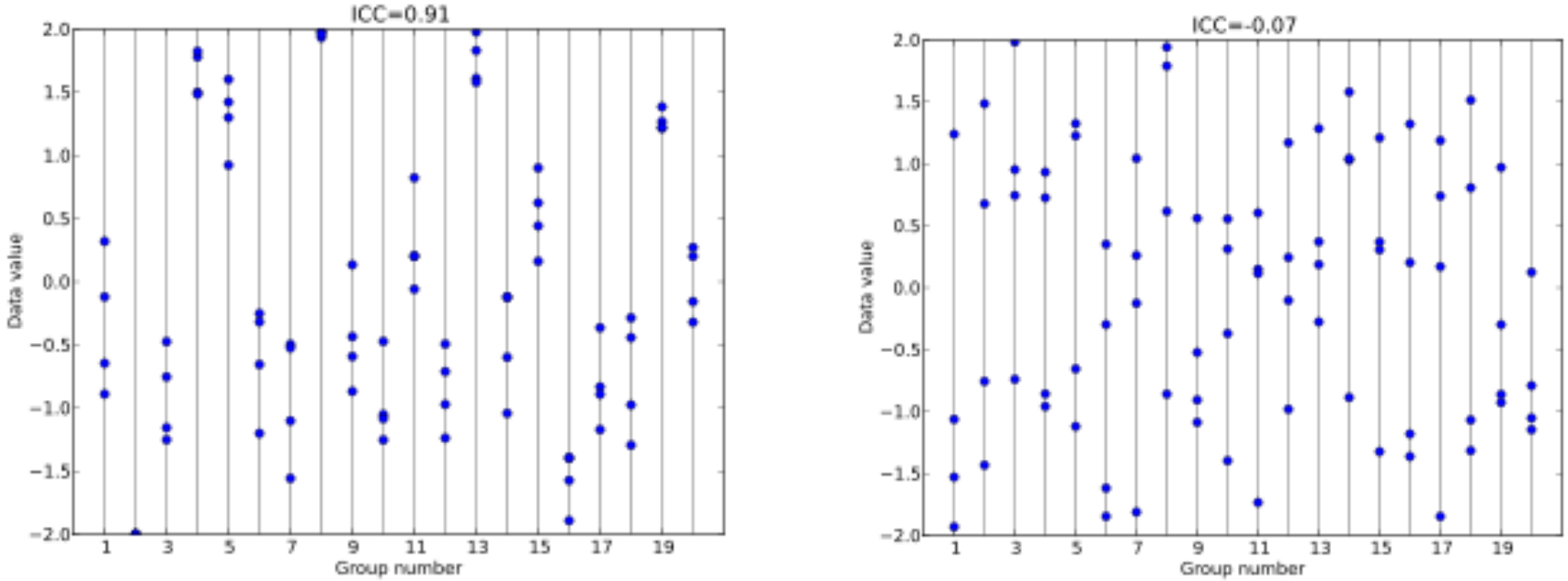
Proband selection

- For a continuous trait a random series of probands from the general population may be enrolled, together with their family members.
- Examples of such traits include blood pressure and height. Familial aggregation can be assessed using a correlation or covariance-based measure

Techniques

- The **intra-family correlation coefficient** (ICC) describes how strongly units in the same group resemble each other and can be interpreted as the proportion of the total variability in a phenotype that can reasonably be attributed to real variability between families
- Linear regression and multilevel modelling analysis of variance (non-random ascertainment unaccounted for can seriously bias ICC), *familial correlation coefficients* with FCOR in the Statistical Analysis for Genetic Epidemiology (SAGE) software package

Example



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

3.c Familial aggregation with dichotomous traits

Proband selection

- 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.
- In a **retrospective type** of analysis, the outcome of interest is disease in the proband. Disease in the relatives serves to define “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.

Techniques

- 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,

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

Techniques

- An alternative algebraic expression for the covariance is

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

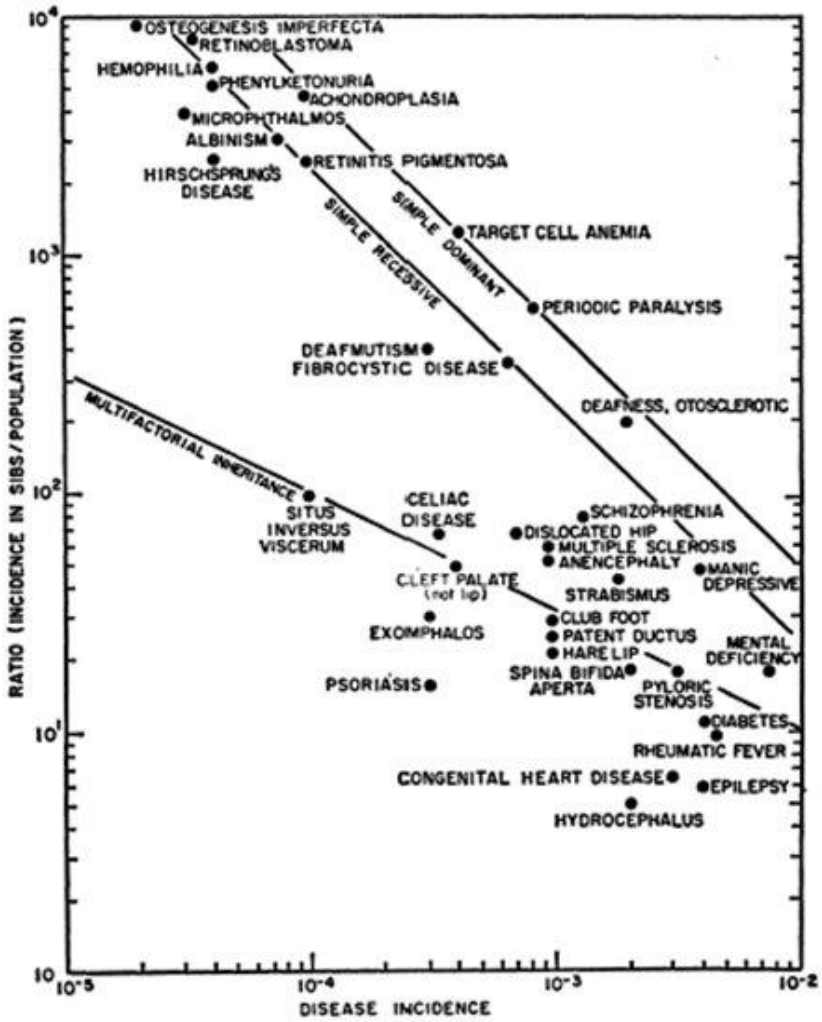
with Mean(Y_1Y_2) the probability that both relatives are affected. From this we derive for the familial risk ratio λ, defined before:

