

DIFFERENT FACES OF GENETIC EPIDEMIOLOGY

Kristel Van Steen, PhD² (*)

kristel.vansteen@uliege.be

(*) GIGA-R Medical Genomics, Systems Genetics Lab, University of Liège, Belgium

Systems Medicine Lab, KU Leuven, Belgium

DIFFERENT FACES OF GENETIC EPIDEMIOLOGY

1 Basic epidemiology

1.a Aims of epidemiology

1.b Designs in epidemiology

1.c Genetics – a primer

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

5 A bird's eye view on genetic epidemiology

5.a Selected topics in genetic epidemiology

5.b Interesting reads

1 Basic epidemiology

Main references:

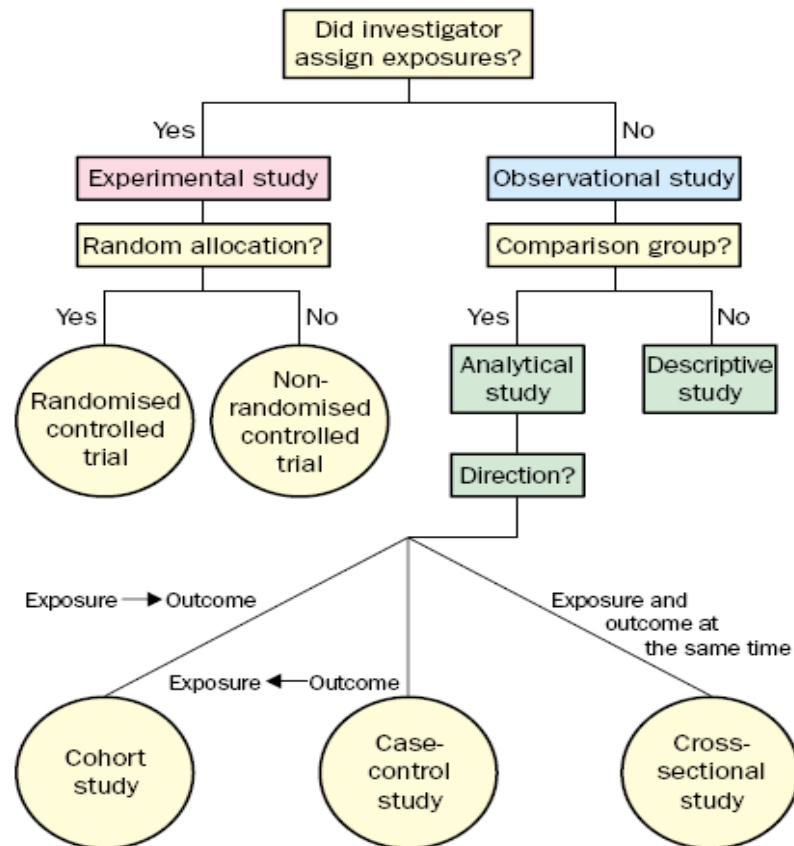
- Burton P, Tobin M and Hopper J. Key concepts in genetic epidemiology. The Lancet, 2005
- Clayton D. Introduction to genetics (course slides Bristol 2003)
- Bonita R, Beaglehole R and Kjellström T. *Basic Epidemiology*. WHO 2nd edition
- URL:
 - <http://www.dorak.info/>

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)

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

(Grimes & Schulz 2002)

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

(Grimes & Schulz 2002)

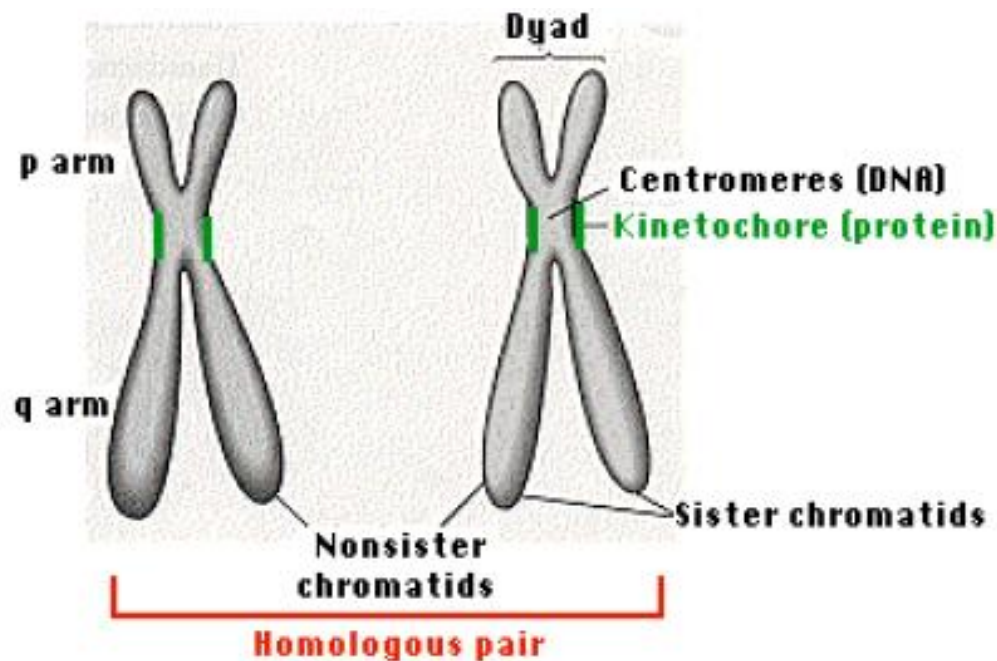
1.c Genetics – a primer

The structure of chromosomes

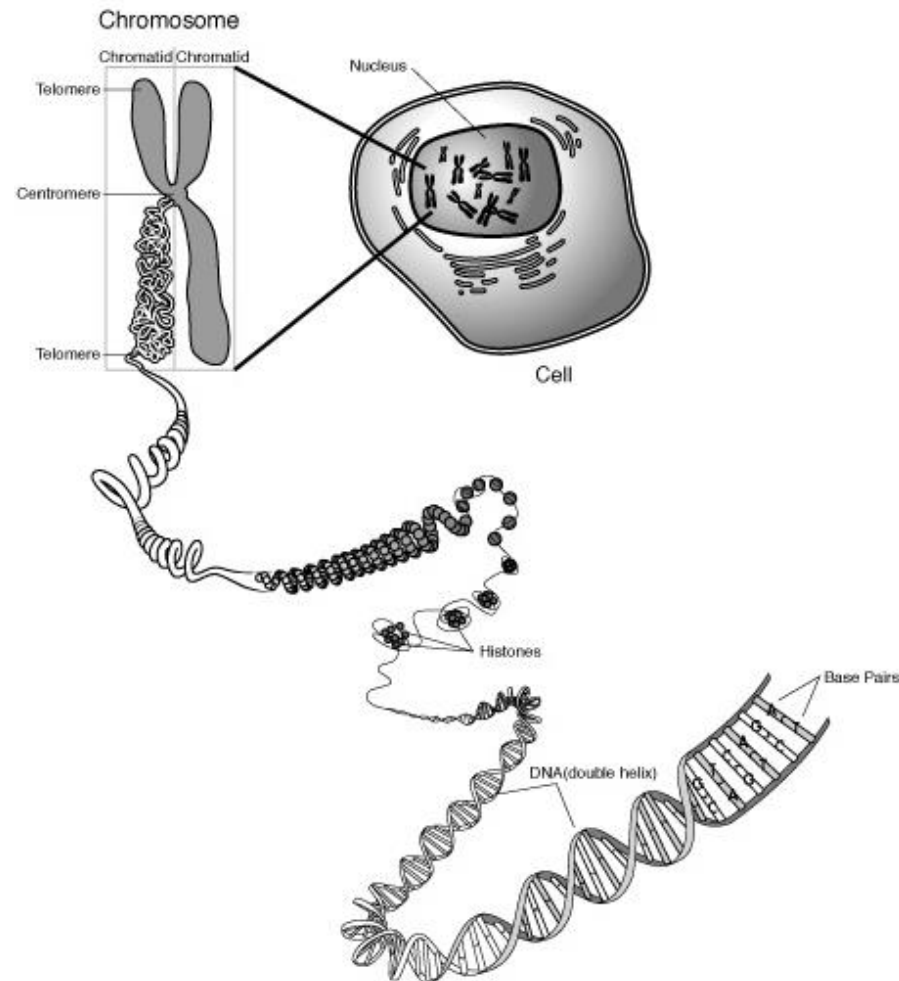
- In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called **chromosomes**. Each chromosome is made up of DNA tightly coiled many times around proteins called *histones* that support its structure.
- Chromosomes are not visible in the cell's nucleus—not even under a microscope—when the cell is not dividing.
- However, the DNA that makes up chromosomes becomes more tightly packed during cell division and is then visible under a microscope → Most of what researchers know about chromosomes was learned by observing chromosomes during cell division.

Chromosomes and chromatids

- A chromatid is one among the two identical copies of DNA making up a replicated chromosome, which are joined at their centromeres, for the process of cell division



Histones: packaging of DNA in the nucleus

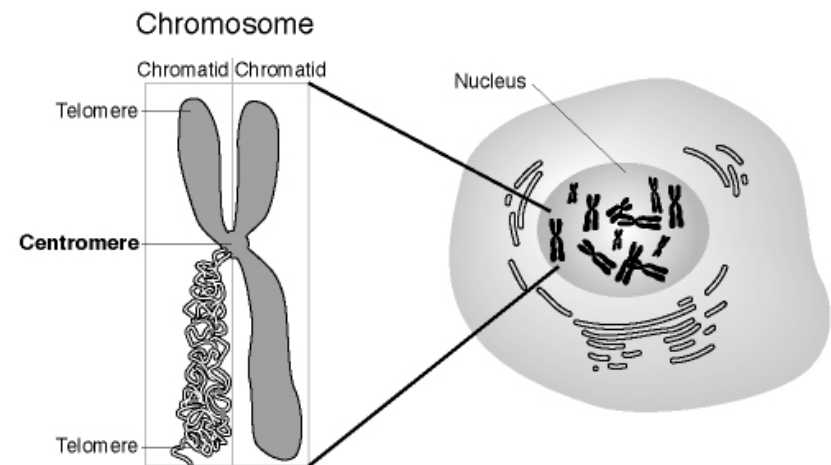


<http://www.accessexcellence.org/AB/GG/chromosome.html>

- *Histones* are proteins rich in lysine and arginine residues and thus positively-charged.
- For this reason they bind tightly to the negatively-charged phosphates in DNA.

The structure of chromosomes

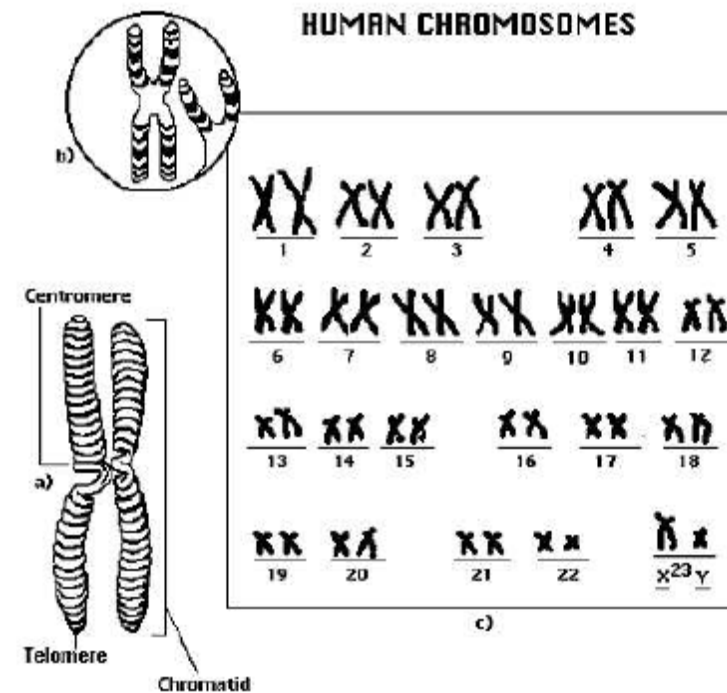
- All chromosomes have a stretch of repetitive DNA called the *centromere*. This plays an important role in chromosomal duplication before cell division.
- If the centromere is located at the extreme end of the chromosome, that chromosome is called acrocentric.
- If the centromere is in the middle of the chromosome, it is termed metacentric



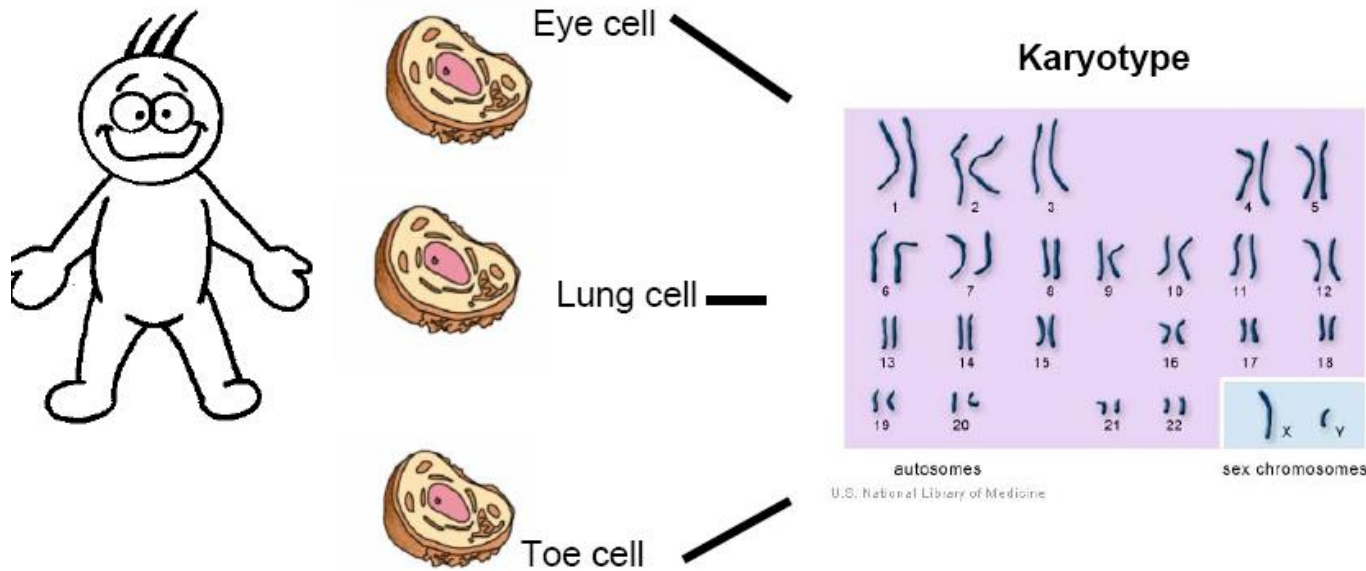
(www.genome.gov)

The structure of chromosomes

- The short arm of the chromosome is usually termed p for petit (small),
- the long arm, q , for queue (tall).



