Genetic Epidemiology

Kristel Van Steen, PhD[2]

Systems Genomics – GIGA-R, ULg

Systems Medicine – Human Genetics, KU Leuven

kristel.vansteen@uliege.be

Topics in Genetic Epidemiology

1 Setting the pace

1.a What can your spit tell you about your DNA?

1.b Speaking the language: relevant questions and concepts

1.c “The Human Genome Project” and its context

2 The rise of Genome-wide association studies

3 Study Design Elements

3.a Marker level

3.b Subject level

4 Pre-analysis Steps

4.a Quality-Control

4.b Linkage disequilibrium

4.c Confounding by shared genetic ancestry

5 Analysis Steps

5.a Association modeling

5.b Causation versus association

6 Post Association Analysis Steps

6.a Replication and validation

6.b GWA Interpretation and follow-up

7 Epistasis analysis

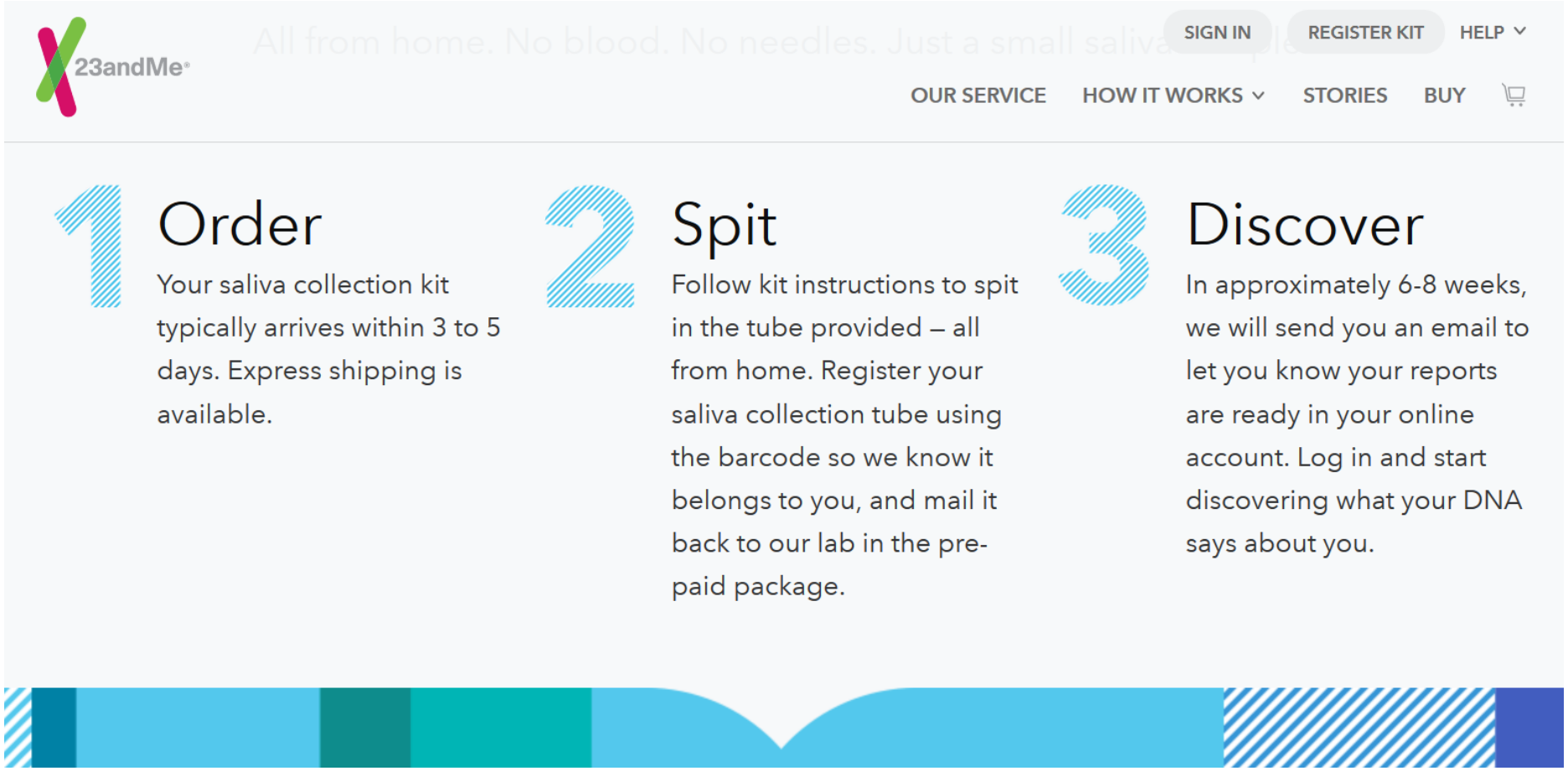
1 Setting the pace

1.a What can your spit tell you about your DNA?

From saliva to DNA

- Your saliva contains a veritable mother load of biological material from which your genetic blueprint can be determined.
- For example, a mouthful of spit contains hundreds of complex protein molecules – enzymes -- that aid in the digestion of food.

- Swirling around with those enzymes are cells sloughed off from the inside of your cheek.
- Inside each of those cells lies a nucleus, and inside each nucleus, chromosomes, which themselves are made up of DNA


23andMe
SIGN IN
REGISTER KIT
HELP
OUR SERVICE
HOW IT WORKS
STORIES
BUY
1
Order
Your saliva collection kit typically arrives within 3 to 5 days. Express shipping is available.
2
Spit
Follow kit instructions to spit in the tube provided – all from home. Register your saliva collection tube using the barcode so we know it belongs to you, and mail it back to our lab in the pre-paid package.
3
Discover
In approximately 6-8 weeks, we will send you an email to let you know your reports are ready in your online account. Log in and start discovering what your DNA says about you.