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

- 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)
- Estimates of conditional probabilities: regression with logit link function

Example

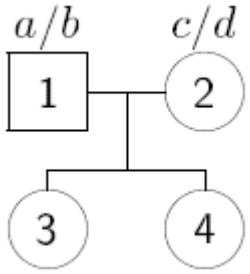
- For λ_S = ratio of risk in sibs compared with population risk.
  - cystic fibrosis: the risk in sibs = 0.25 and the risk in the population = 0.0004, and therefore λ_S =500
  - Huntington disease: the risk in sibs = 0.5 and the risk in the population = 0.0001, and therefore λ_S =5000
- Higher value indicates greater proportion of risk in family compared with population.
- Note that 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).

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 (4) 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:

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

- F.i.: z_0 = probability that none of the two alleles in the second relative are *identical by descent* (IBD), at the locus of interest, and conditional on the genetic make-up of the first relative

- Now, consider an allele 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 alleles are IBD.

Kinship coefficients (continued)

- Given there is no inbreeding (there are no loops in the pedigree graphical representation),
  - Under 2-IBD, *prob that two randomly selected alleles are IBD* = ½
  - Under 1-IBD, *prob that two randomly selected alleles are IBD* = ¼
  - Under 0-IBD, *prob that two randomly selected alleles are IBD* = 0
- So the kinship coefficient is