Every cell in the body has “the same” DNA



- One base pair is 0.00000000034 meters
- DNA sequence in any two people is 99.9% identical
- The residual 0.1% leads to several million spelling differences; variations leading to dramatically higher risks of certain cancers and other diseases

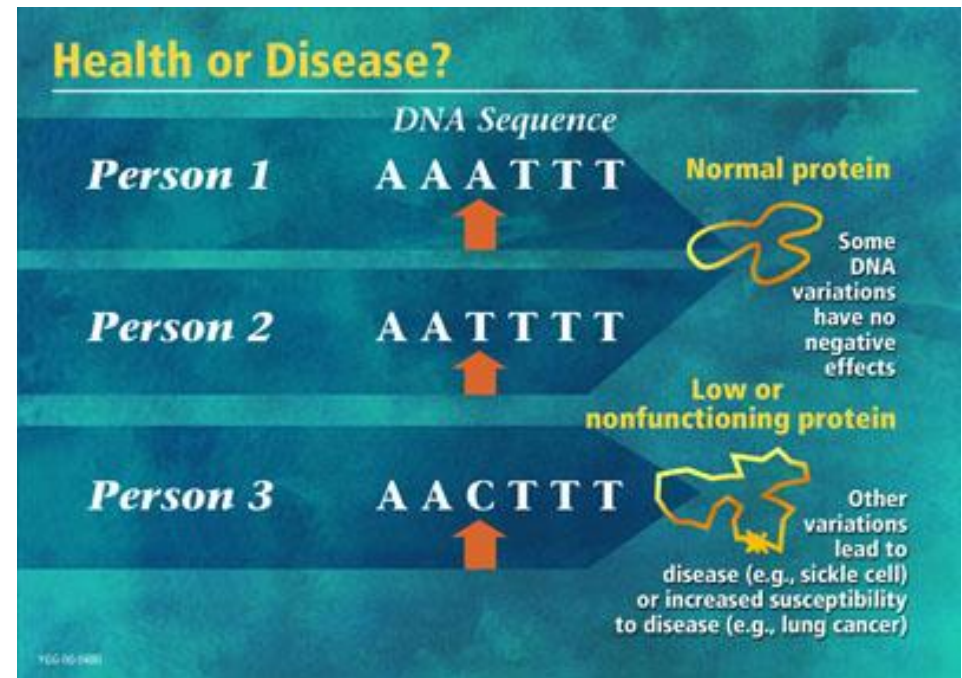
Every cell in the body has the same DNA: differential expression

- The **determination** of different cell types (*cell fates*) involves progressive restrictions in their developmental potentials. When a cell “chooses” a particular fate, it is said to be determined, although it still “looks” just like its undetermined neighbors. Determination implies a stable change - the fate of determined cells does not change.
- **Differentiation** follows determination, as the cell elaborates a cell-specific developmental program. Differentiation results in the presence of cell types that have clear-cut identities, such as muscle cells, nerve cells, and skin cells.
- Differentiation results from differential *gene expression*

DNA mutations

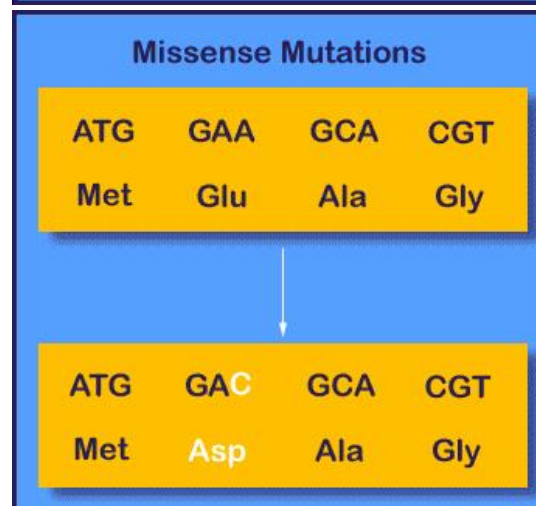
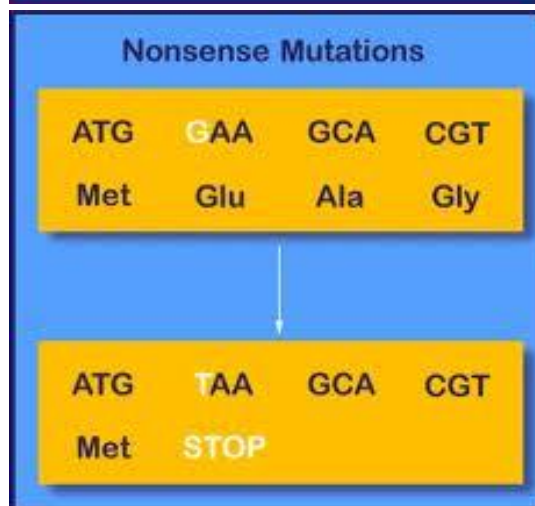
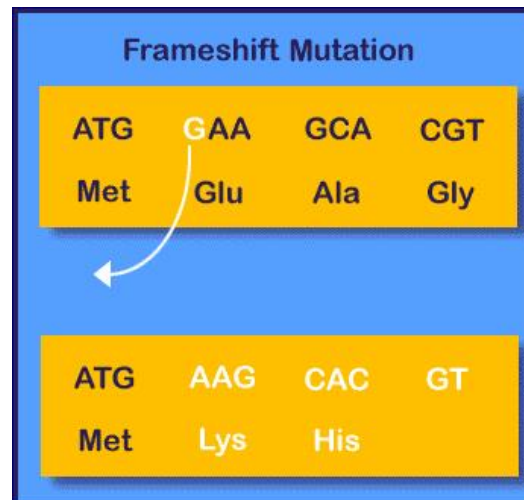
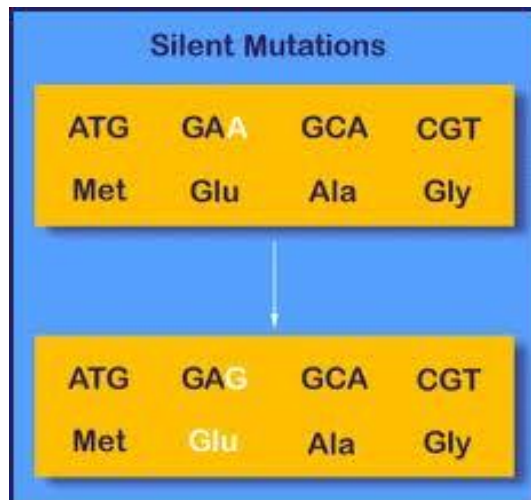
A source of variation

- As DNA polymerase copies the DNA sequence, some mistakes may occur.
- For example, one DNA base in a gene might get substituted for another. This is called a **mutation** (specifically a **point mutation**) or variation in the gene.
- The genetic code has built-in repair mechanisms which may work or not



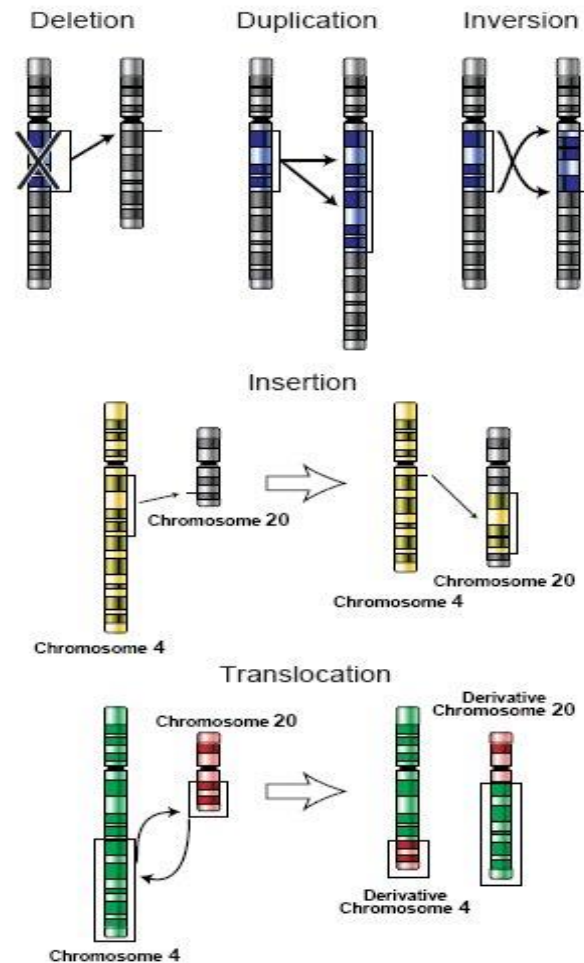
(Photo courtesy U.S. Department of Energy Human Genome Program)

Types of mutations



Types of mutations

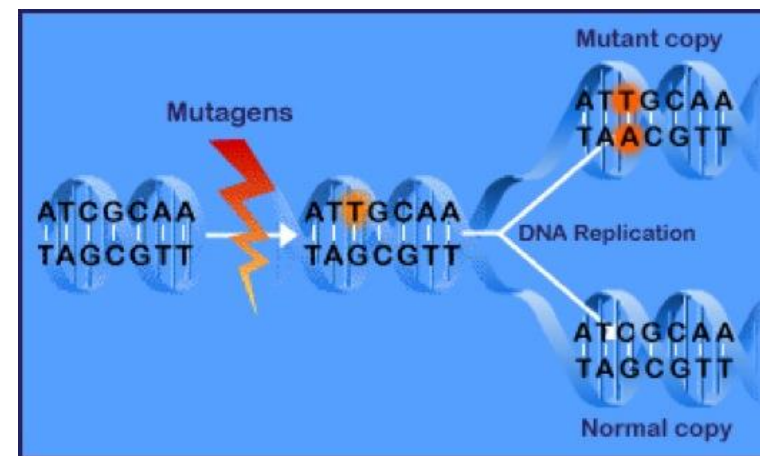
- Deletion
- Duplication
- Inversion
- Insertion
- Translocation



(National Human Genome Research Institute)

DNA repair mechanisms

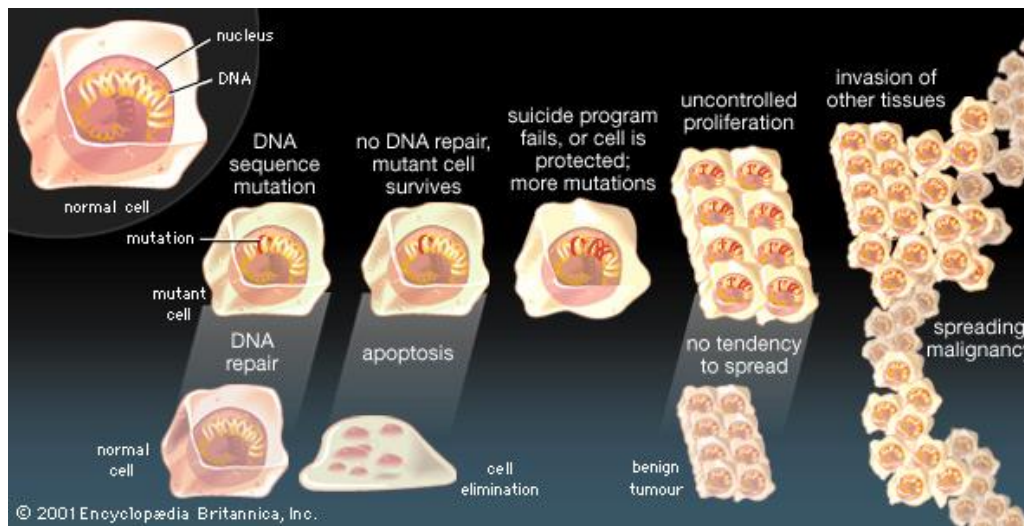
- In biology, a **mutagen** (Latin, literally origin of change) is a physical or chemical agent that changes the genetic material (usually DNA) of an organism and thus increases the frequency of mutations above the natural background level.
- As many mutations cause cancer, mutagens are typically also carcinogens.
- Not all mutations are caused by mutagens: so-called "**spontaneous mutations**" occur due to errors in DNA replication, repair and recombination.



(Roche genetics)

DNA repair mechanisms

- **damage reversal:** simplest; enzymatic action restores normal structure without breaking backbone
- **damage removal:** involves cutting out and replacing a damaged or inappropriate base or section of nucleotides
- **damage tolerance:** not truly repair but a way of coping with damage so that life can go on



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?

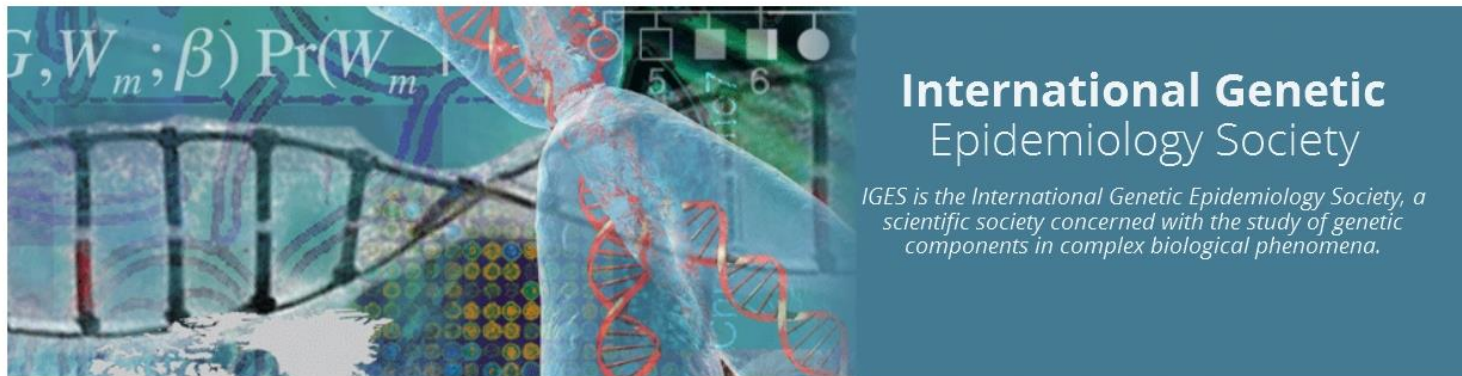


INTERNATIONAL GENETIC
EPIDEMIOLOGY SOCIETY

Contact Us: iges@geneticcepi.org

Search... 

[Home](#) [About IGES](#) [Scientific Meetings](#) [Publications](#) [Organization](#) [Membership](#) [Position Openings](#) [Latest News](#) [Login](#) [Abstracts](#)



Welcome!

Upcoming Events

The **23rd Annual IGES Meeting** will be held in Vienna, Austria, **August 28-30, 2014**. The meeting will be held in conjunction with two other major international scientific events – the *35th Annual Conference of the International Society for Clinical Biostatistics* and the *Genetic Analysis Workshop 19*. Details [here](#).

The **35th Annual Conference of the International Society for Clinical Biostatistics** will be held in Vienna, Austria, **August 24-28, 2014**. Details on the scientific topics of the conference, the invited sessions, the conference courses, and on the mini-symposia on Thursday morning are available on this [website](#).



Not an IGES Member? [Sign up!](#)

[Online registration for IGES](#)

Follow us!



Towards a definition for genetic epidemiology ...

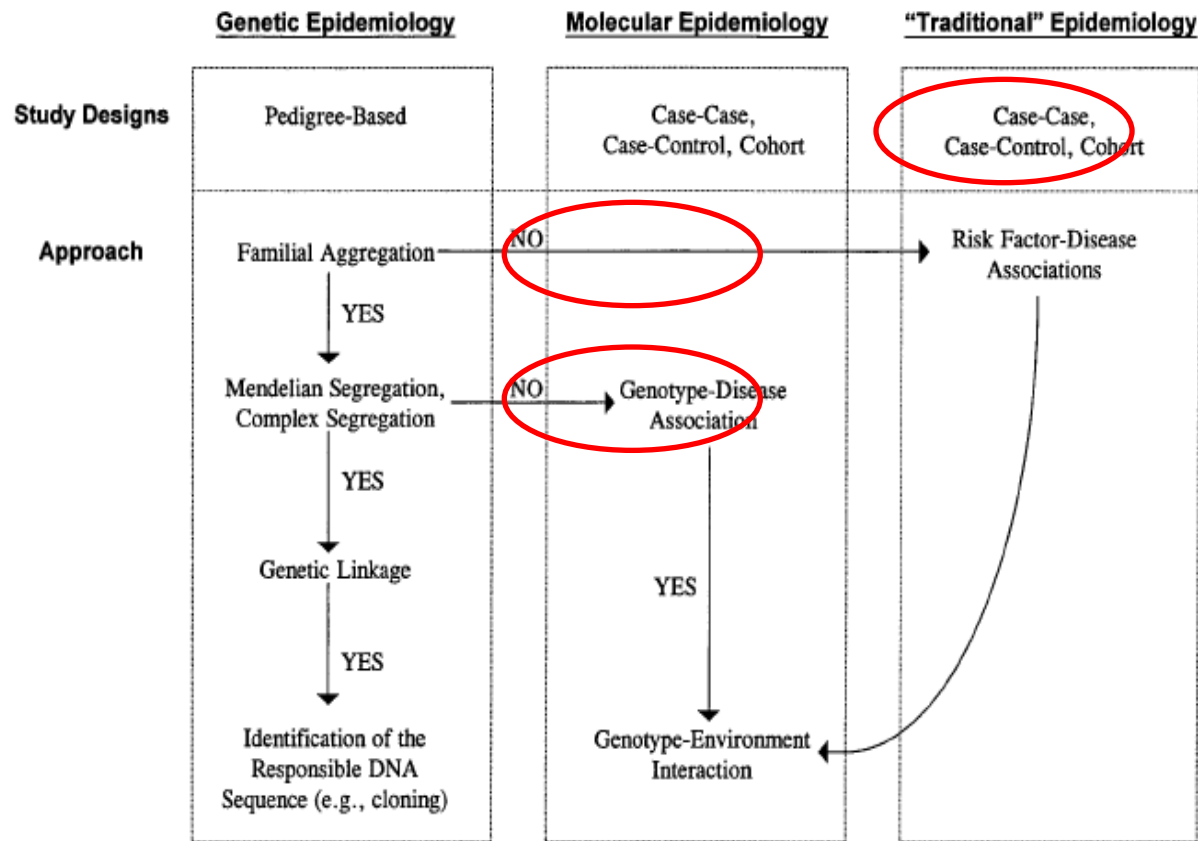
No agreement

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 via X – epidemiology



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

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).