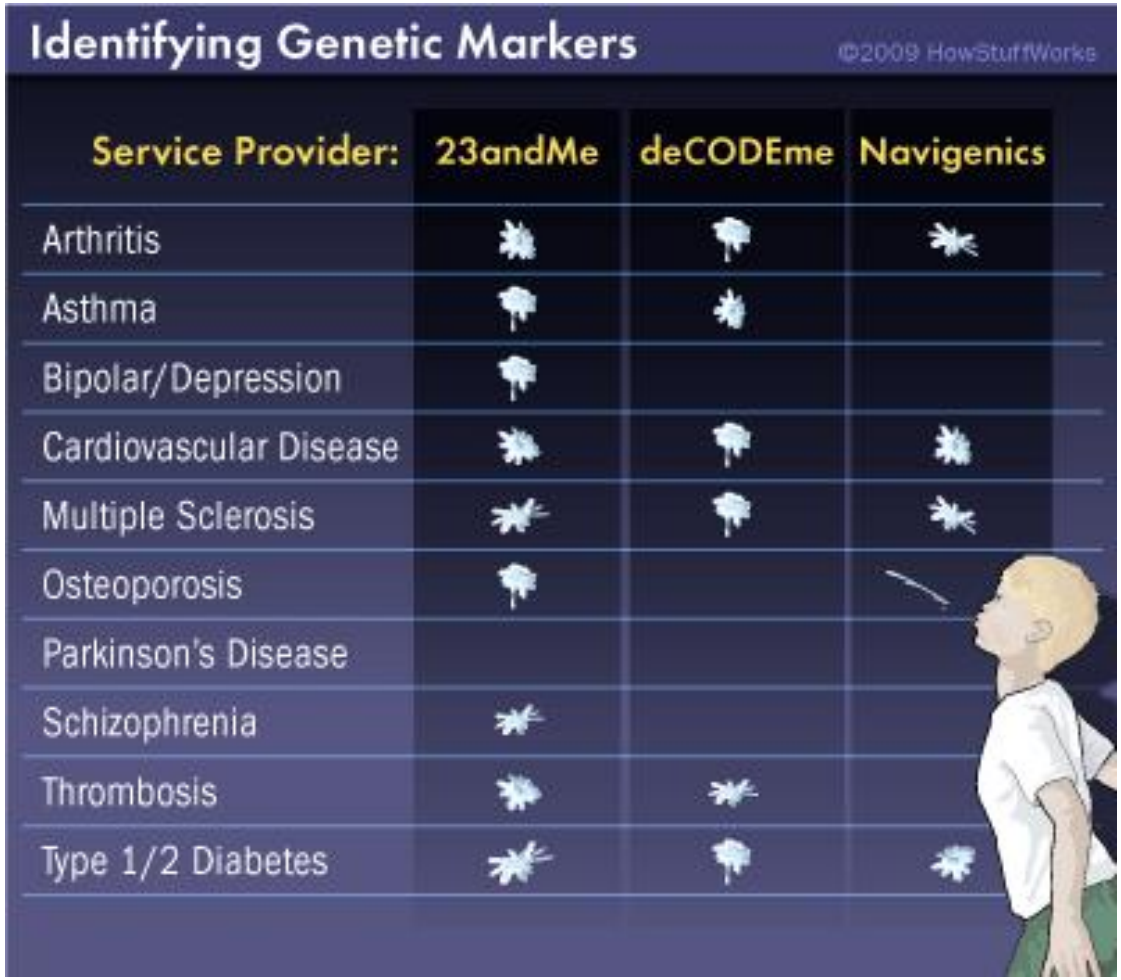

Print this page
Your estimated lifetime risk
Click anywhere on the colored boxes below to access in-depth information about each health condition, your genetic predispositions, what you can do, your specific genetic markers, and much more.
0 - 1%
>1 - 10%
>10 - 25%
>25 - 50%
>50 - 100%
Brain aneurysm
You: 0.91%
Avg: 0.90%
Alzheimer's disease
You: 10%
Avg: 17%
Heart attack
You: 22%
Avg: 25%
Osteoarthritis
You: 47%
Avg: 47%
You have no results in this range
Rheumatoid arthritis
You: 0.88%
Avg: 3.3%
Lung cancer
You: 9%
Avg: 6%
Breast cancer
You: 14%
Avg: 13%
Obesity
You: 38%
Avg: 32%
NEW
Sarcoidosis
You: 0.55%
Avg: 1.0%
Psoriasis
You: 6%
Avg: 4.0%
Diabetes, type 2
You: 37%
Avg: 30%
Macular degeneration
You: 0.44%
Avg: 3.1%
Colon cancer
You: 3.6%
Avg: 5%
Atrial fibrillation
You: 29%
Avg: 23%
NEW
Multiple sclerosis
You: 0.28%
Deep vein thrombosis
You: 2.9%
Key to your results
Condition name
Diabetes type 2
You: 45%
Avg: 30%
Your results
Population Average
Why orange & gray boxes?
Video: Understanding your results
Tutorial: Review the tutorial
More: How we estimate your risk
Your genetic counselor
Counselors are available weekdays from 9am to 5pm PST, or you can schedule another time convenient for you.
Call (866) 522-1585
International:
+1 (650) 585-7743
Sharing results with your doctor