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

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

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

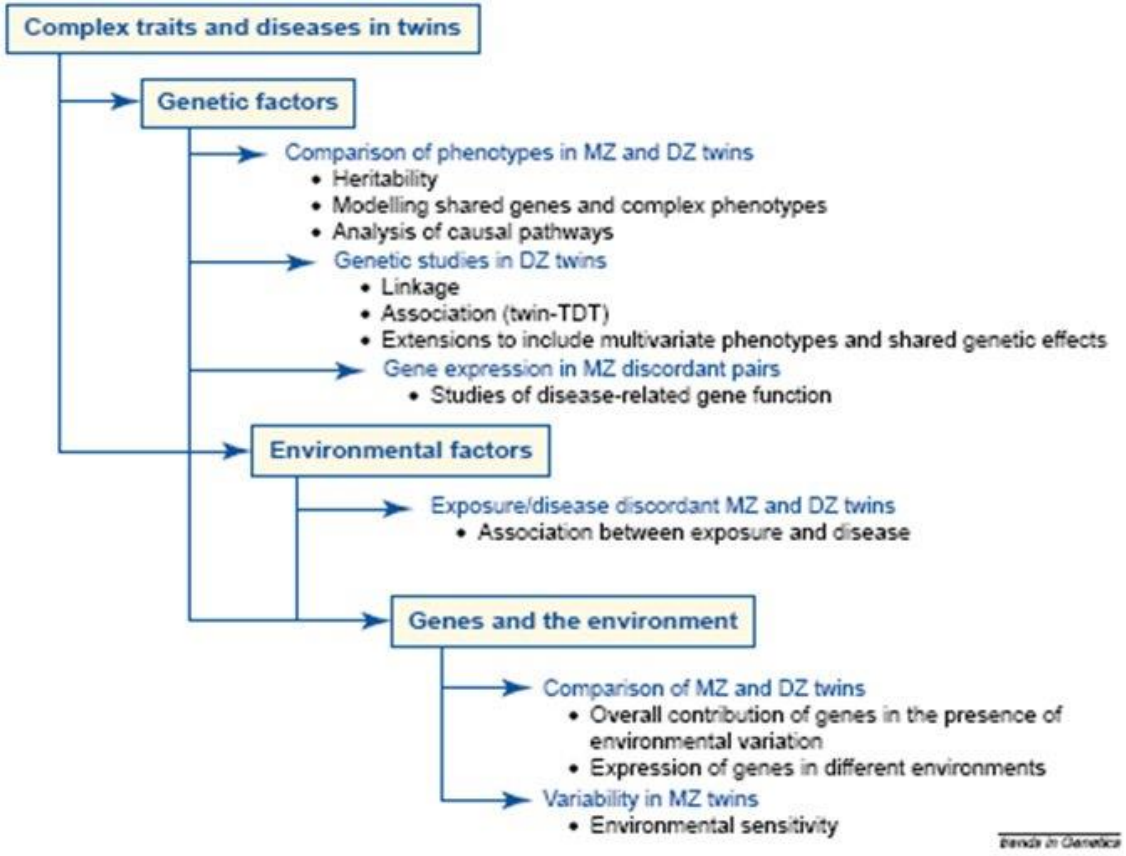
IBD sharing and kinship by relationship

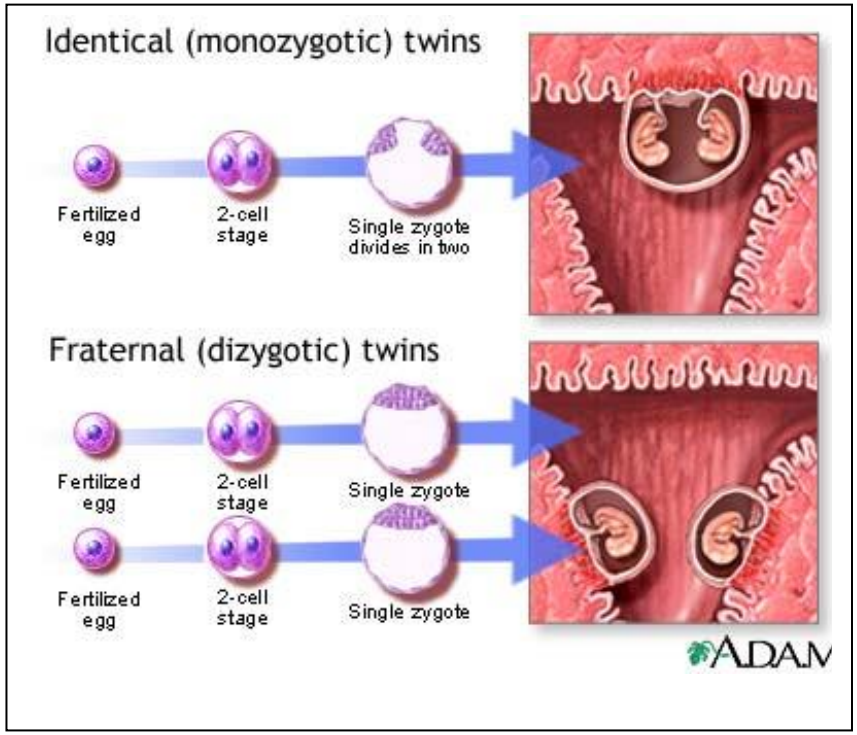
	No. alleles shared IBD			
	2	1	0	
Relationship	z_2	z_1	z_0	Φ
Self, MZ twins	1	0	0	1/2
Parent–Offspring	0	1	0	1/4
Full siblings	1/4	1/2	1/4	1/4
Half siblings	0	1/2	1/2	1/8
Uncle–nephew	0	1/2	1/2	1/8
Double 1st cousins	1/16	6/16	9/16	1/8
Grandchild–grandparent	0	1/4	3/4	1/16
First cousins	0	1/4	3/4	1/16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

(assuming no inbreeding)

- **Technique** : see before SAGE or R package GenABEL (pkin, in contrast to gkin)

3.e Quantifying genetics versus environment





Interpretation and follow-up of familial aggregation analysis results

- The presence of familial aggregation can be due to many factors, including shared family environment; Familial aggregation alone is not sufficient to demonstrate a genetic basis for the disease.
- Methods exist to estimate the proportion of phenotypic variance that is due to genetics (linked to concepts of “heritability”)
- In general, when wishing to decompose trait variance into
  - Genetic variance
  - Shared environmental variance
  - Unique environmental variance

  a twin design can be used.

Heritability

- We can measure the variance in a trait (call it variance in liability, L, and assume that it corresponds to a normally distributed variable) as a mixture of different effects: variance due to genetics (which we will call A, for "additive"), and variation due to environment; $L = A + E$
- The **heritability**, which is called h^2 is the proportion of the total variance that is genetic, and therefore $h^2 = A/(A + E)$
- As both genetics and environment vary between families, the variance between families is A + E. We can measure A from identical (monozygotic, or MZ) twins, by assuming that they have perfectly correlated genetics, but non-correlated environment, so the shared variance (the Covariance) is A

$$h^2 = [\text{covariance within MZ twinships}]/[\text{variance between families}]$$

- So far, we have assumed that MZ twins do not share a common environment; this is a bad assumption, because often they will. So, instead, we model the liability as having some shared environmental component C (for common), so that L = A + E + C
- Assuming monozygotic and dizygotic twins share the same environment, the covariance between monozygotic twins is A + C, and between dizygotic twins is 0.5 x A + C (as they have the same environment, but half the same DNA).
- We can thus recalculate the heritability as follows:

$$h2 = A / (A + C + E)$$

$$= 2 \times ([A + C] - [0.5\ A + C]) / (A + C + E)$$

$$= 2 \times ([\text{Covariance within MZs}] - [\text{Covariance within DZs}]) / [\text{Variance between families}]$$

Heritability questions

- What if we have a dichotomous trait and cannot assume a normal distribution?
  - In this case we can use liability threshold modeling
- How accurate are these estimates?
  - Error bars from twin studies for rare diseases tend to be pretty large, due to the inability to find enough twins with the disease. For example, in Crohn's disease (a common disease!) we generally find error bars that place h^2 between 40 and 80%
- How are heritability estimates used in practice?
  - They may indicate best case scenarios for prediction
  - They are used in estimates about how much of the genetic effect (A) we have accounted for with our GWAS results (see later)
  -

Missing heritability

- For virtually all diseases we find that the majority of genetic risk is still left undiscovered….