Important 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)

Modern genetic epidemiology

- **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).**

Statistical methodology GWAS (genome-wide association studies) NGS
(Next Generation Sequencing)

GxE (gene by environment) interactions Family studies

Risk score development Predictive markers

Pharmacogenetics Microbiome Epigenetics

Transcriptomics eQTL analyses

Metabolomics Lipidomics

Ethics Data ownership

Modern genetic epidemiology in practice

- 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.

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 series¹²), 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.

Lancet 2005; 366: 941–51

See Comment page 880

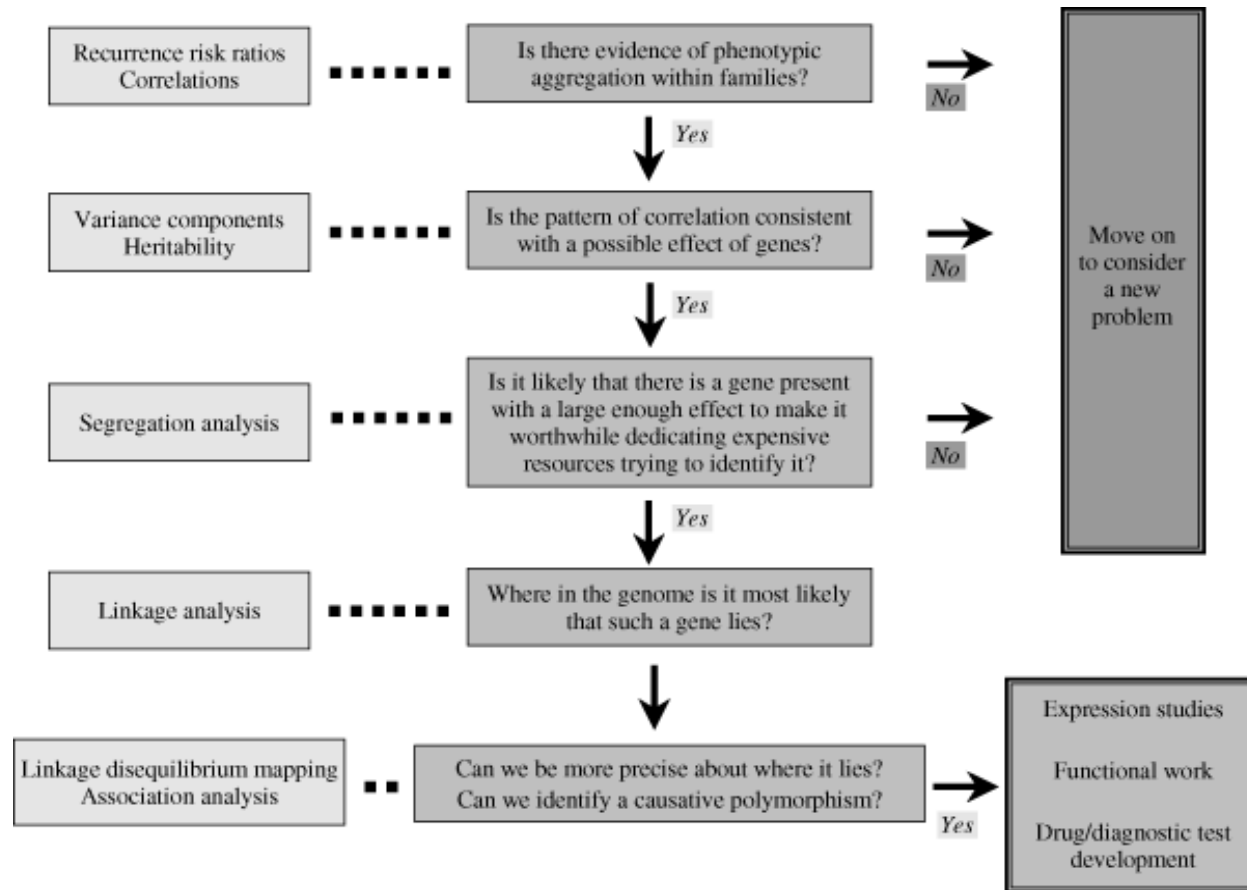
This is the first in a Series of seven papers on genetic epidemiology.

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)

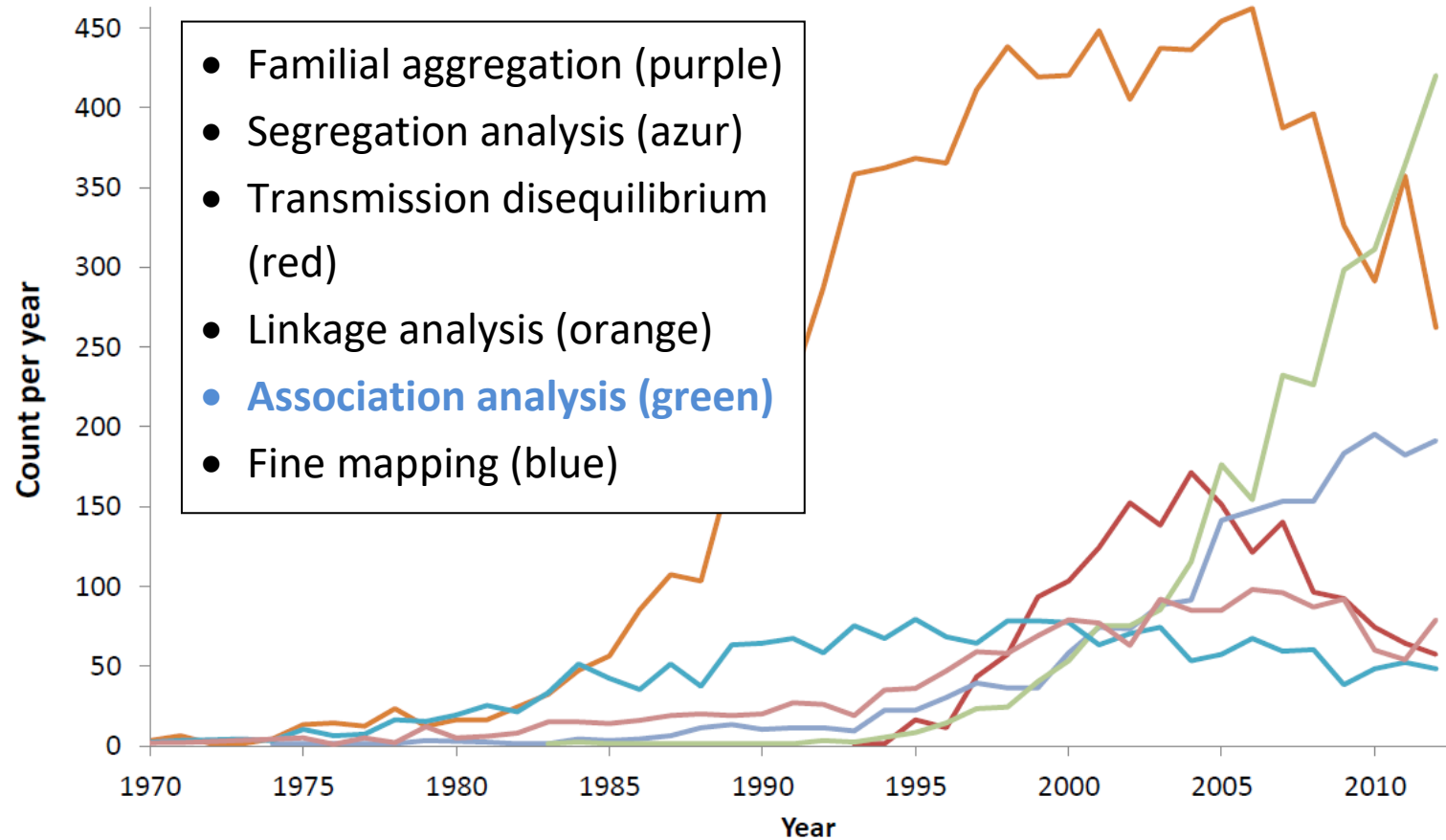
Correspondence to: Prof Paul R Burton, Department of Health Sciences, University of Leicester, 22–28 Princess Road West, Leicester LE1 6TP, UK
pb51@le.ac.uk

Relevant questions in genetic epidemiology



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

Use of genetic terms over time



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

Positioning compared to public health genomics

- The field of public health genomics by Khoury (2010)

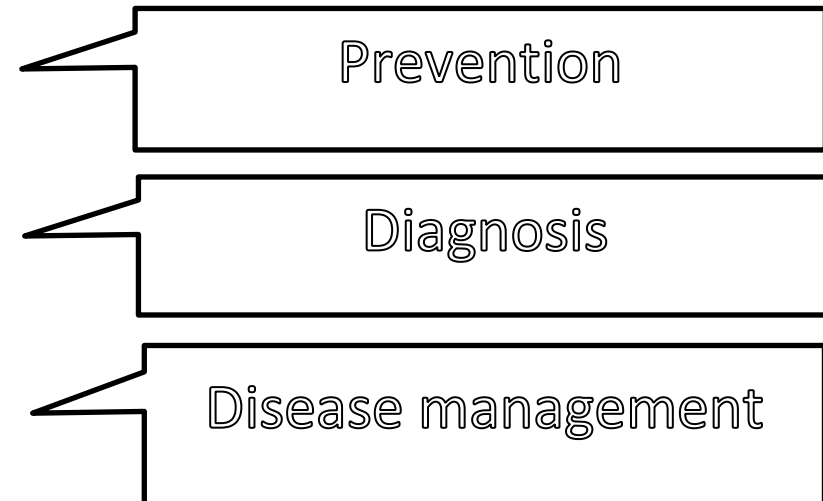
Khoury et al.: Public health genomics, a multidisciplinary field concerned with the **effective and responsible translation of genome-based knowledge and technologies to improve population health**. ... Public health genomics uses population-based data on genetic variation and gene-environment interactions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

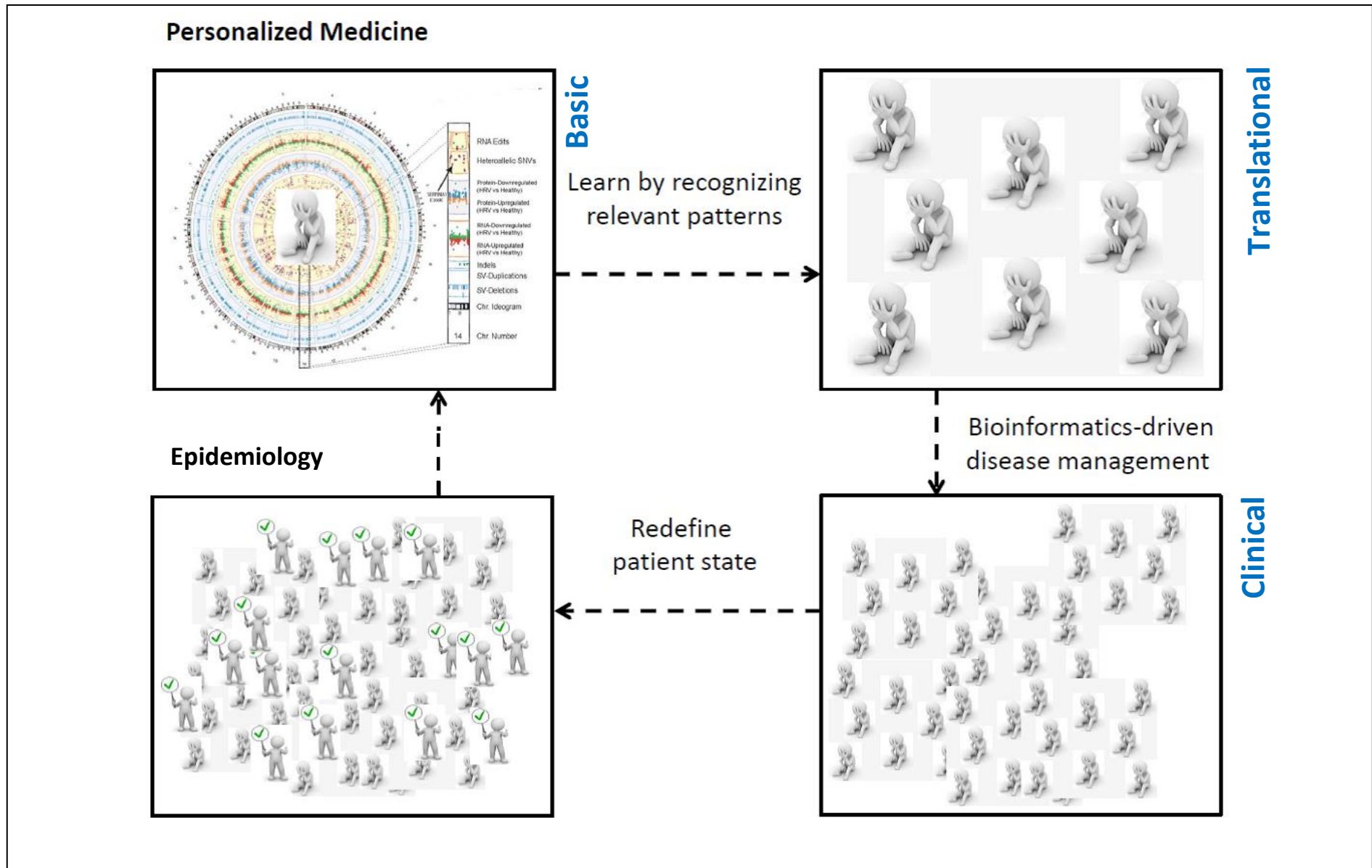
(IGES presidential address A Ziegler, Chicago 2013)

Positioning compared to precision medicine

“a medical model using characterization of individual’s phenotypes and genotypes (e.g., molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention.”

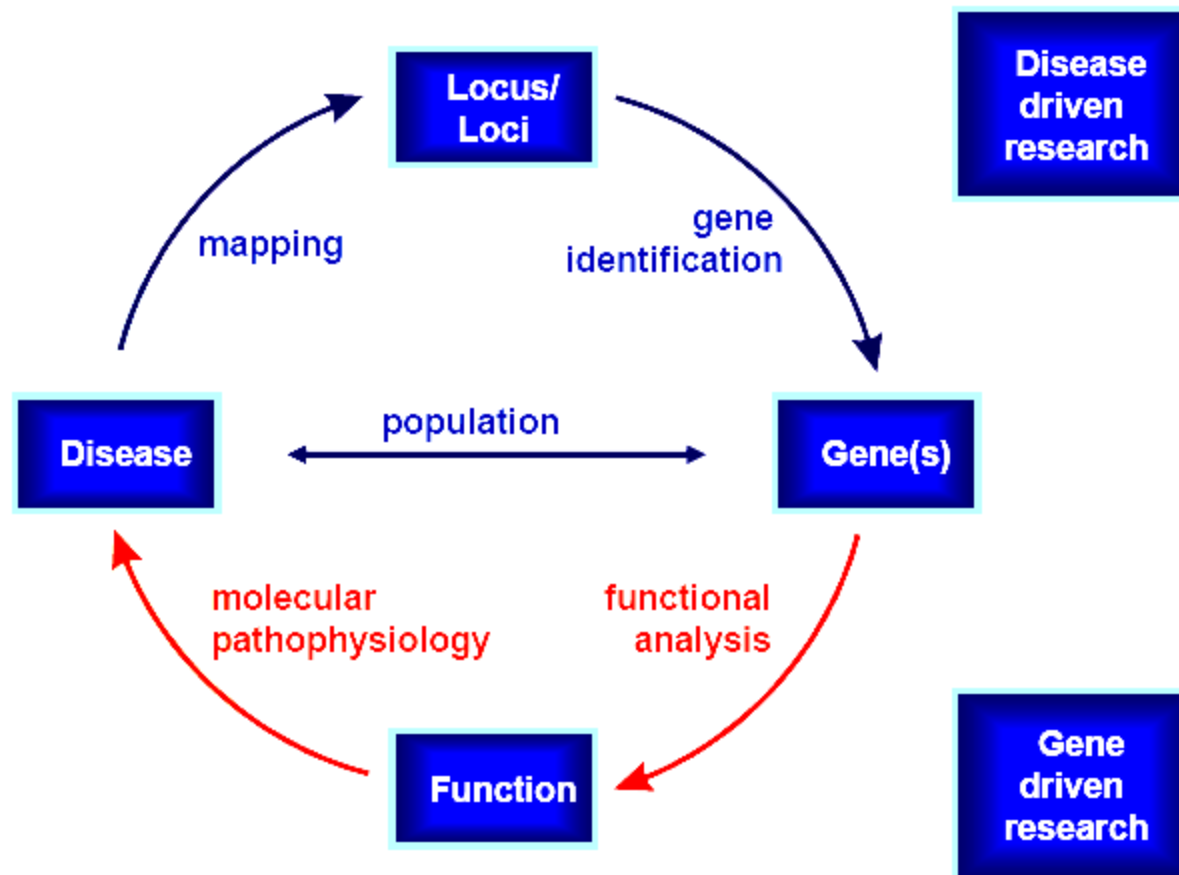
(HORIZON2020 Advisory Group; EU Health Ministers – December 2015)



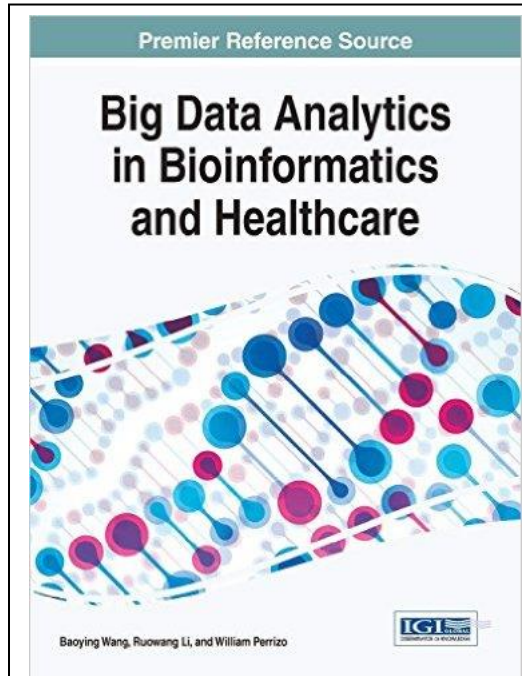


Previous: PM & disease diagnosis & disease management

Here: PM & understanding biological mechanisms of disease



Biological challenge: omics data are related



“As of 2006 there were 1,062 papers explicitly mentioning "*data integration*" in their abstract or title, whereas this number has more than doubled in 2013 (2,365).”