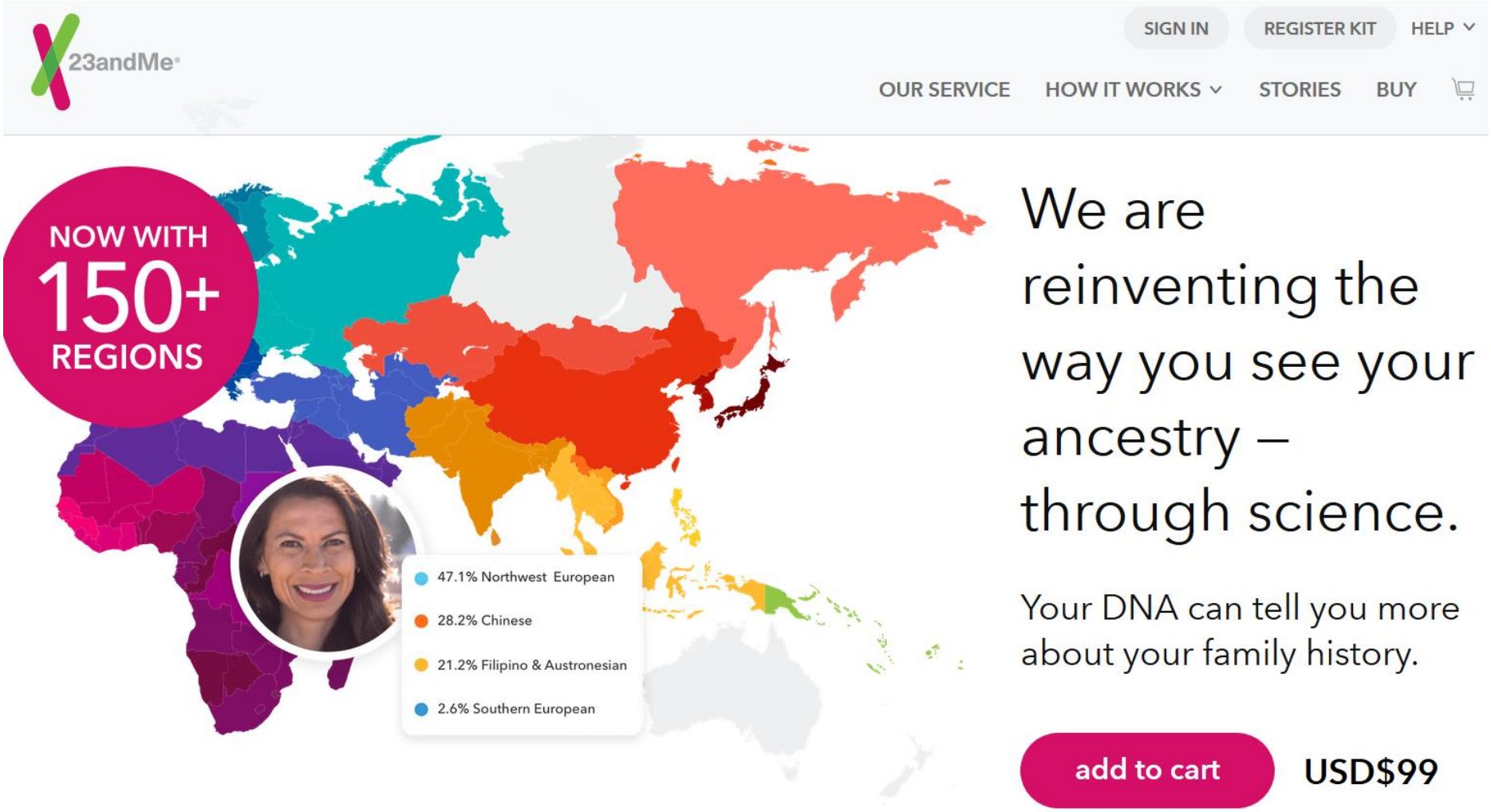
The 23andMe story

- Wojcicki founded 23andme in 2006 with Linda Avey and Paul Cusenza with a goal of upending conventional models of health care:
  - put sophisticated DNA analyses into the hands of consumers,
  - giving them information about health, disease and ancestry,
  - and allowing the company to sell access to the genetic data to fuel research.
- In 2013, that vision hit a snag. Wojcicki didn't think she needed regulatory approval to provide information about her customers' health risks. The US Food and Drug Administration (FDA) disagreed, and ordered the company to stop.

(source: https://www.nature.com/news/the-rise-and-fall-and-rise-again-of-23andme-1.22801)

Can you handle the truth?




23andMe
SIGN IN
REGISTER KIT
HELP
OUR SERVICE
HOW IT WORKS
STORIES
BUY
NOW WITH
150+
REGIONS
47.1% Northwest European
28.2% Chinese
21.2% Filipino & Austronesian
2.6% Southern European
We are reinventing the way you see your ancestry – through science.
Your DNA can tell you more about your family history.
add to cart
USD$99

The 23andMe story

- After years of effort, the pay-off came in April 2017, when the FDA agreed to allow 23andme to tell consumers their risks of developing ten medical conditions, including Parkinson's disease and late-onset Alzheimer's disease.

- With more than 2 million customers, the company hosts by far the largest collection of gene-linked health data anywhere

(source: https://www.nature.com/news/the-rise-and-fall-and-rise-again-of-23andme-1.22801)

How many types of genetic tests exist?

- There are >2000 genetic tests available to physicians to aid in the diagnosis and therapy for >1000 different diseases. Genetic testing is performed for the following reasons:

  - conformational diagnosis of a symptomatic individual
  - pre-symptomatic testing for estimating risk developing disease
  - pre-symptomatic testing for predicting disease
  - **prenatal screening**
  - newborn screening
  - preimplantation genetic diagnosis
  - carrier screening
  - **forensic testing**
  - **paternal testing**

How is genetic testing used clinically?

- **Diagnostic medicine**: identify whether an individual has a certain genetic disease. This type of test commonly detects a specific gene alteration but is often not able to determine disease severity or age of onset. It is estimated that there are >4000 diseases caused by a mutation in a single gene. Examples of diseases that can be diagnosed by genetic testing includes cystic fibrosis and Huntington's disease.

- **Predictive medicine**: determine whether an individual has an increased risk for a particular disease. Results from this type of test are usually expressed in terms of probability and are therefore less definitive since disease susceptibility may also be influenced by other genetic and non-genetic (e.g. environmental, lifestyle) factors. Examples of diseases that use genetic testing to identify individuals with increased risk include certain forms of breast cancer (BRCA) and colorectal cancer.

How is genetic testing used clinically?

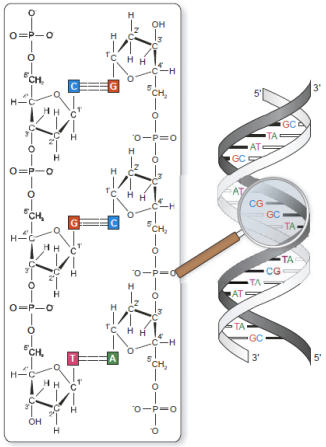
- **Pharmacogenomics**: classifies subtle variations in an individual's genetic makeup to determine whether a drug is suitable for a particular patient, and if so, what would be the safest and most effective dose. Learn more about pharmacogenomics & precision medicine → DNA passports … are no longer science fiction!