(Maher 2008)

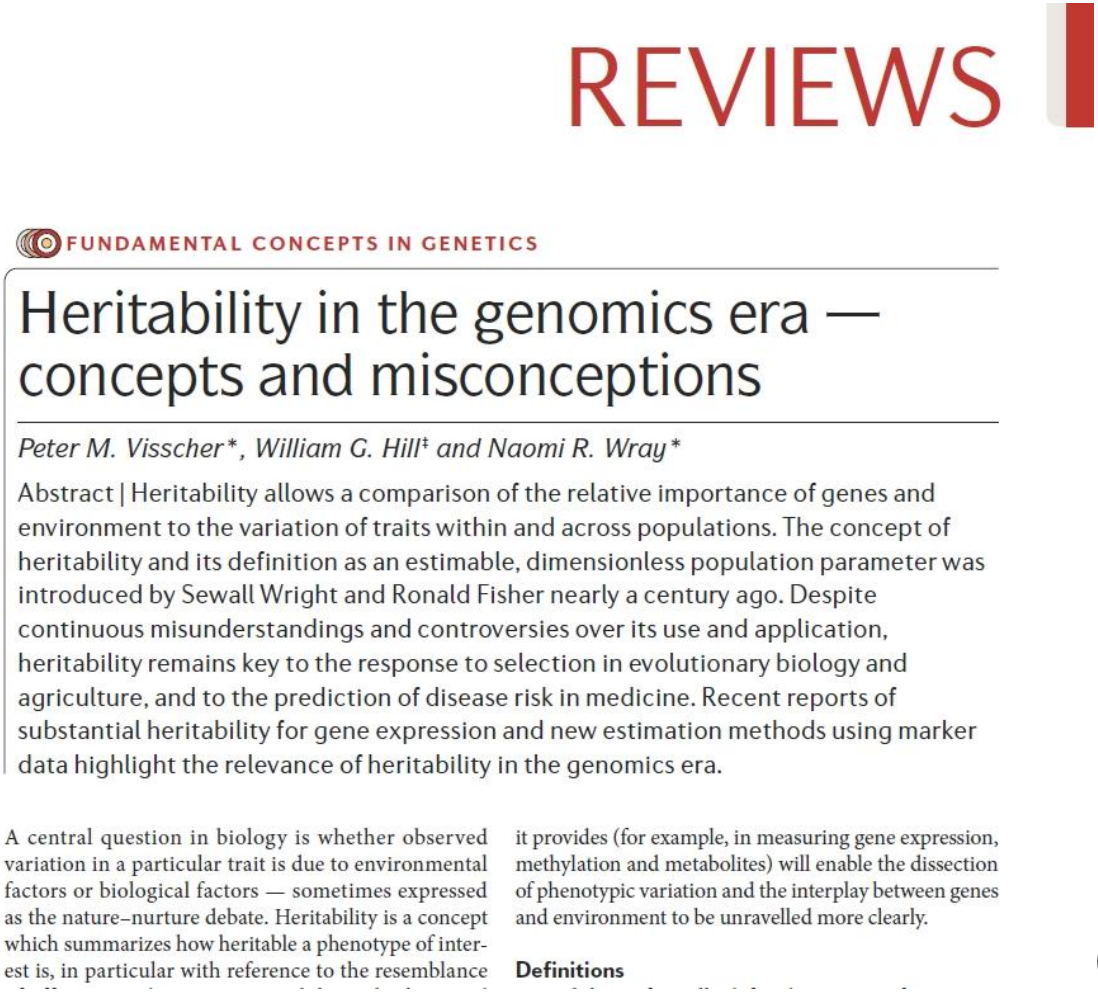
Missing heritability

- Are unreasonable assumptions made regarding estimating heritability?
  - We assume MZ twins share no environment that DZ twins do not also share (MZ: shared placenta, different social environment than DZ?)
  - We assume that we can disregard gene/environment interaction, which can have complicated twin-sharing properties
  - We assume that DZ twins share half the genetic effect, i.e. no gene-gene interactions occur. If this is false, heritability can be overestimated.

    In fact: The genetic variance can be partitioned into the variance of additive genetic effects (breeding values; σ_A^2), of dominance (interactions between alleles at the same locus) genetic effects (σ_D^2), and of epistatic (interactions between alleles at different loci) genetic effects (σ_I^2)

Background reading

(http://genomesunzipped.org/2010/12/estimating-heritability-using-twins.php)



REVIEWS

FUNDAMENTAL CONCEPTS IN GENETICS

Heritability in the genomics era — concepts and misconceptions

Peter M. Visscher, William G. Hill‡ and Naomi R. Wray**

Abstract | Heritability allows a comparison of the relative importance of genes and environment to the variation of traits within and across populations. The concept of heritability and its definition as an estimable, dimensionless population parameter was introduced by Sewall Wright and Ronald Fisher nearly a century ago. Despite continuous misunderstandings and controversies over its use and application, heritability remains key to the response to selection in evolutionary biology and agriculture, and to the prediction of disease risk in medicine. Recent reports of substantial heritability for gene expression and new estimation methods using marker data highlight the relevance of heritability in the genomics era.

A central question in biology is whether observed variation in a particular trait is due to environmental factors or biological factors — sometimes expressed as the nature–nurture debate. Heritability is a concept which summarizes how heritable a phenotype of interest is, in particular with reference to the resemblance

it provides (for example, in measuring gene expression, methylation and metabolites) will enable the dissection of phenotypic variation and the interplay between genes and environment to be unravelled more clearly.

Definitions

(Visscher et al. 2008)

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?

Introduction

- Segregation analysis moves beyond aggregation of disease and seeks to more precisely identify the factors responsible for familial aggregation.
- For instance:
  - Is the aggregation due to environmental, cultural or genetic factors?
  - What proportion of the trait is due to genetic factors?
  - What mode of inheritance best represents the genetic factors?
  - Does there appear to be genetic heterogeneity?