(Gomez-Cabrero et al. 2014)



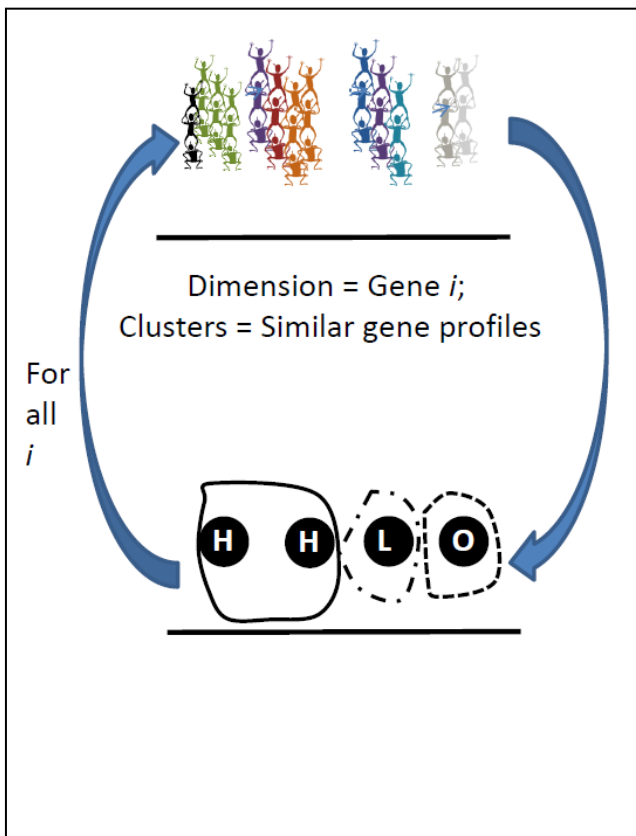
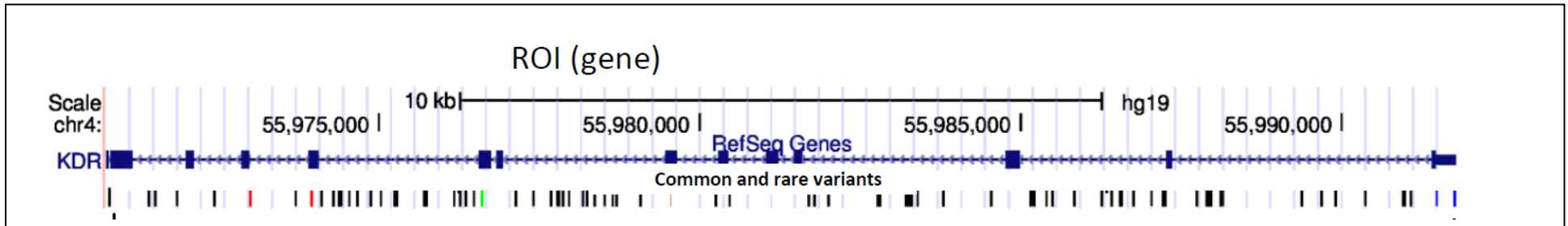
Chapter 13: Perspectives on Data Integration in Human Complex Disease Analysis

(Van Steen and Malats 2015)

BIO3 Route 1: Understanding biological mechanisms via integrated networks

- Genomics, epigenomics and transcriptomics integration with penalized regression (Pineda et al. 2015)
- Integrated gene expression and methylation in glioblastoma with LABNet (Gadaleta et al. - submitted)
- Gene networks via conditional inference forests (Bessonov et al. – submitted)
- Integrated gene expression and genomics with adapted conditional inference forests (Bessonov et al. – circulating among co-authors)
- Epistasis-based network analysis of *cis* and *trans* regulatory effects in asthma (Bessonov et al. – circulating among co-authors)

BIO3 Route 2: Gene identification via association (interact.) analyses

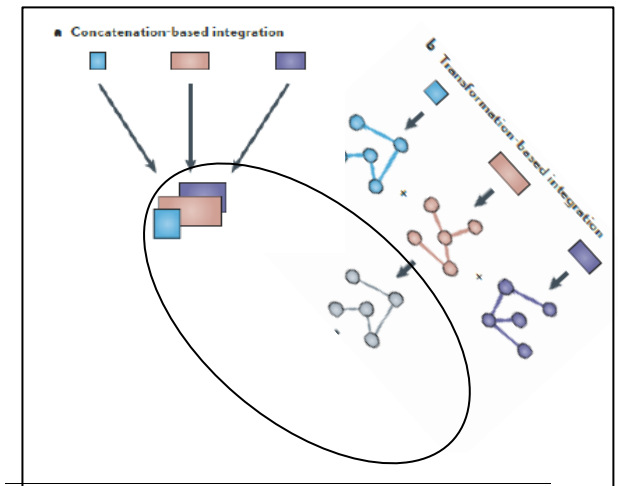


Diffusion kernel PCA

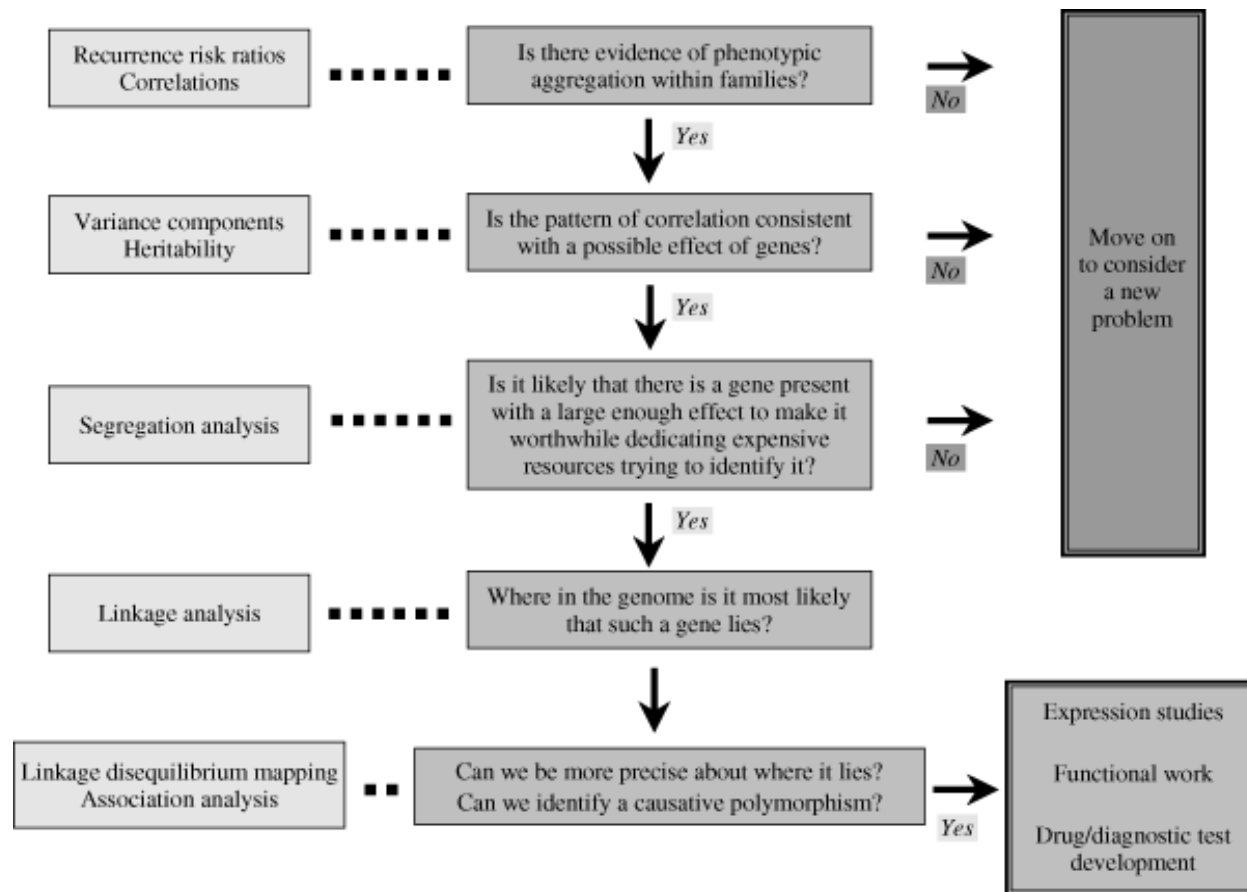
to perform omics integrated gene-based sample clustering

- Component-based
- Kernel-based
- Network-based

(Fouladi et al. 2015, 2016+)



2.b Designs in genetic epidemiology



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

Different flows of research in genetic epidemiology require specific designs

In summary, 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

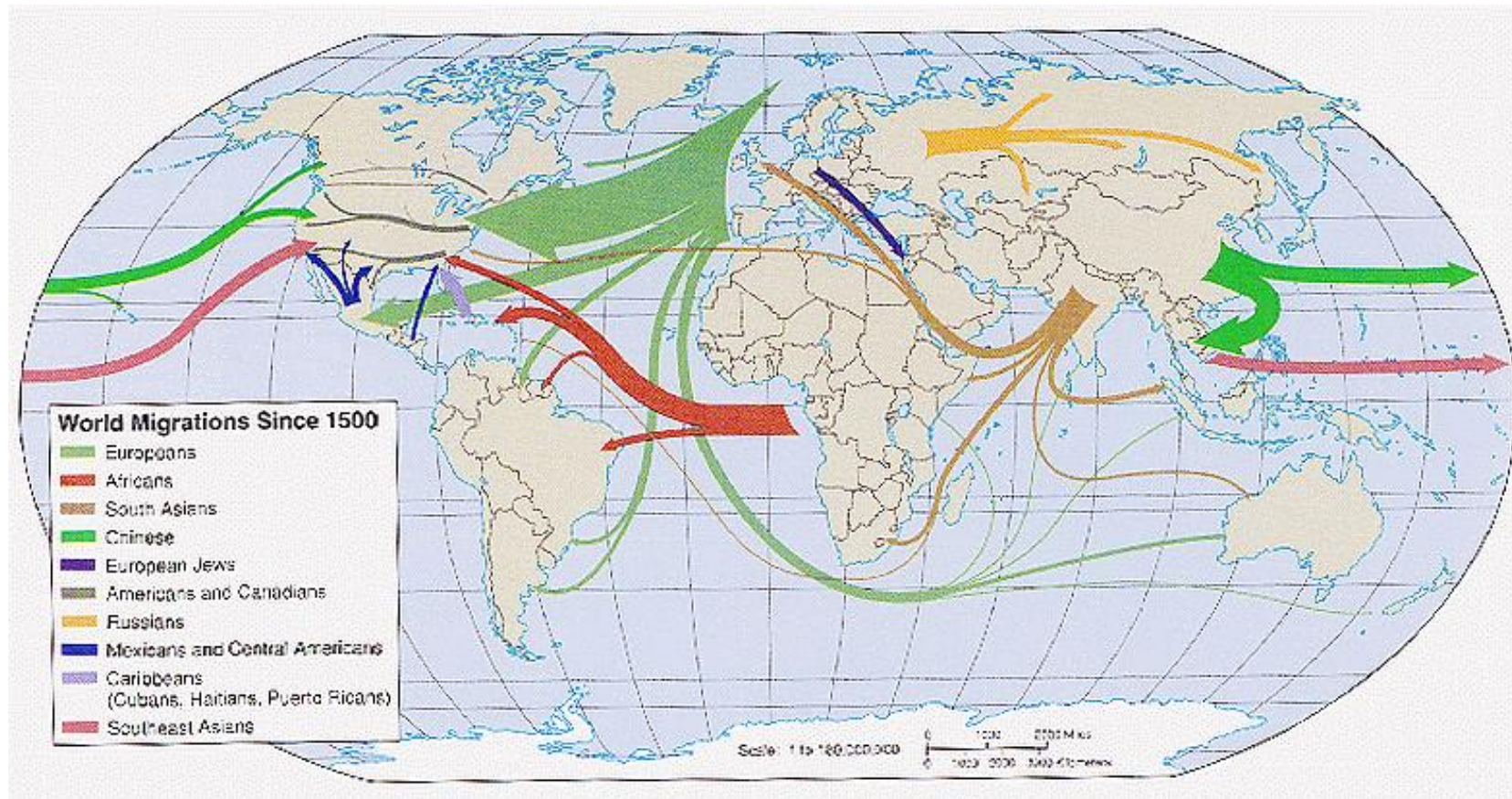
2.c Study types in genetic epidemiology

Main methods in genetic epidemiology

A) 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

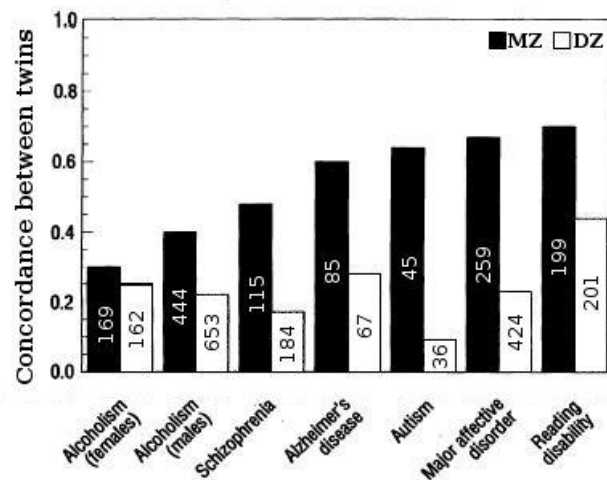
- As one of the initial steps in the process of genetic epidemiology, one could use information on populations who migrate to countries with different genetic and environmental backgrounds - as well as rates of the disease of interest - than the country they came from.
- Here, one compares people who 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>)

B) Segregation studies:

- 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.

C) Linkage studies:

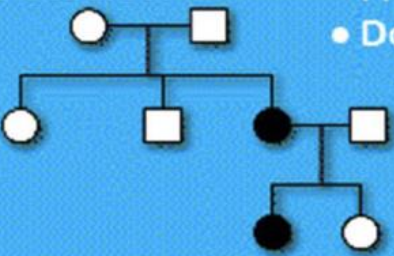
- What is the location of the disease gene(s)? Linkage studies screen the whole genome and use parametric or nonparametric methods such as allele sharing methods {affected sibling-pairs method} with no assumptions on the mode of inheritance, penetrance or disease allele frequency (the parameters). The underlying principle of linkage studies is the cosegregation of two genes (one of which is the disease locus).