- **Whole-genome and whole-exome sequencing**: examines the entire genome or exome to discover genetic alterations that may be the cause of disease. Currently, this type of test is most often used in complex diagnostic cases, but it is being explored for use in asymptomatic individuals to predict future disease → increasingly feasible by improved technology + reduced costs → more adequate reference genomes

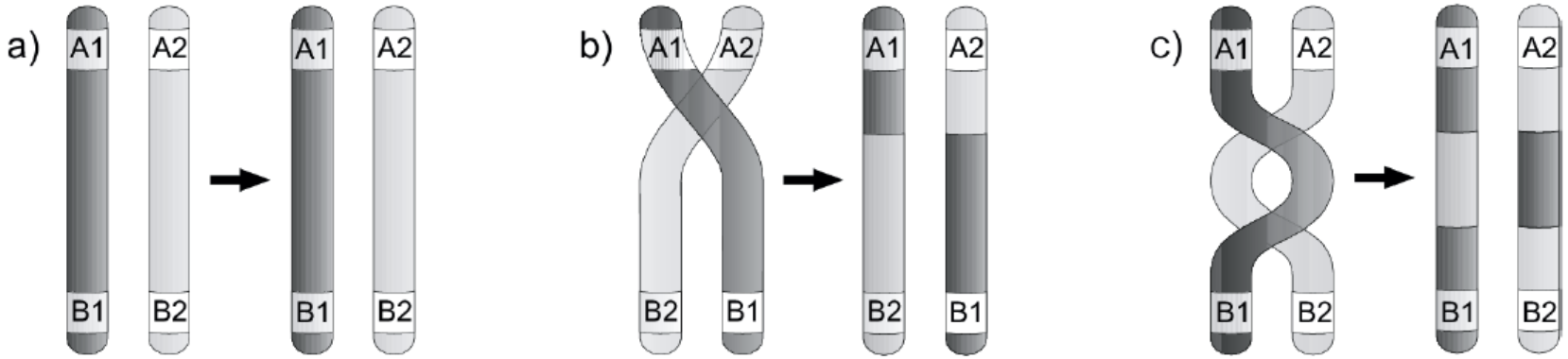
1.b Speaking the language

Where is the genetic information located?

- Cell has nucleus
- Nucleus carries genetic information in chromosomes
- Chromsomes composed of desoxyribonucleic acid (DNA) and proteins
- DNA large molecule consisting in two strands
- Each strand has backbone of sugar and phosphate residues
- Sequence of bases attached to backbone
- Bases: adenine (A), guanine (G), cytosine (C), thymine (T)
- Strands connected through hydrogen bonds
  - A with T (2 hydrogen bonds)
  - C with G (3 hydrogen bonds)