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 (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
- **Technique:**

  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.

Example: 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 (i.e. non-sex chromosome).
- Let D be the allele causing the disorder and let d represent be the normal allele.
- There are 9 possible mating types (can collapse to six mating types due to symmetry): for instance DDxdd
- 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 and can be used to formally test whether a disease is caused by a single autosomal dominant gene

Example: 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.
- Questions:
  - 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

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: $X \sim B(n, p)$

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

- 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

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

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

The binomial distribution applied to Marfan

- 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 sample
- 52 of the offspring are affected and 60 are unaffected
- Are these observations consistent with an autosomal dominant disease.
- The p-value is

$$= \left(\frac{1}{2}\right)^{112-1} \sum_{x=0}^{52} \binom{112}{x} = 0.5085$$

- What if only 42 of the offspring are affected?

$$= \left(\frac{1}{2}\right)^{112-1} \sum_{x=0}^{42} \binom{112}{x} = 0.0104$$

Normal approximation to the binomial applied to Marfan

- If $X \sim B(n, p)$, and n is large enough such that

$$np \geqslant 10 \qquad \text{and} \qquad n(1-p) \geqslant 10$$

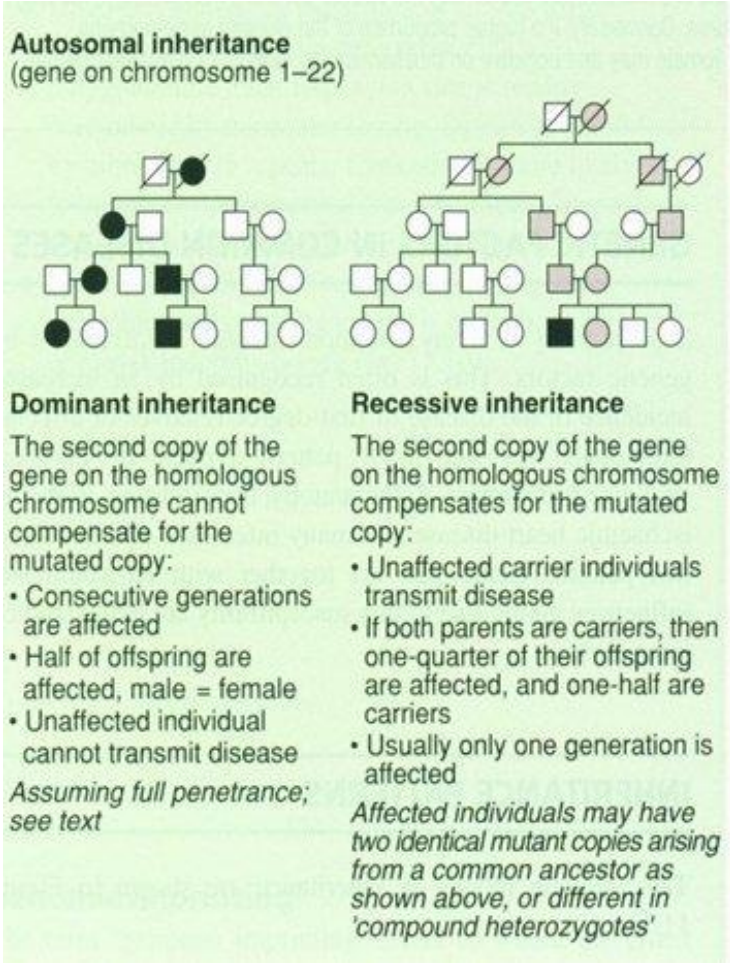
- Then X is approximately $N\left(\mu_X = np, \sigma_X = \sqrt{np(1-p)}\right)$
- For the Marfan syndrome data with 52 offspring affected,

$$z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{52.5 - (112)(.5)}{\sqrt{112(.5)(.5)}} = -.661$$

P-value is $2P(Z \geqslant |z|) = 2(0.2539) = .5079$, where Z follows a standard normal distribution

- For the Marfan syndrome data with 42 offspring affected, the p-value is .0107.

Modes of inheritance



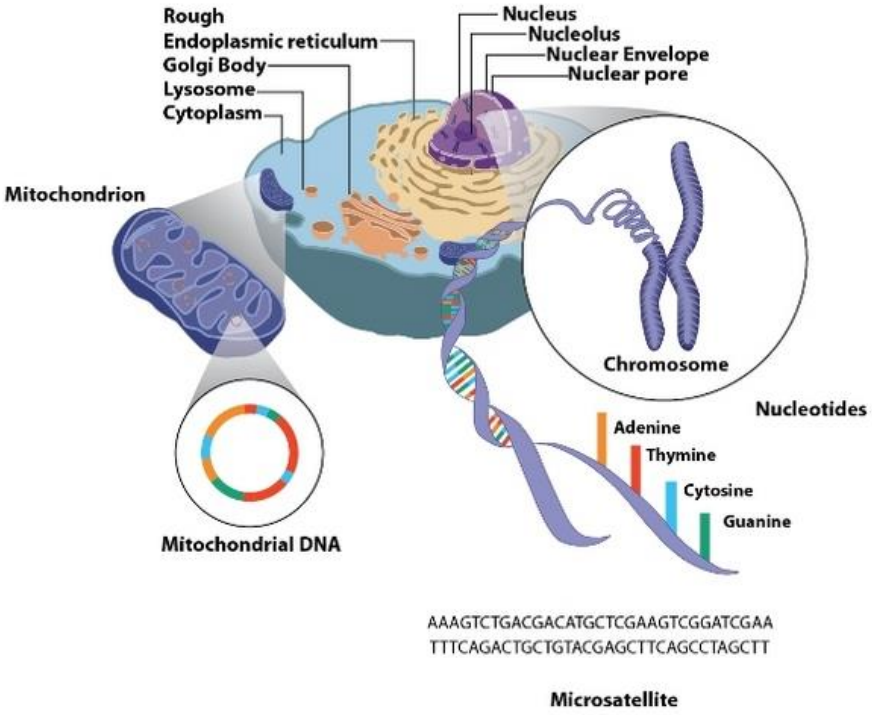
Left: single gene and Mendelian inheritance