Linkage and Association

Approaches to Identifying Susceptibility Genes for Common Diseases



Linkage Studies:

- A relation between loci
- Done within families



Association Studies:

- A relation between alleles
- Done in populations

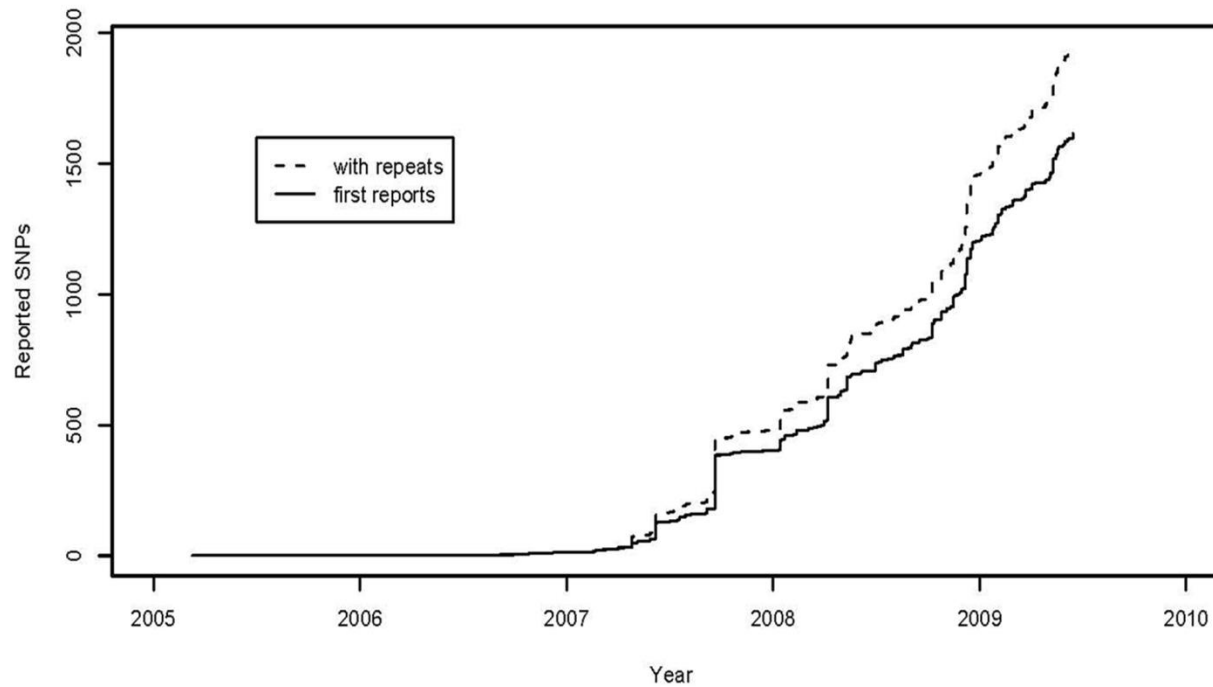
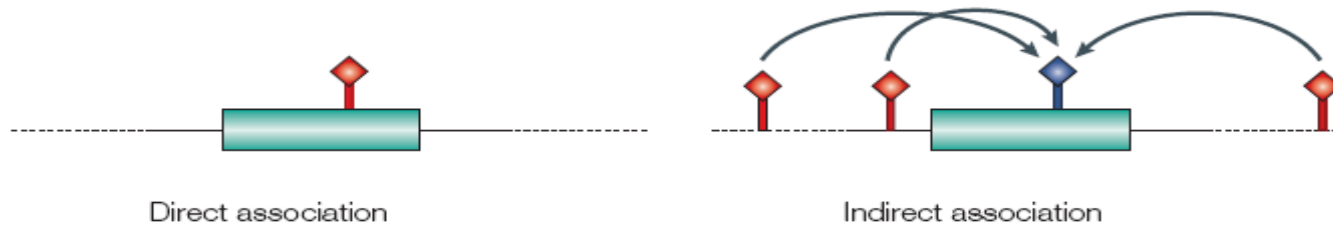
Type 1 Diabetes Patients	Controls
	

(Roche Genetics Education)

D) Association studies:

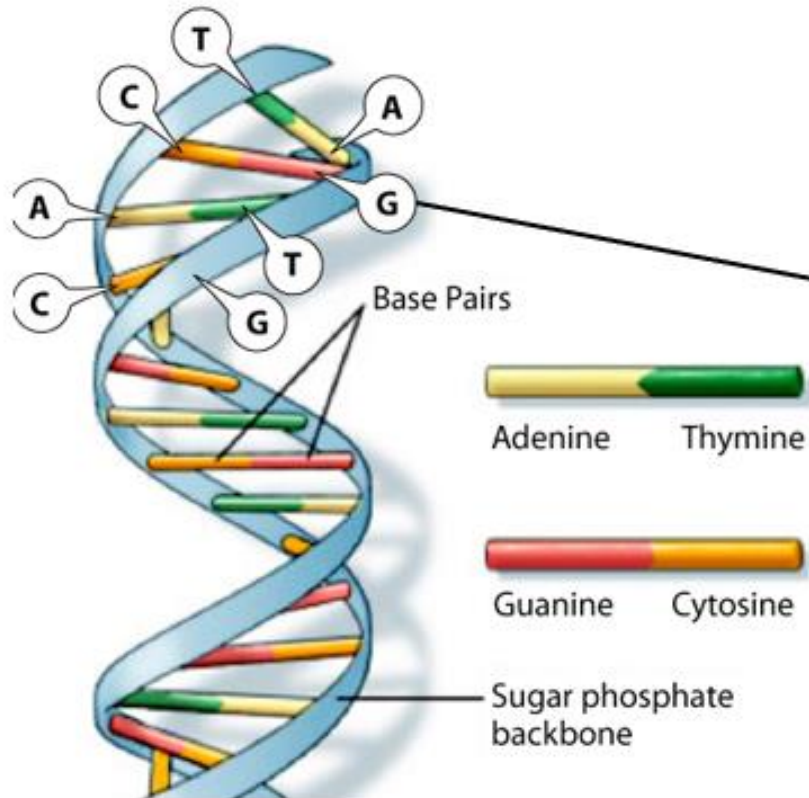
- What is the allele associated with the disease susceptibility? The principle is the coexistence of the same marker on the same chromosome in affected individuals (due to linkage disequilibrium). Association studies may be family-based (TDT) or population-based. Alleles or haplotypes may be used. Genome-wide 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

GWAS most popular “variable”: SNPs (Single Nucleotide Polymorphisms)



Single Nucleotide Polymorphisms (SNPs)	Frequency in general population
G	95%
A	5% > 1%

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

The logo for 23andMe features a stylized 'X' shape formed by two overlapping, thick, rounded bars. The left bar is pink and the right bar is green. To the right of the 'X' is the text '23andMe' in a grey, sans-serif font.

- deCODEme

The logo for deCODEme consists of the word 'deCODE' in a bold, black, sans-serif font above the word 'ME' in a larger, blue, sans-serif font. A white DNA double helix is integrated into the letter 'E' of 'ME'. The entire logo is set against a light blue background with a reflection effect below it.

- Navigenics

The screenshot shows the 23andMe website interface. At the top left is the 23andMe logo. To the right are buttons for 'sign in', 'register kit', and a shopping cart icon with '0' items. Below this is a navigation bar with links for 'welcome', 'ancestry', 'health', 'how it works' (highlighted), 'store', 'search', and 'help'. On the left side, there is a vertical menu with 'Order', 'Register', 'Spit', 'Send', and 'Discover' (highlighted with a white arrow). The main content area features a central promotional graphic for 'finding your roots with HENRY LOUIS GATES, JR.' and 'Ancestry Painting'. The graphic includes a bar chart showing ancestry percentages: European 60%, African 34%, and Asian 6%. Below this, there are two sections: 'Ancestry' and 'Health'. The 'Ancestry' section includes a description and a list of features: 'Relative Finder', 'Global Origins', and 'Ancestral Lineages'. The 'Health' section includes a description and a list of features: 'Carrier Status', 'Disease Risk', and 'Drug Response'. To the right of the main content is a mobile app interface showing a 'Results' screen with categories like 'New & Updated' (56), 'Disease Risk' (30), 'Carrier Status' (44), 'Traits' (13), and 'Drug Response' (8). At the bottom of the page, there is a grey bar with the text 'Start exploring your DNA...' and a red 'Add to Cart' button.

23andMe

welcome ancestry health **how it works** store search help

sign in register kit 0

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%

New & Updated 56 >

Disease Risk 30 >

Carrier Status 44 >

Traits 13 >

Drug Response 8 >

Start exploring your DNA...

Add to Cart

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

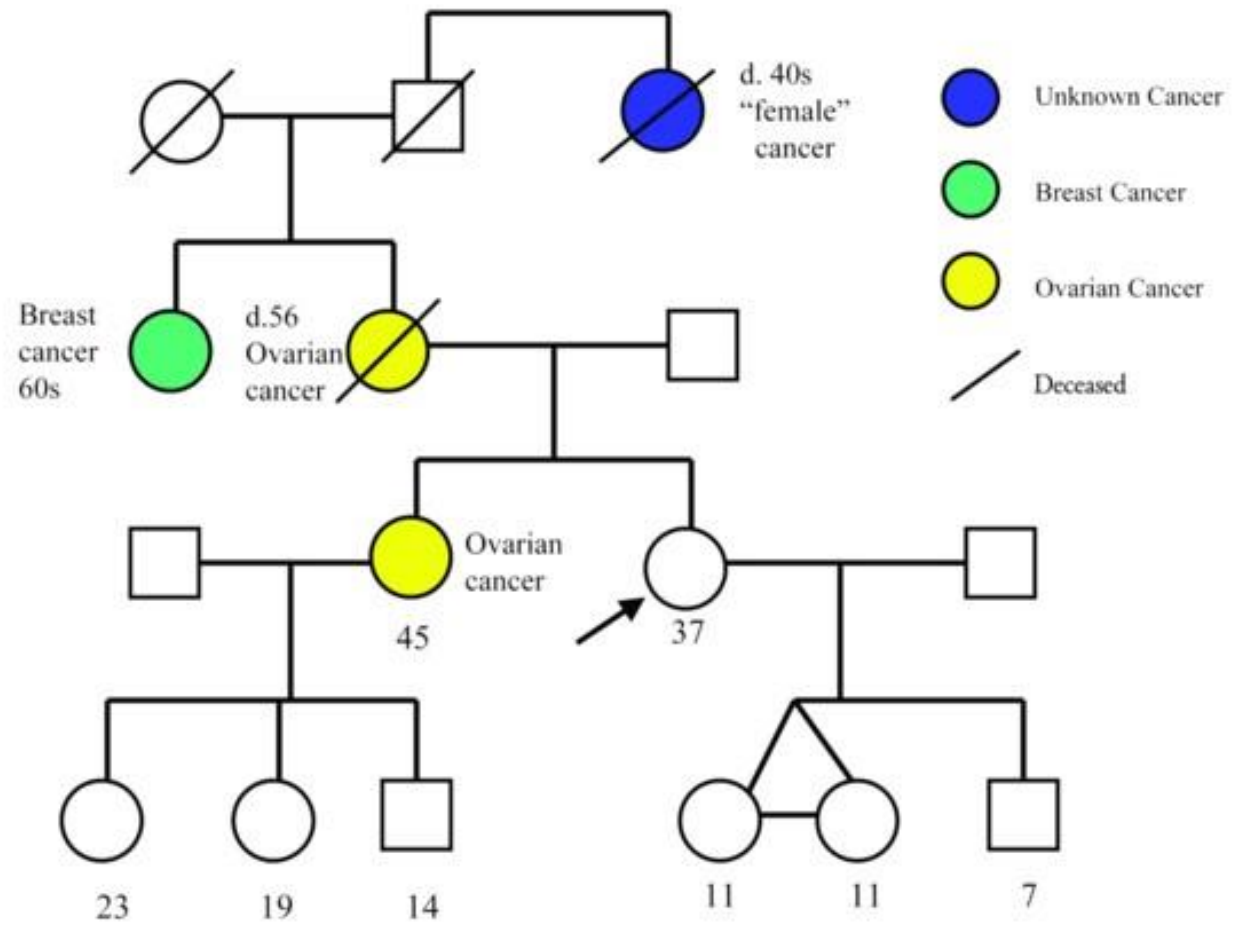
Aggregation and segregation studies in human genetics

- 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?

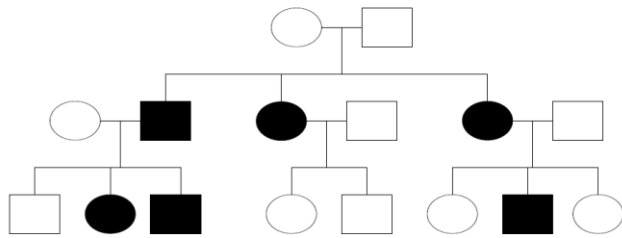
Definition of familial aggregation

- 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*.

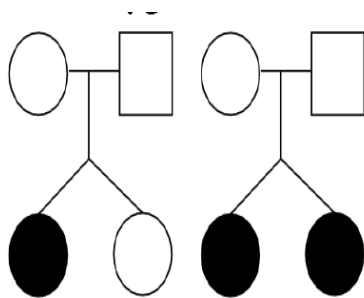
Example 1: does the phenotype run in the family?



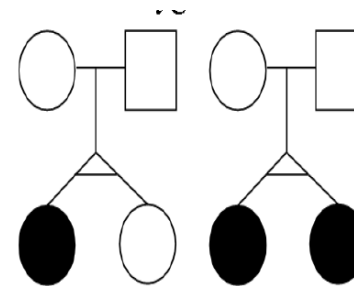
- **Pedigree** - A diagram of the genetic relationships and medical history of a family using standardized symbols and terminology
- **Founder** - Individuals in a pedigree whose parents are not part of the pedigree
- **Extended pedigrees**



Dizygotic twins

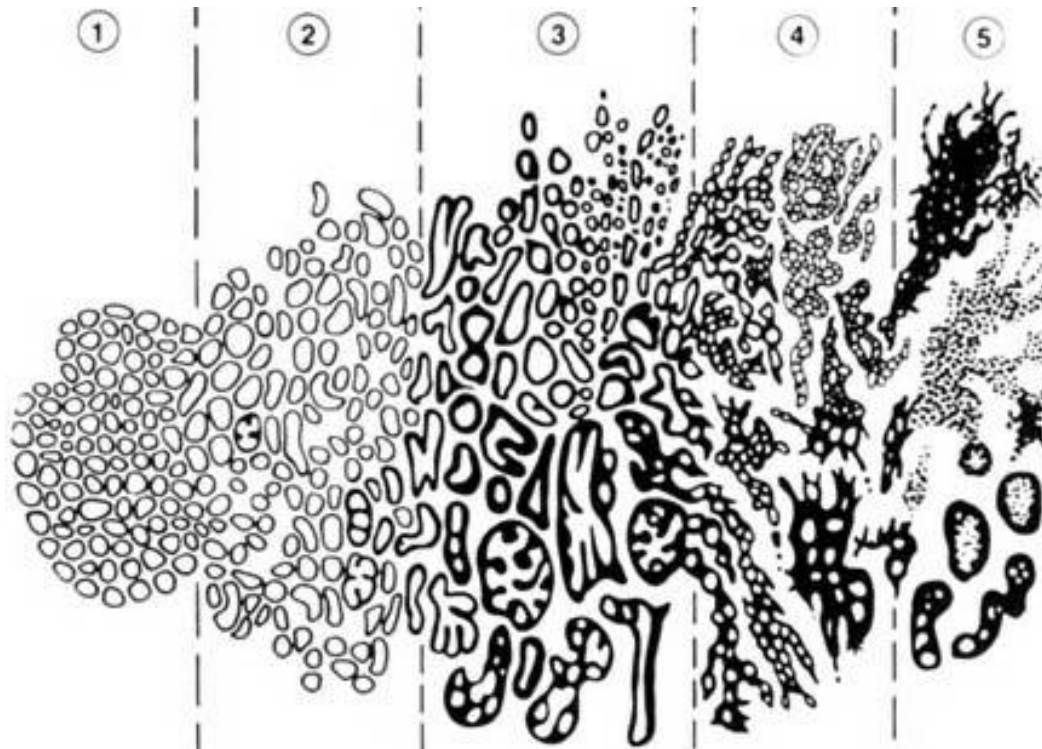


Monozygotic twins



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

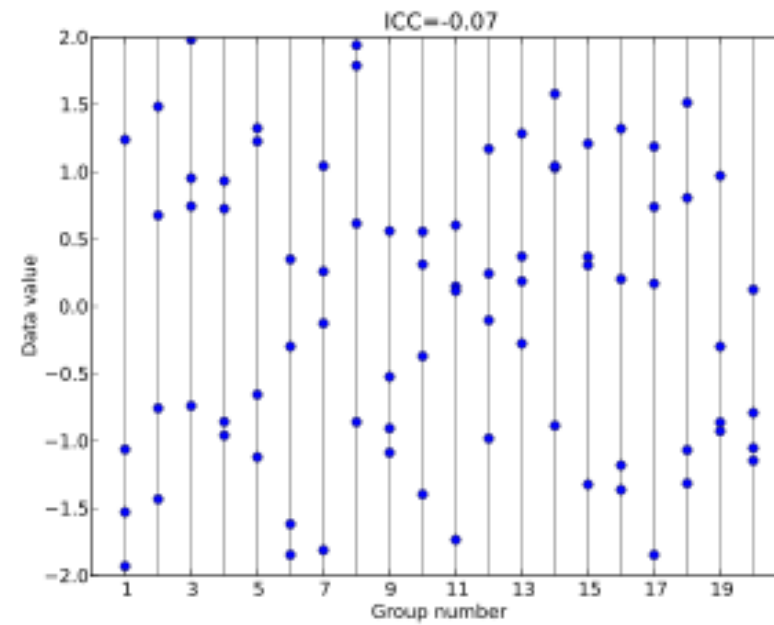
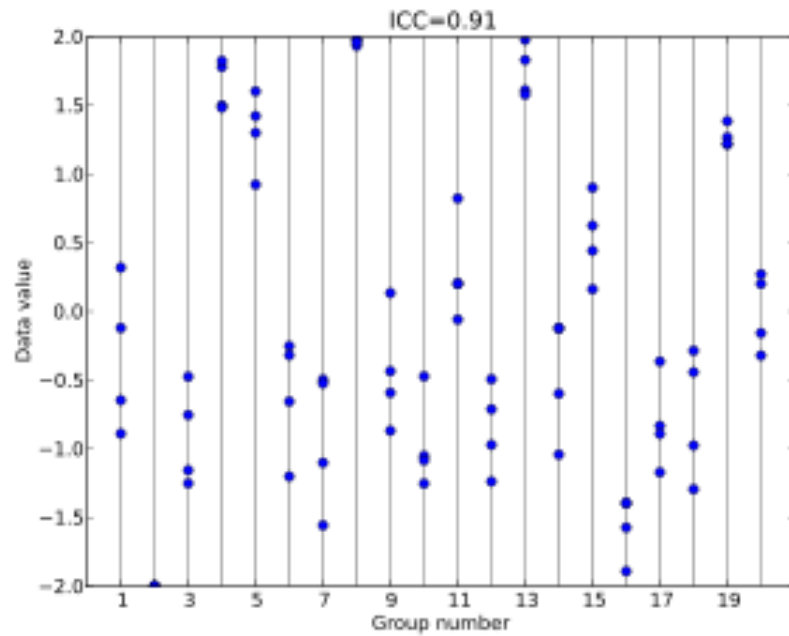
Proband selection