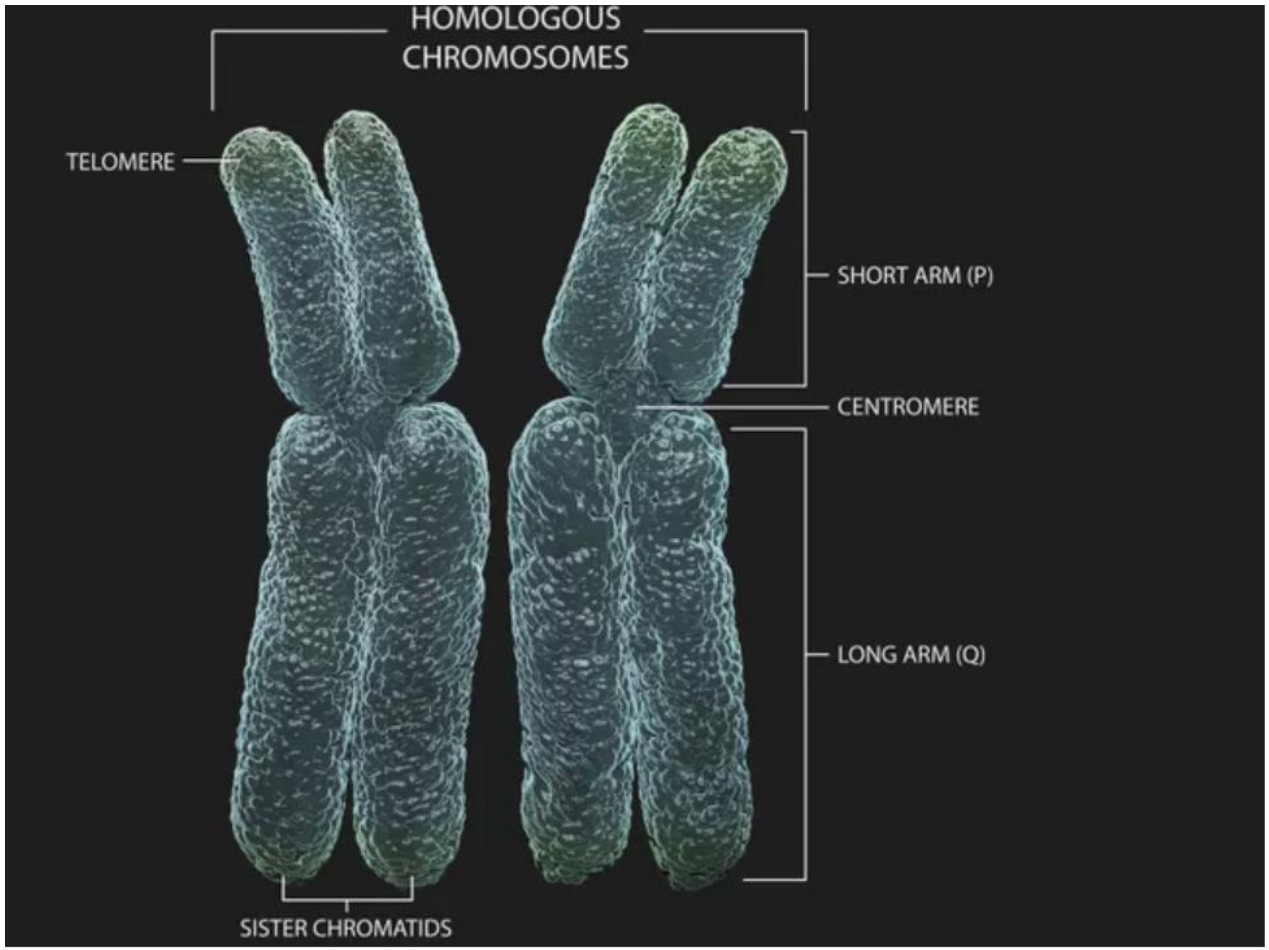


Result of crossover: recombination in meiotic products



- Relevant measure: recombination fraction (probability of odd number of crossovers) between two chromosomal positions
- Strong correlation between recombination fraction and distance in base pairs

(Ziegler and Van Steen, Brazil 2010)



Photon Illustration/Stocktrek Images/Getty Images

In humans, males have lower recombination rates than females over the majority of the genome, but the opposite is usually true near the telomeres

How much do individuals differ with respect to genetic information?

- Allele: one of several alternative forms of DNA sequence at specific chromosomal location (locus)
- Genetic marker: polymorphic DNA sequence at single locus
- Polymorphism: existence of ≥ 2 alleles at single locus
- Homozygosity (homozygous): both alleles identical at locus
- Heterozygosity (heterozygous): different alleles at locus
- Mutation:
  - Changes allele at specific chromosomal position
  - Frequency ≈ 10^{-4} to 10^{-6} ⇒ Individuals differ with freq. of 1/1000 bases

(Ziegler and Van Steen, Brazil 2010)

How much do individuals differ with respect to genetic information?

- **Genotype**: The two alleles inherited at a specific locus. If the alleles are the same, the genotype is homozygous, if different, heterozygous. In genetic association studies, genotypes can be used for analysis as well as alleles or haplotypes.
- **Haplotype**: Linear arrangements of alleles on the same chromosome that have been inherited as a unit. A person has two haplotypes for any such series of loci, one inherited maternally and the other paternally. A haplotype may be characterized by a single allele

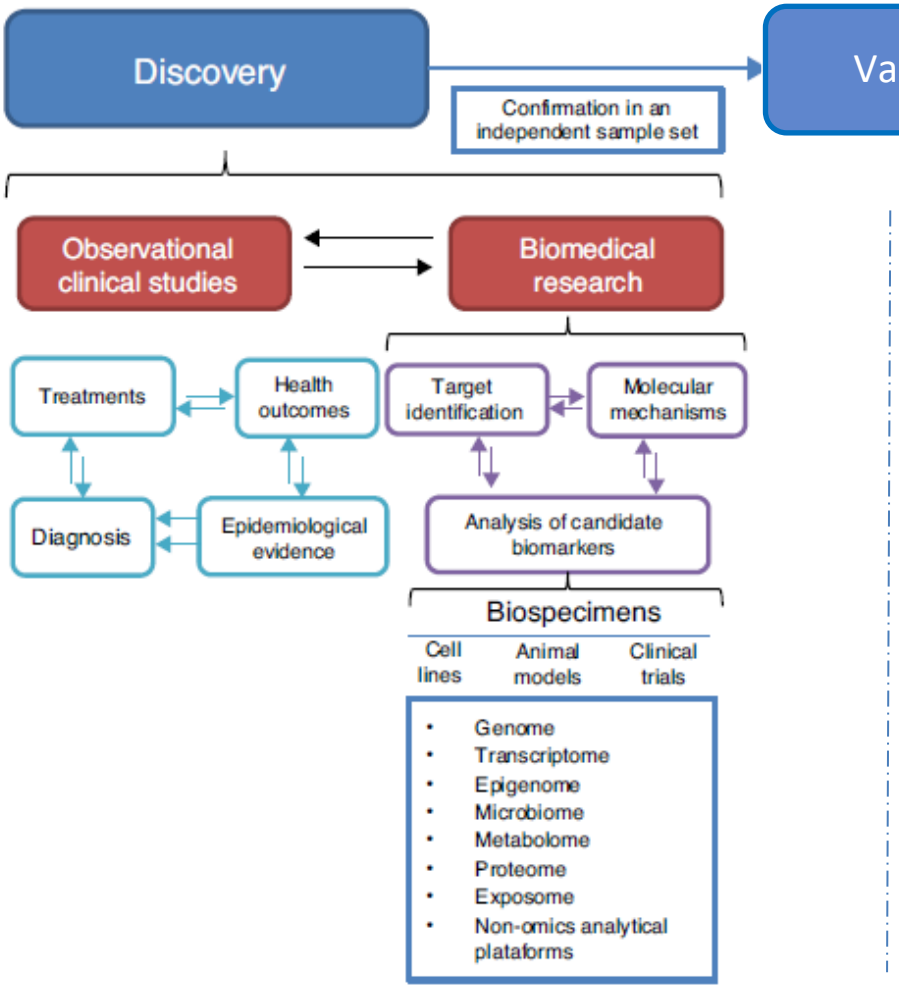
http://www.dorak.info/epi/glosge.html

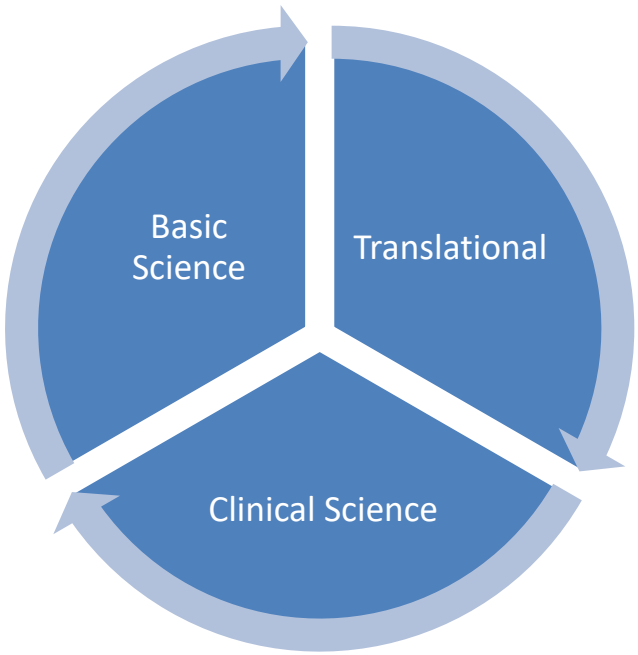
What are biomarkers?