Increasing levels of complexity:

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

(See also Roche Genetics)

Mitochondrial DNA

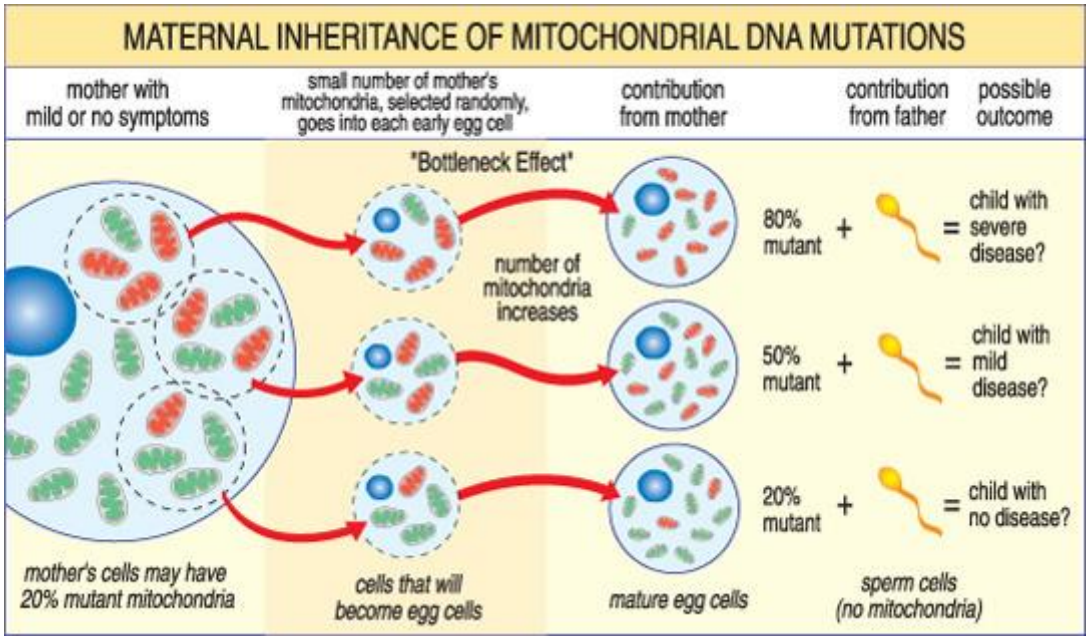


(http://www.nature.com/scitable/knowledge/library/)

- Mitochondrial DNA (mtDNA) is the DNA located in the mitochondria, structures within eukaryotic cells that convert the chemical energy from food into a form that cells can use, adenosine triphosphate (ATP). Most of the rest of human DNA present in eukaryotic cells can be found in the cell nucleus. In most species, including humans, mtDNA is inherited solely from the mother (i.e., maternally inherited).

Mitochondrial DNA

- In humans, mitochondrial DNA can be regarded as the smallest chromosome coding for only 37 genes and containing only about 16,600 base pairs.
- Human mitochondrial DNA was the first significant part of the human genome to be sequenced.



(http://mda.org/disease/mitochondrial-myopathies/causes-inheritance)

Complex segregation analysis

- For more complicated structures, segregation models are generally 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 (i.e., Prob(Disease | Genotype))
  - the population genotype
  - the transmission probabilities within families
  - the method of ascertainment
- Especially for extended pedigrees (multiple generations) a numerical procedure is needed for all probability calculations involved.

Segregation analysis involves computing (often very complicated!) probabilities

- 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, $i = 1,\ldots,N$, where *N* denotes the number of individuals in the pedigree:

$$L(Y) = \sum_{g1} \sum_{g2} \cdots \sum_{gN} P(Y|g_1 g_2 \cdots g_N) P(g_1 g_2 \cdots g_N) \ .$$

- Widely used in segregation analysis is the **Elston–Stuart** peeling algorithm (Elston and Stuart 1971), a recursive formula for the computation of the likelihood *L* given as

$$L = \sum_{g1} \sum_{g2} \cdots \sum_{gN} \prod_{j=1}^{N} f(g_j) \prod_{k=1}^{N_1} P(g_k) \prod_{m=1}^{N_2} \tau(g_m|g_{m1} g_{m2}) \ .$$

(Bickeböller – Genetic Epidemiology)

Background information about the formula

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 $f\ (g_i)$, $i = 1,\ldots,N$, parametrize the genotype-phenotype correlation for each individual i.