- 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
- Available via multilevel modelling analysis of variance, *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,

Population mean(Y) = Prob($Y = 1$) = Population risk, K

Techniques

- An alternative algebraic expression for the covariance is

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

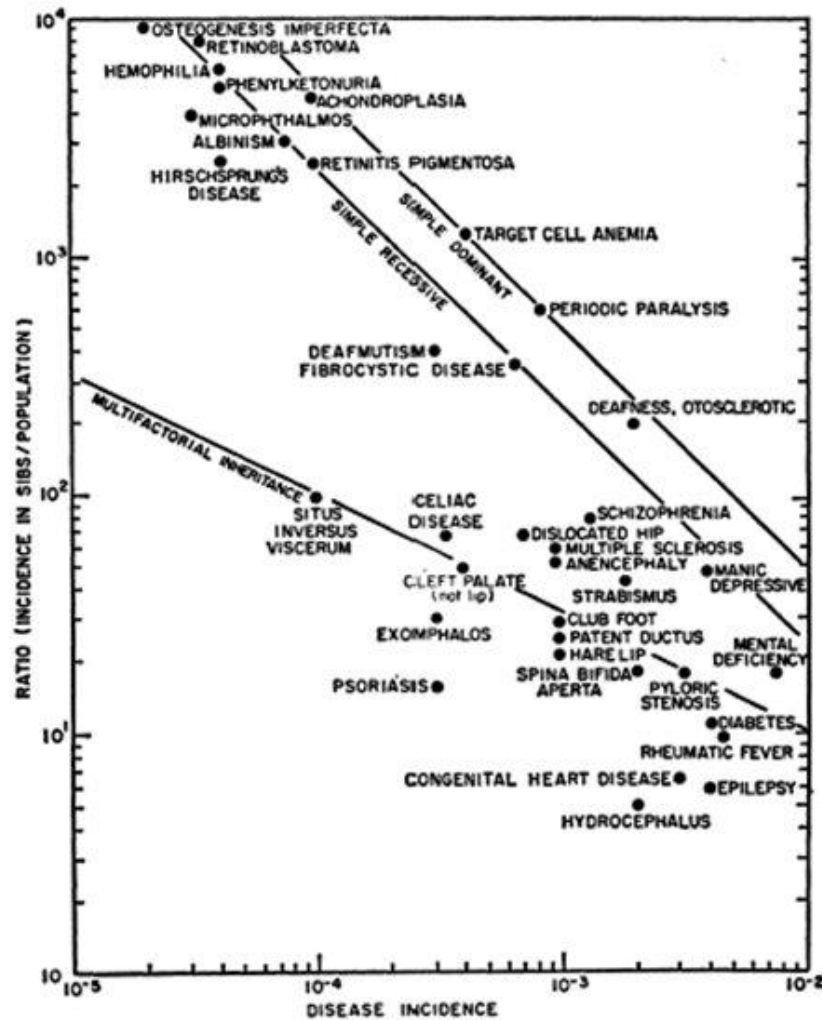
with $\text{Mean}(Y_1 Y_2)$ the probability that both relatives are affected. From this we derive for the familial risk ratio λ , defined before:

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

- 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

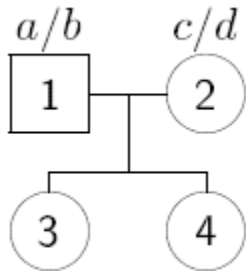
- For λ_S = ratio of risk in sibs compared with population risk.
 - cystic fibrosis: the risk in sibs = 0.25 and the risk in the population = 0.0004, and therefore $\lambda_S = 500$
 - Huntington disease: the risk in sibs = 0.5 and the risk in the population = 0.0001, and therefore $\lambda_S = 5000$
- 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 $1/4 = 1/2 \times 1/2$
- 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:

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

- 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* = $\frac{1}{2}$
 - Under 1-IBD, *prob that two randomly selected alleles are IBD* = $\frac{1}{4}$
 - Under 0-IBD, *prob that two randomly selected alleles are IBD* = 0
- So the kinship coefficient is

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

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

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

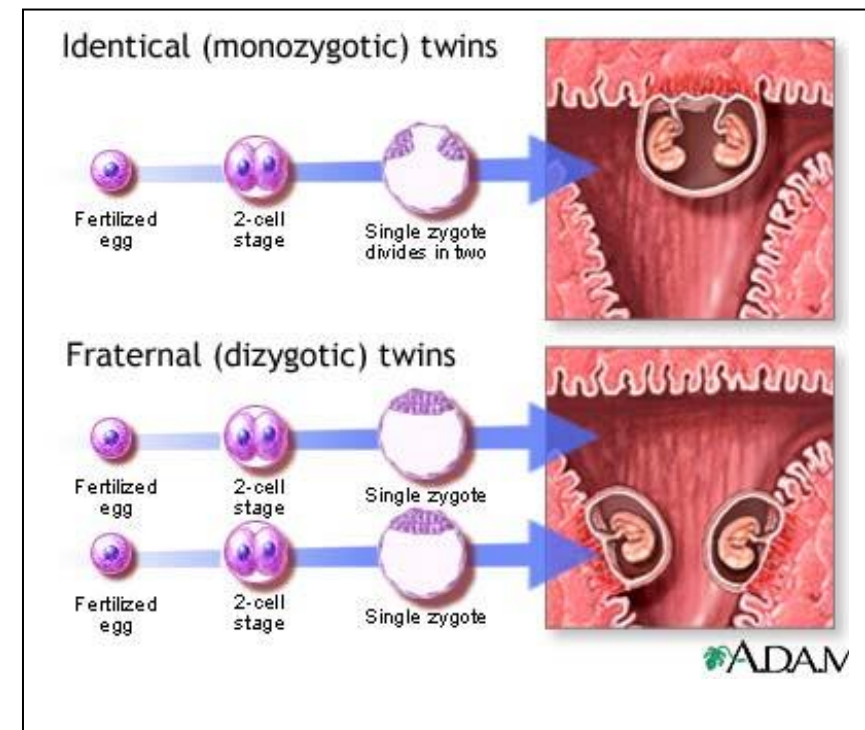
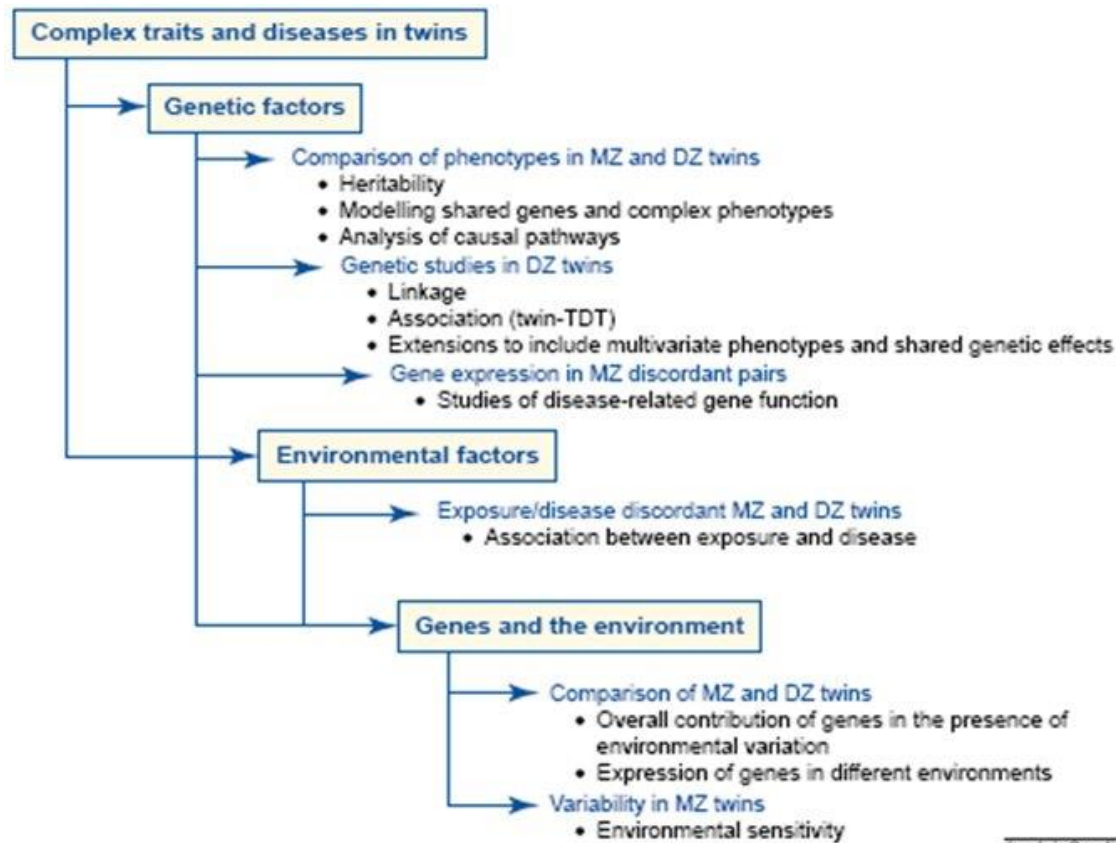
IBD sharing and kinship by relationship

Relationship	No. alleles shared IBD			Φ
	2	1	0	
	z_2	z_1	z_0	
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 variancea twin design can be used.

Missing heritability

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



(Maher 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 (see next section)
- 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.

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

Penetrances – explained via binary traits (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 $k - 1 \leq c_{12} \leq k$ and $k - 1 \leq c_{22} \leq k$, with $0 \leq k \leq 1$, $c_{12} \geq 0$, and $c_{22} \geq 0$

- 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 $0 \leq k \leq 1$, $r \geq 1$, and $0 \leq r^2 k \leq 1$

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).

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

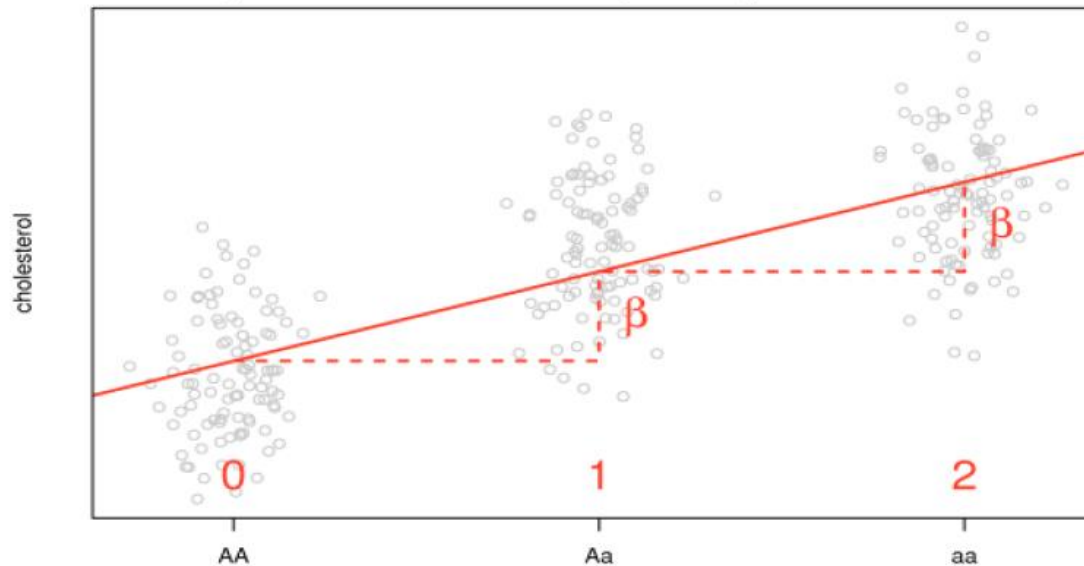
Quantitative traits

- 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.

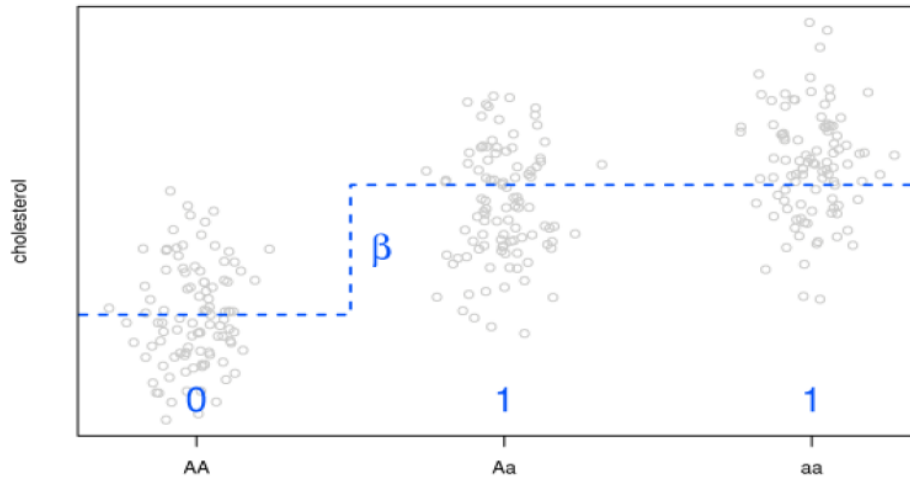
Technique

- 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

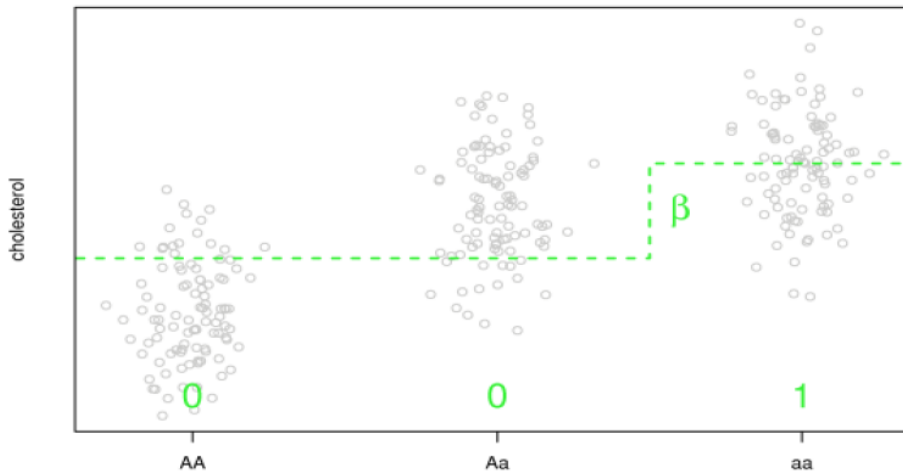
Additive model (the most commonly used)



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.