- A biological marker, or biomarker, is something that can be measured, which points to the presence of a disease, a physiological change, response to a treatment, or a psychological condition.
- A molecular biomarker is a molecule that can be used in this way. Recall that **DNA is a molecule**! → **genetic markers** = polymorphic DNA sequences at a locus
- Biomarkers are used in different ways at different stages of medicines development, including in some cases as a surrogate endpoint to indicate and measure the effect of medicines in clinical trials → also genetic markers

(www.eupati.eu)

The biomarker development process





(Quezada et al. 2017)

What are the most popular genetic markers?

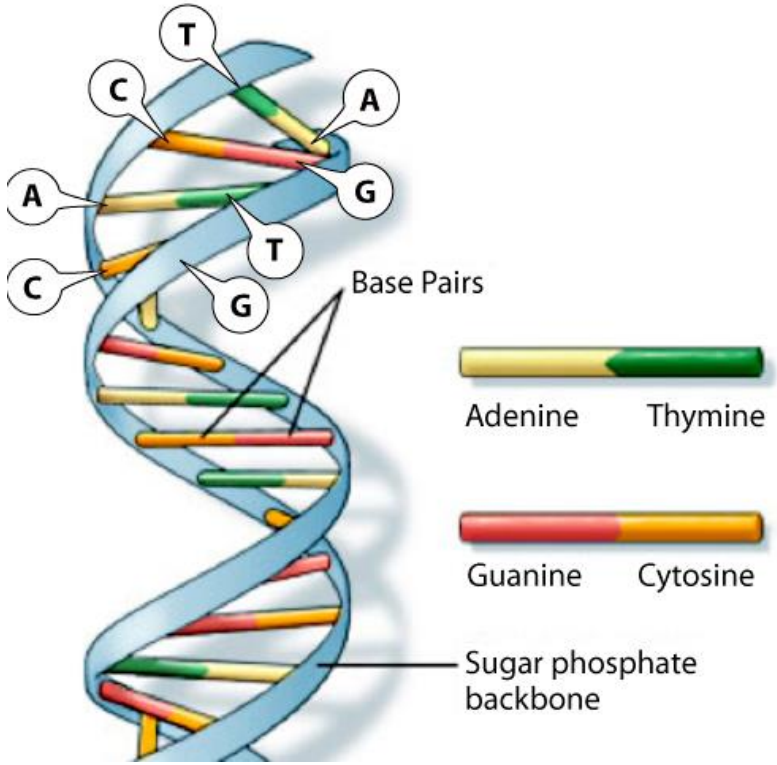
Single Nucleotide Polymorphisms (SNPs)

- Variations in single base, i.e., one base substituted by another base
- In theory: four different nucleotides possible at base
- In practice: generally only two different nucleotides observed
- Definition strict and loose:
  - Strict: minor allele frequency ≥ 1%
  - Loose: ≥ 2 nucleotides observed in two individuals at position
- Nomenclature:
  - ss-number (submitted SNP number)
  - rs-number: searchable in dbSNP, mapped to external resources, unique
  - rs-numbers do not provide information about possible function of SNP
  - Alternative: nomenclature of Human Genome Variation Society

(Ziegler and Van Steen, Brazil 2010)

Do SNPs capture differences between human genomes?

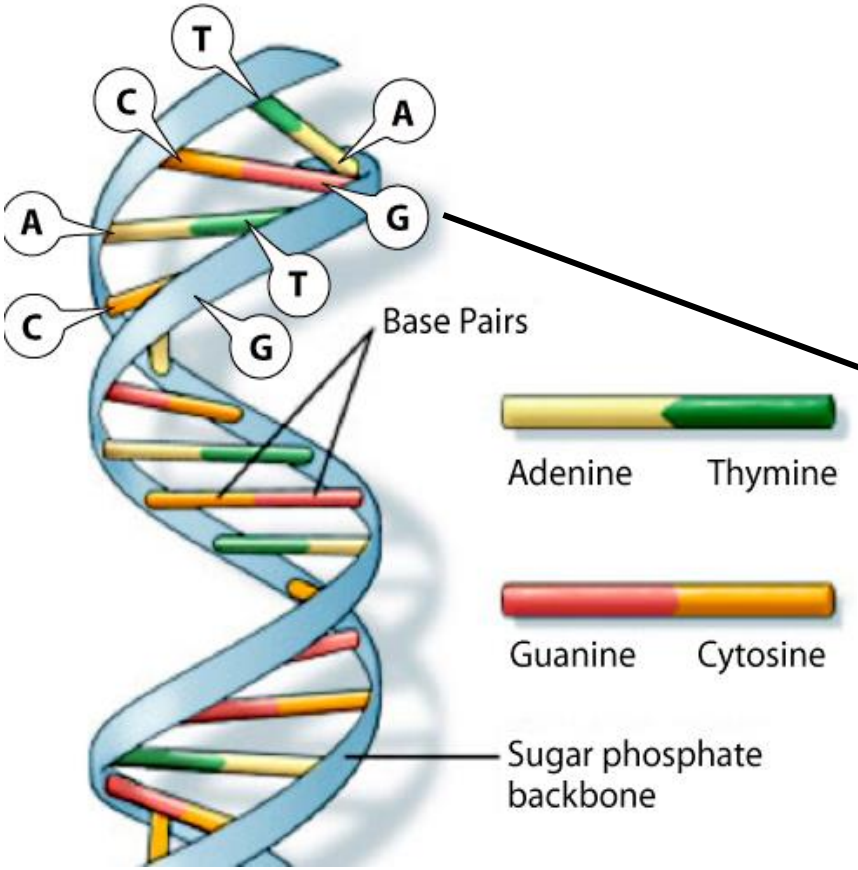
- Any two people plucked at random off the street are on average 99.9 percent the same, DNA-wise (> 3 million positional differences)
- Most genome variations are relatively small and simple, involving only a few bases—an A substituted for a T here, a G left out there, a short sequence such as CG added somewhere else