4.b Genetic models

From easy to complex modes of inheritance

- Single major locus: Simple Traits / Diseases
  - Dominant model
  - Recessive model
  - Additive
  - Multiplicative
- Multifactorial/polygenic: Complex Traits / Diseases
  - Multifactorial (many factors)
  - Polygenic (many genes)
  - General assumption: each of the factors and genes contribute a small amount to phenotypic variability
- Mixed model - single major locus with a polygenic background

Single major locus

- **Monogenic diseases** are those in which defects in a single gene produce disease. Often these disease are severe and appear early in life, e.g., cystic fibrosis. For the population as a whole, they are relatively rare. In a sense, these are pure genetic diseases: They do not require any environmental factors to elicit them. Although nutrition is not involved in the causation of monogenic diseases, these diseases can have implications for nutrition. They reveal the effects of particular proteins or enzymes that also are influenced by nutritional factors

  (http://www.utsouthwestern.edu)

- In this scenario, a single gene, usually assumed to have only 2 alleles, contributes to the phenotypic variability.

Binary traits (where an individual can be either affected or unaffected)

- q_1 = frequency of allele increasing risk of disease, where $q_1 + q_2 = 1$
- Penetrance parameters
  - f_{11} = probability of being affected given 11 genotype
  - f_{12} = probability of being affected given 12 genotype
  - f_{22} = probability of being affected given 22 genotype
- K_p=population prevalence of the disease
- $K_p = q_1^2 f_{11} + 2q_1 q_2 f_{12} + q_2^2 f_{22}$
- Genotype Relative Risk - It is common to represent the risk of a genetic variants relative to the average population
  - $R_{11} = \frac{P(\text{affected}|11)}{K_p} = \frac{f_{11}}{K_p}$
  - $R_{12} = \frac{f_{12}}{K_p}$
  - $R_{22} = \frac{f_{22}}{K_p}$

Penetrance parameters

- The penetrance parameters determines the model type
- Consider the following parameterization
  - $f_{11} = k$
  - $f_{12} = k - c_{12}$
  - $f_{22} = k - c_{22}$

  where $k - 1 \leqslant c_{12} \leqslant k$ and $k - 1 \leqslant c_{22} \leqslant k$, with $0 \leqslant k \leqslant 1$, $c_{12} \geqslant 0$, and $c_{22} \geqslant 0$
- What is the relationship between c_{12} and c_{13} for an additive model?
- What are the parameter values for a fully penetrant dominant disease?
- Note that if both $c_{12} = 0$ and $c_{22} = 0$, then the locus is not involved with the phenotype, and k would be equal to K_p.

Another example: penetrance parameters determine model type

A multiplicative model is given below

- $f_{11} = r^2 k$
- $f_{12} = rk$
- $f_{22} = k$

where with $0 \leqslant k \leqslant 1$, $r \geqslant 1$, and $0 \leqslant r^2 k \leqslant 1$

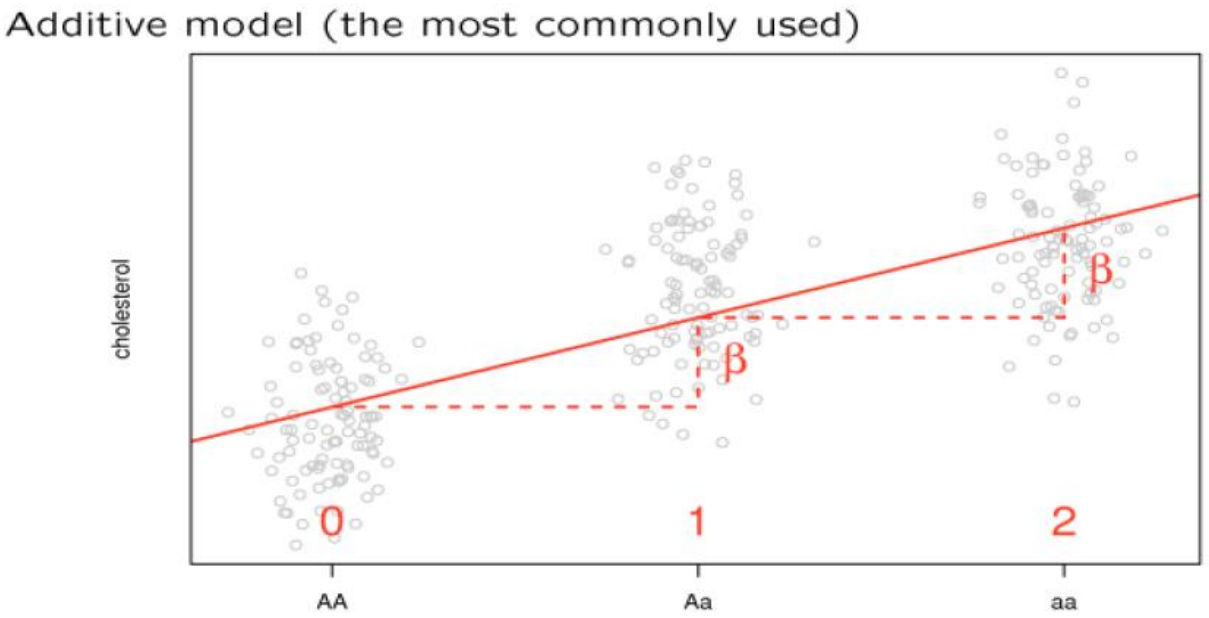
Codominant genetic model: If the risk conferred by the heterozygote individuals lies between that of wildtype homozygote and minor allele homozygote individuals, but not in the specific relationship of a multiplicative or additive model (Lewis, 2002; Minelli, 2005). This model is the most powerful one (over additive, recessive or dominant) to detect associations when the inheritance model is not known (Lettre, 2007).