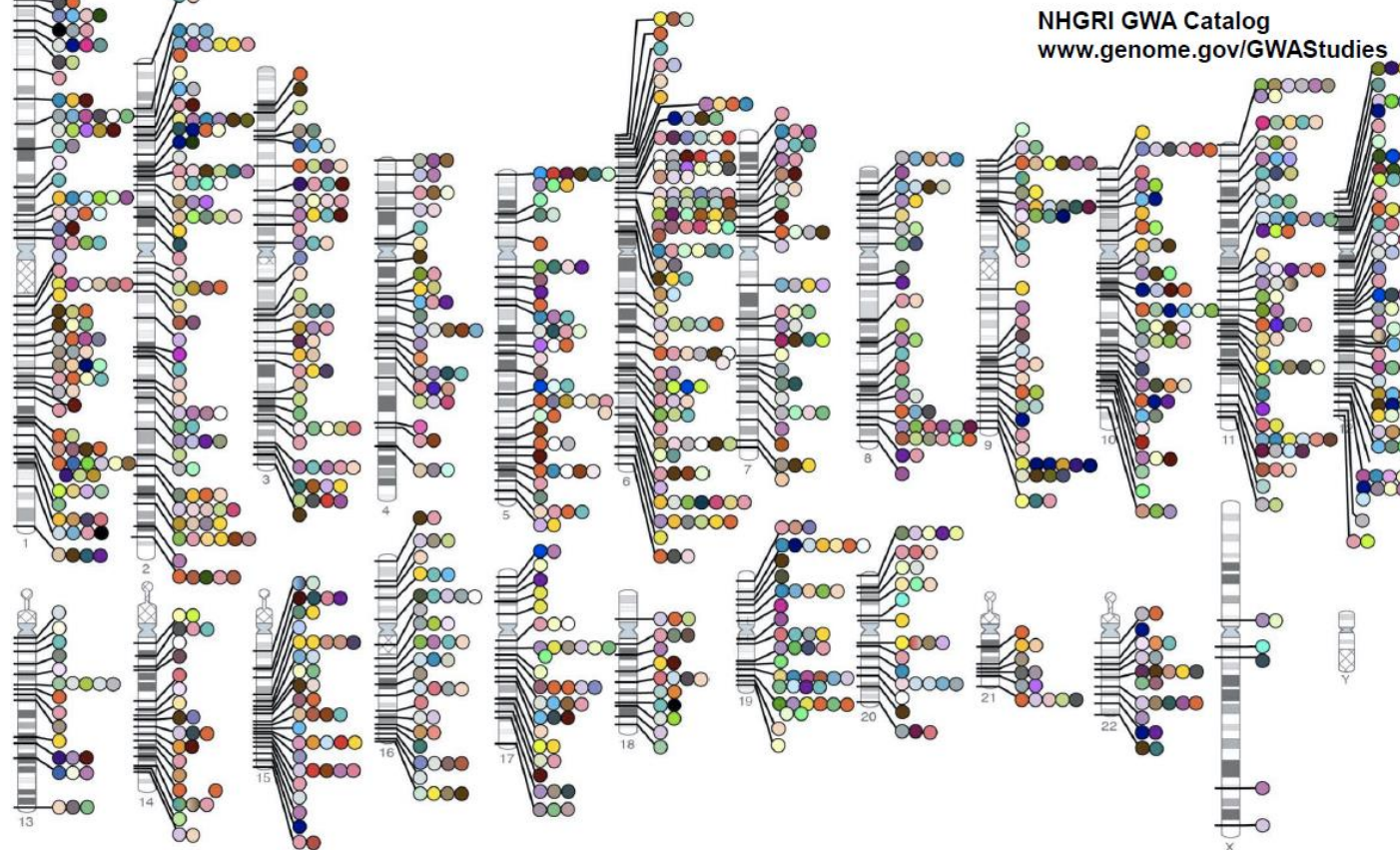
Historical overview: GWAs as a tool to “map” diseases - locus heterogeneity

2008 third

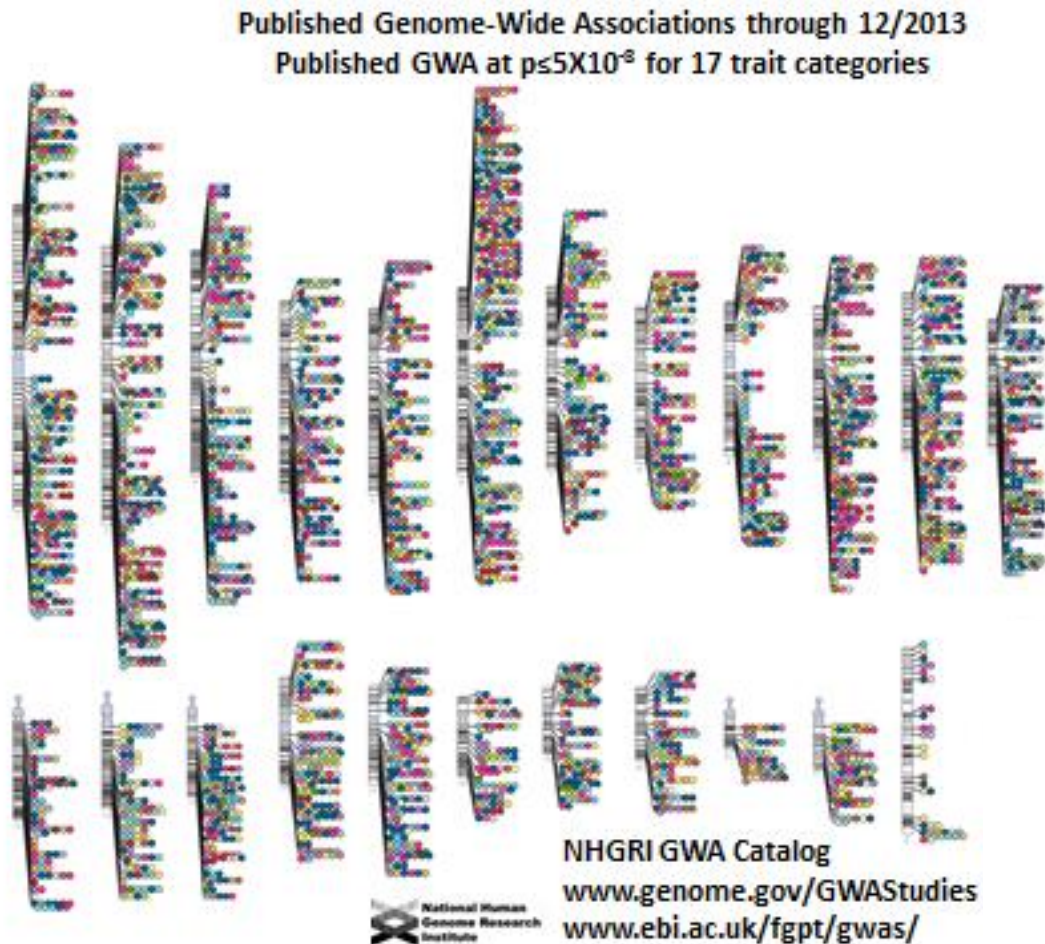


Historical overview: 210 traits – multiple loci (sites, locations)

Published Genome-Wide Associations through 12/2010,
1212 published GWA at $p \leq 5 \times 10^{-8}$ for 210 traits



Historical overview: trait categories



- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

- **Pleiotropy** in contrast: 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>)
- **Epistasis (~genetic interactions, GxG interactions)**: 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.

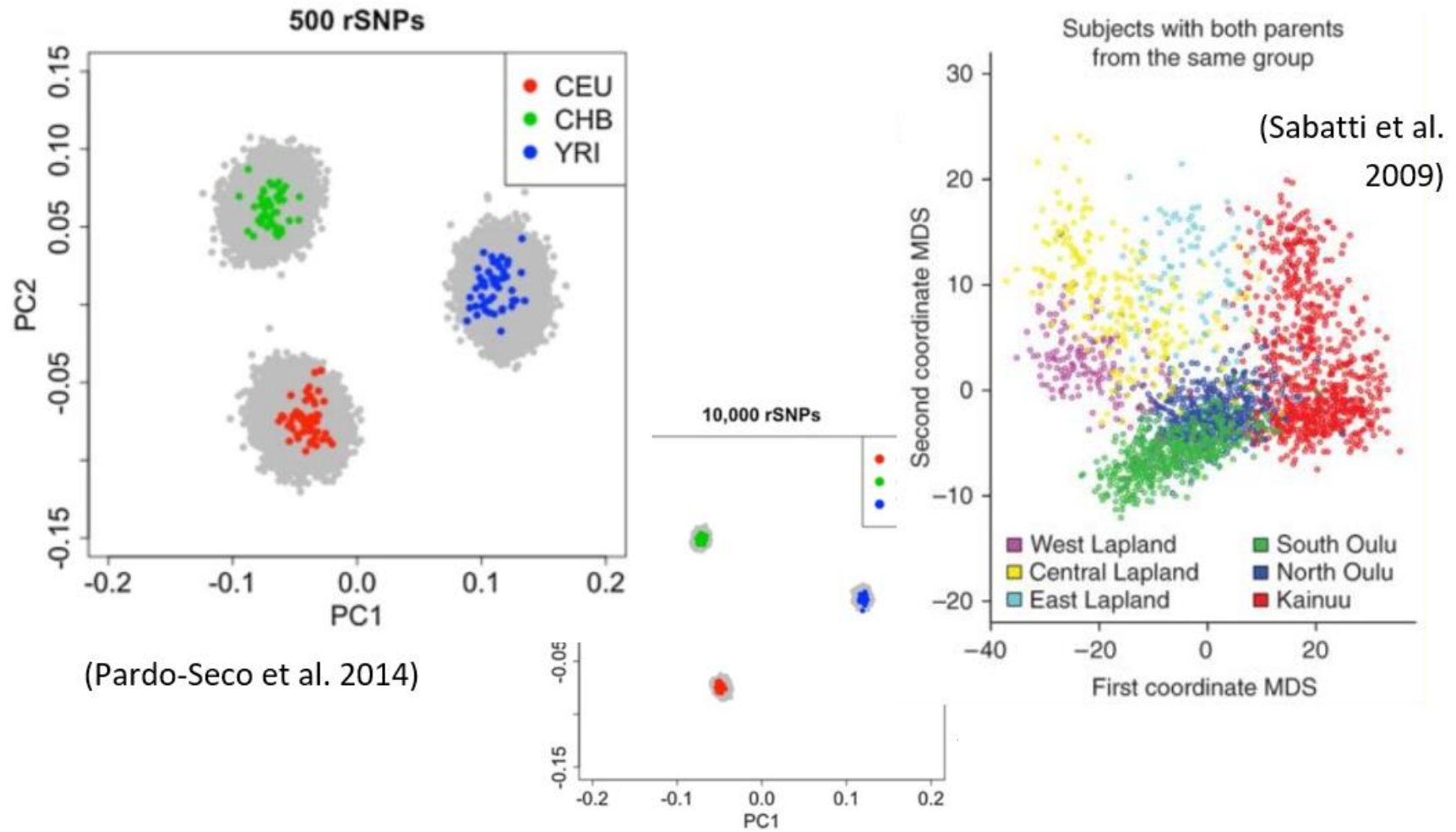
5 A bird's eye view on genetic epidemiology

5.a Selected topics in genetic epidemiology

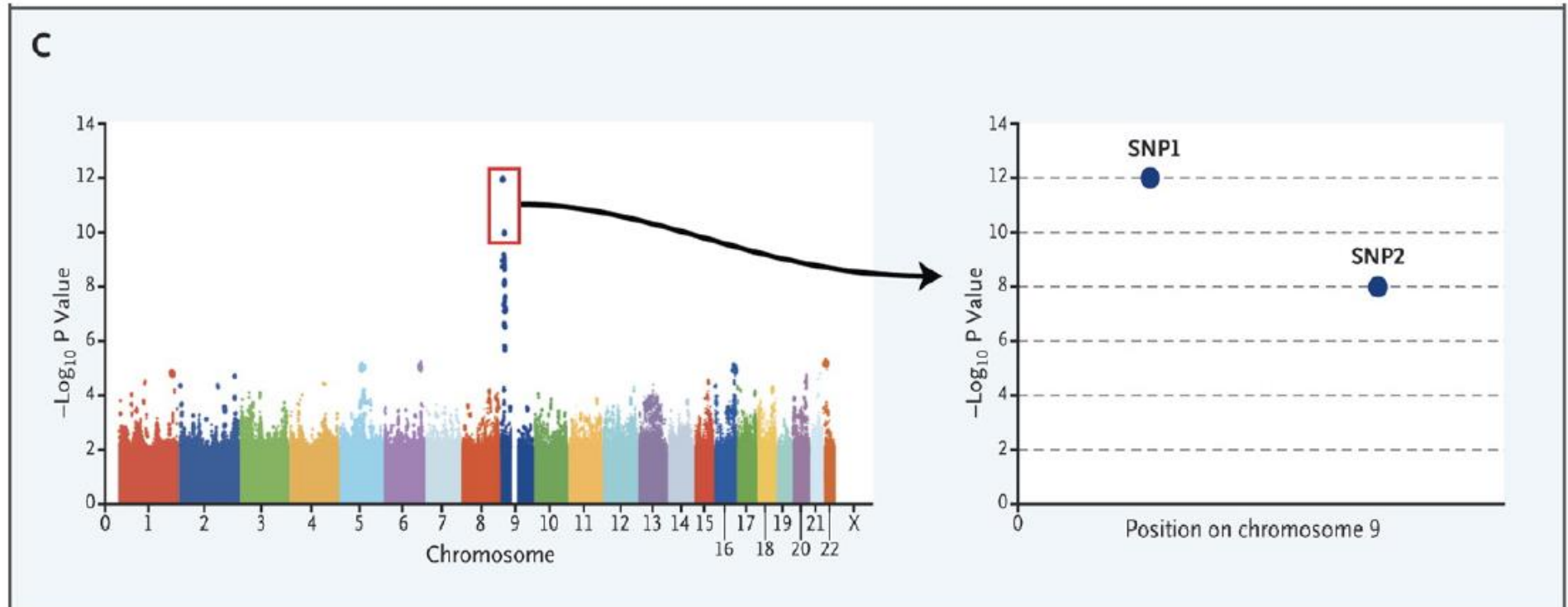


(slide Doug Brutlag 2010)

Population Genetics



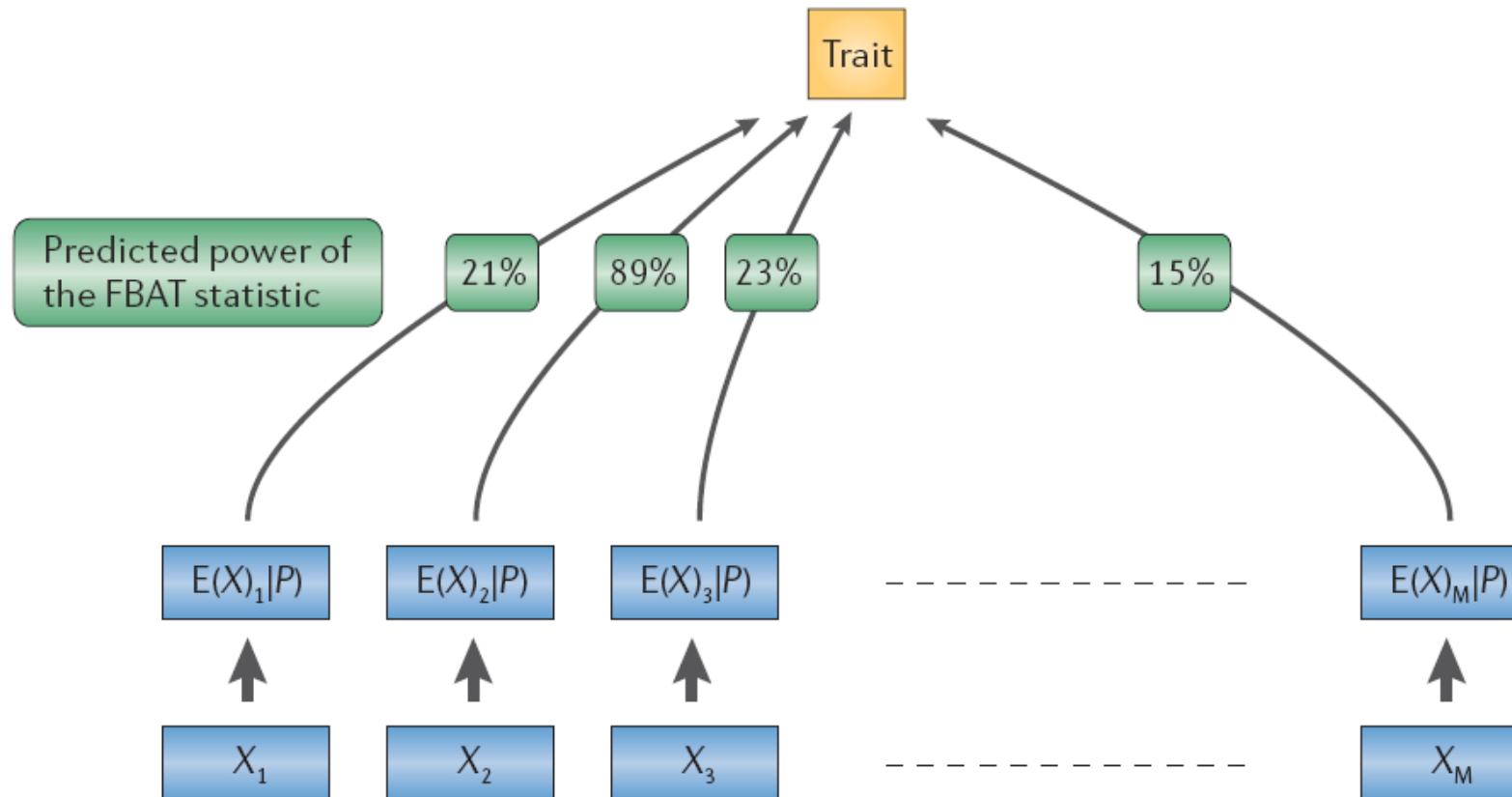
GWAS with SNPs



Exemplar Manhattan plot

Family-based GWAS with SNPs

PBAT screening



(Lange and Laird 2006)