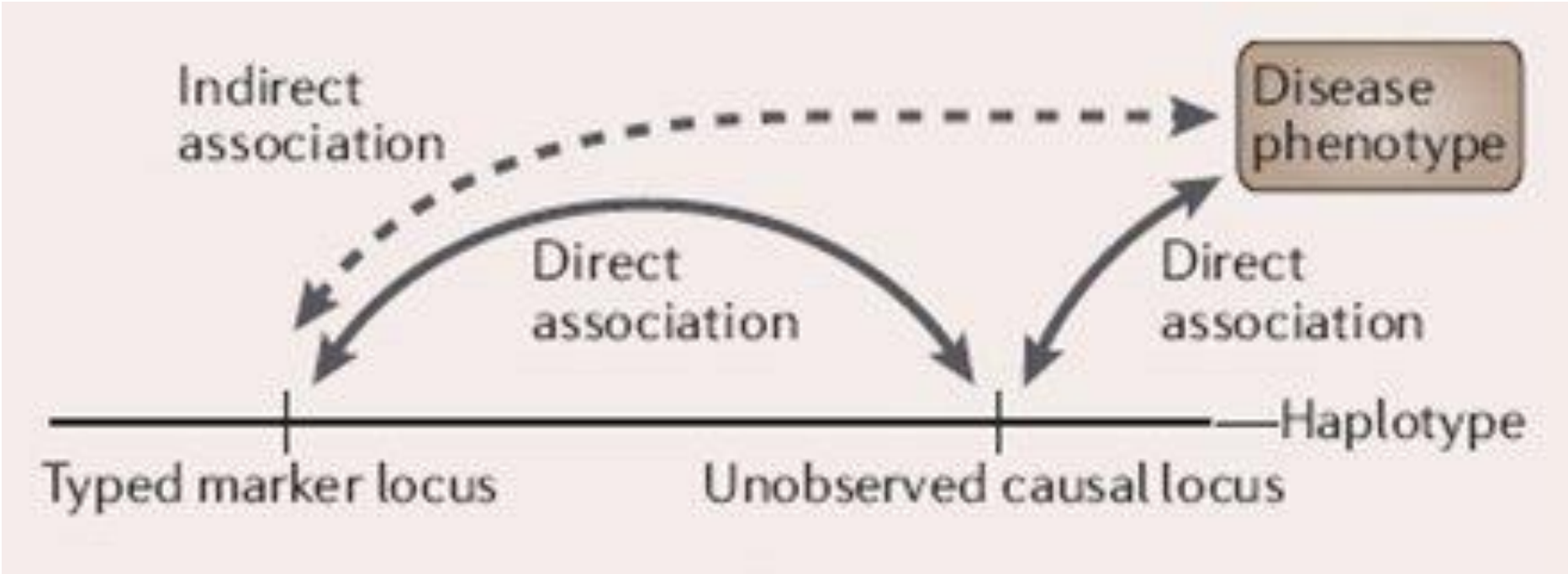
(U.S. National Library of Medicine)

Common genetic variations

MAF (minor allele frequency)



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

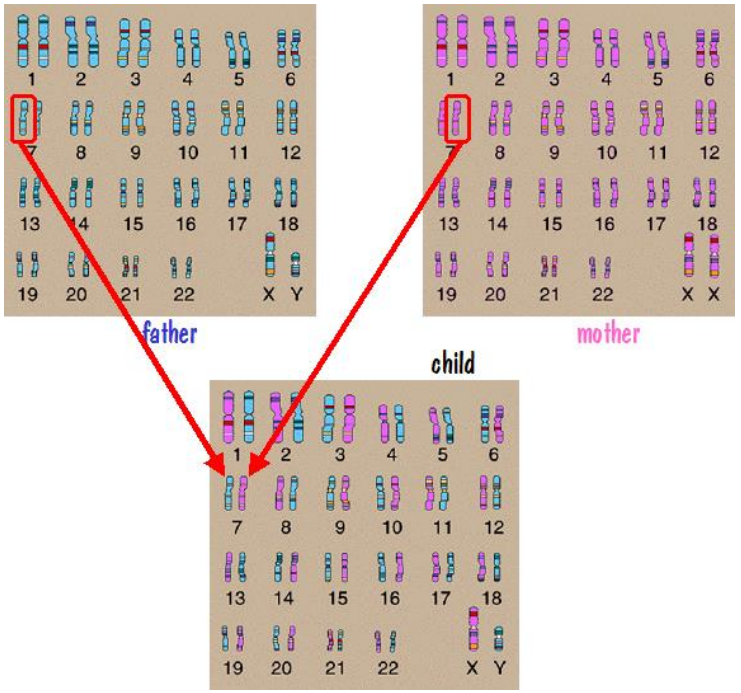


(Balding 2006)

What are genes?