Quantitative traits

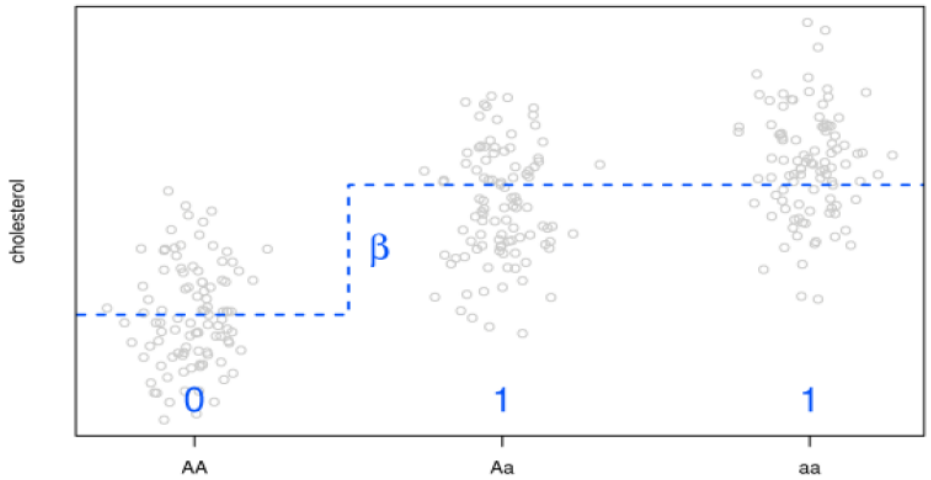
- For a quantitative trait, Y, the penetrance function describes the distribution of the trait conditional on an individual's genotype, $f(Y|\text{genotype})$.
- Location of the heterozygote mean determines whether the allele increasing susceptibility to the disease or increasing the value of the phenotype is dominant, additive, recessive, or etc.

Technique

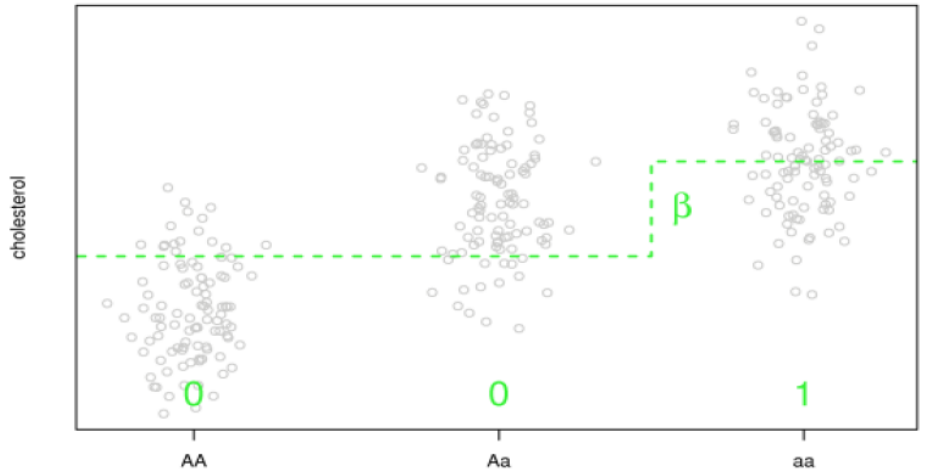
- Regression framework : e.g., logistic regression for binary traits and linear regression for quantitative traits. Depending on the coding of the "genetic effect" a particular genetic model is implicitly assumed



Dominant model (best fit to this data)



Recessive model (least stable for rare aa)



Multiple loci

- **Oligogenic diseases** are conditions produced by the combination of two, three, or four defective genes. Often a defect in one gene is not enough to elicit a full-blown disease; but when it occurs in the presence of other moderate defects, a disease becomes clinically manifest. It is the expectation of human geneticists that many chronic diseases can be explained by the combination of defects in a few (major) genes.

- A third category of genetic disorder is **polygenic disease**. According to the polygenic hypothesis, many mild defects in genes conspire to produce some chronic diseases. To date the full genetic basis of polygenic diseases has not been worked out; multiple interacting defects are highly complex!!!

(http://www.utsouthwestern.edu)

- **Complex diseases** refer to conditions caused by many contributing factors. Such a disease is also called a multifactorial disease.
  - Some disorders, such as sickle cell anemia and cystic fibrosis, are caused by mutations in a single gene.
  - Common medical problems such as heart disease, diabetes, and obesity likely associated with the effects of multiple genes in combination with lifestyle and environmental factors, all of them possibly interacting.

4.c Genetic heterogeneity

What's in a name?

- **Allelic heterogeneity**: In some instances different alleles at the same locus cause the same disorder, a situation called allelic heterogeneity. A notable example is cystic fibrosis, where more than 600 different alleles can cause the associated symptoms.
- **Locus heterogeneity**: Contrast allelic heterogeneity with a situation where mutations in genes at different loci cause the same disease. An example of this locus heterogeneity is familial hypercholesterolemia, a single-gene disorder that causes very high cholesterol levels and high risk for coronary artery disease. Mutations in the APOB and LDLR genes are the most common cause of familial hypercholesterolemia, though other genes have been implicated.

- **Epistasis**: Sometimes the products of one gene mask or alter the expression of one or more other genes, a phenomenon called epistasis. In humans, a classic example is the mutation that causes albinism. The expression of that variant overrides the expression of other genes that control pigmentation, including those associated with eye and hair color. In more common examples, researchers are finding that epistasis plays a role in increasing or decreasing risk for the development of a wide array of cancers, Alzheimer disease, and cardiovascular disease. The extent of epistatic heterogeneity needs further research.

In contrast:

- **Pleiotropy**: Cystic fibrosis is a good of example of pleiotropy, where a mutation in a single gene affects multiple systems in this case the lungs, pancreas, and sweat glands.

(http://www.nchpeg.org/nutrition)