Complexities – as of interest via group work

REVIEWS

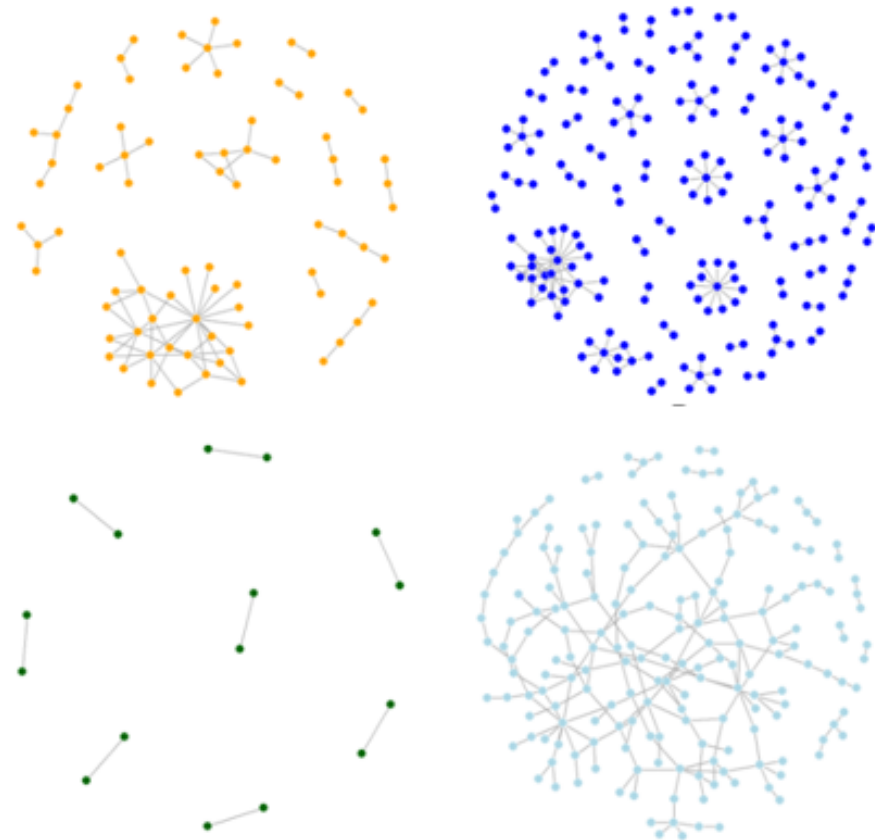
GENOME-WIDE ASSOCIATION STUDIES

Meta-analysis methods for genome-wide association studies and beyond

Evangelos Evangelou¹ and John P. A. Ioannidis^{2,3}

Abstract | Meta-analysis of genome-wide association studies (GWASs) has become a popular method for discovering genetic risk variants. Here, we overview both widely applied and newer statistical methods for GWAS meta-analysis, including issues of interpretation and assessment of sources of heterogeneity. We also discuss extensions of these meta-analysis methods to complex data. Where possible, we provide guidelines for researchers who are planning to use these methods. Furthermore, we address special issues that may arise for meta-analysis of sequencing data and rare variants. Finally, we discuss challenges and solutions surrounding the goals of making meta-analysis data publicly available and building powerful consortia.

GWIS – GxG interactions with SNPs



(Melograna et al. 2021 - ongoing)

Applications of genetic epidemiology in precision medicine

A) Risk scoring (prevention)

natureprotocols

[Explore content](#) ▾ [Journal information](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature protocols](#) > [review articles](#) > [article](#)

Review Article | [Published: 24 July 2020](#)

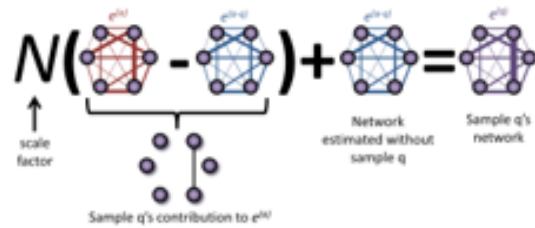
Tutorial: a guide to performing polygenic risk score analyses

[Shing Wan Choi](#), [Timothy Shin-Heng Mak](#) & [Paul F. O'Reilly](#) 

Nature Protocols **15**, 2759–2772(2020) | [Cite this article](#)

20k Accesses | **36** Citations | **86** Altmetric | [Metrics](#)

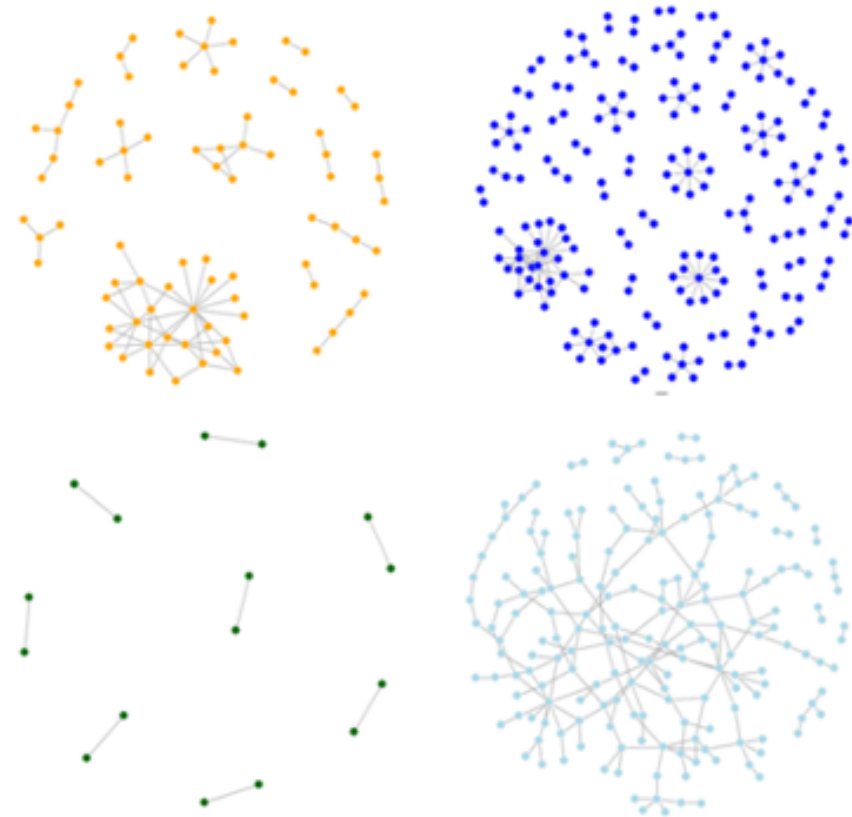
B) Individual-specific networks (diagnosis & disease management)



Attention points :

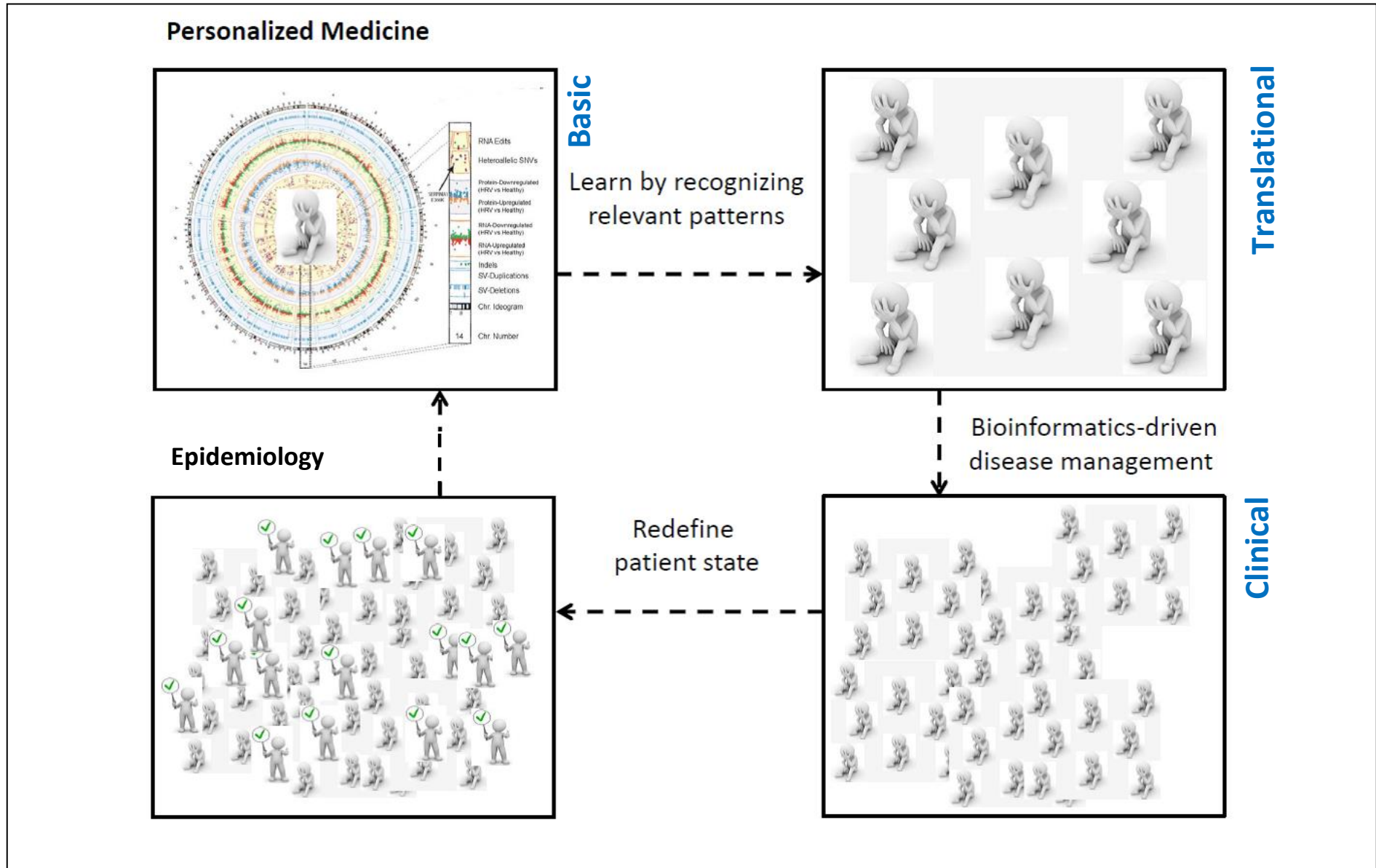
- Strong assumptions on the type of edge weights
- Limiting definitions of a sample's "influence"
- Limiting assessments of how "significant" an individual's influence is

(Kuijjer et al. 2019)

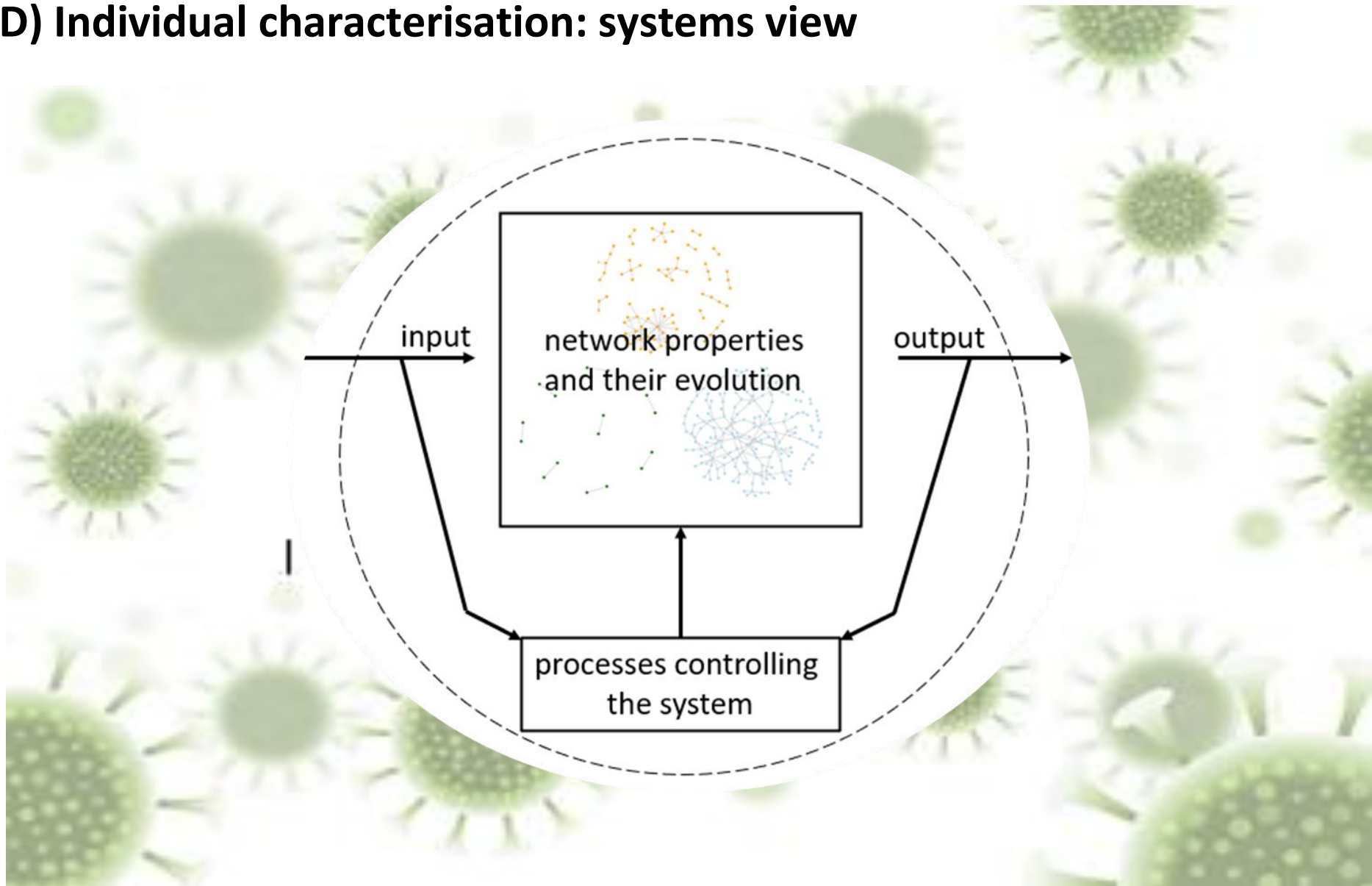


(Melograna et al. 2021 - ongoing)

C) Molecular reclassification of disease (disease management)



D) Individual characterisation: systems view



E) Debate

Focus on particular aspects, including:

Prosperi et al. *BMC Medical Informatics and Decision Making* (2018) 18:139
<https://doi.org/10.1186/s12911-018-0719-2>

BMC Medical Informatics and
Decision Making

DEBATE

Open Access

Big data hurdles in precision medicine and precision public health



Mattia Prospero^{1*} , Jae S. Min¹, Jiang Bian² and François Modave³

5.b Interesting reads

TOPIC 1 Methods in genetic epi via DOI: 10.1023/a:1024933620315

- Focus on population-based studies

TOPIC 2 Case study Asthma GWA

via <https://doi.org/10.4168/aair.2019.11.2.170>

- Focus on pros and cons rather than the medically relevant results

TOPIC 3 Case study Asthma GxE via <https://doi.org/10.1016/j.jaci.2012.10.038>

- Focus on state of the art

TOPIC 4 Case study Asthma GxE via <https://doi.org/10.1016/j.jaci.2012.10.038>

- Focus on challenges

TOPIC 5 Novel developments in genetic epidemiology, start with DOI:
10.1016/S0140-6736(05)67601-5

- Focus on precision medicines, biobanks and apps

Questions & Answers