Defining genes as units of inheritance

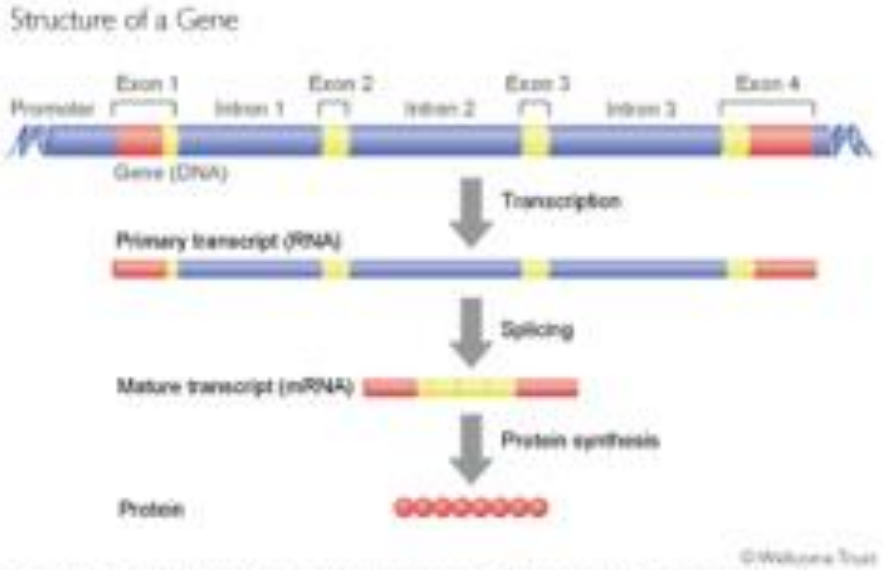
- The **gene** is the basic physical unit of inheritance.
- Genes are passed from parents to offspring and contain the information needed to specify traits.
- They are arranged, one after another, on the chromosomes
- Chromosomes are not taken entirely by genes.



(Figure : Human chromosomes)

What are genes?

Defining genes by their structure



Simplified overview of gene structure and expression (for eukaryotes). A protein-coding gene is defined by the extent of the primary transcript. The gene is first transcribed to yield a primary transcript, which is processed to remove the introns. The mature transcript (messenger RNA, mRNA) is then translated into a sequence of amino acids, which defines the protein. The chain of amino acids must fold up to generate the final tertiary structure of the protein

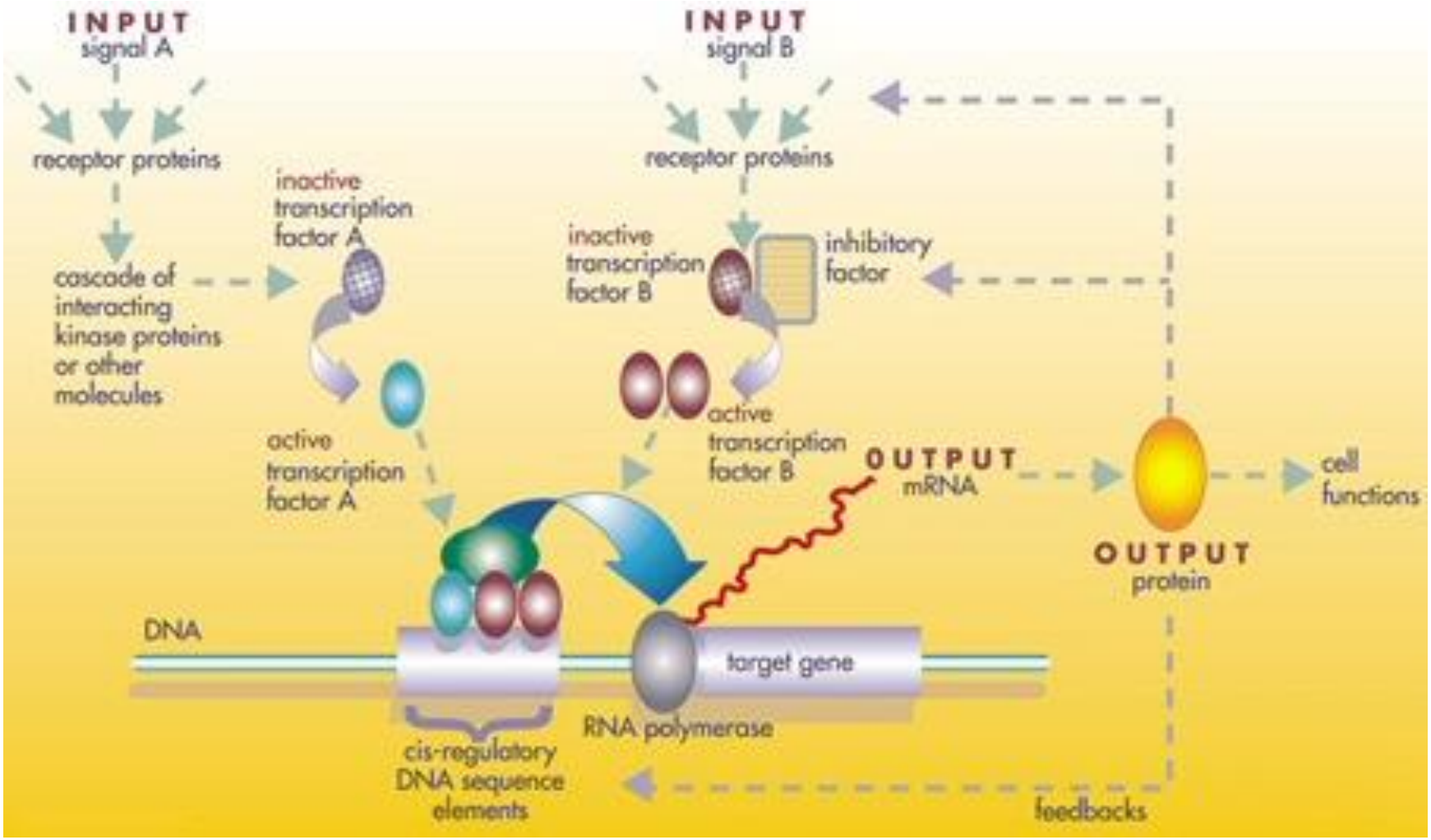
Splicing is carried out by a very complex enzyme machinery: **the spliceosome**. In the spliceosome, proteins as well as RNA molecules are found that form complexes: the small nuclear ribonucleoproteins or snRNPS (snurps). These recognize specific sequences on the borders of an intron, cut the ends, release the intron and ligate the remaining exons.

What is gene annotation?

- An annotation (irrespective of the context) is a note added by way of explanation or commentary.
- **Genome annotation** is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do.
- Once a genome is sequenced, it needs to be annotated to make sense of it

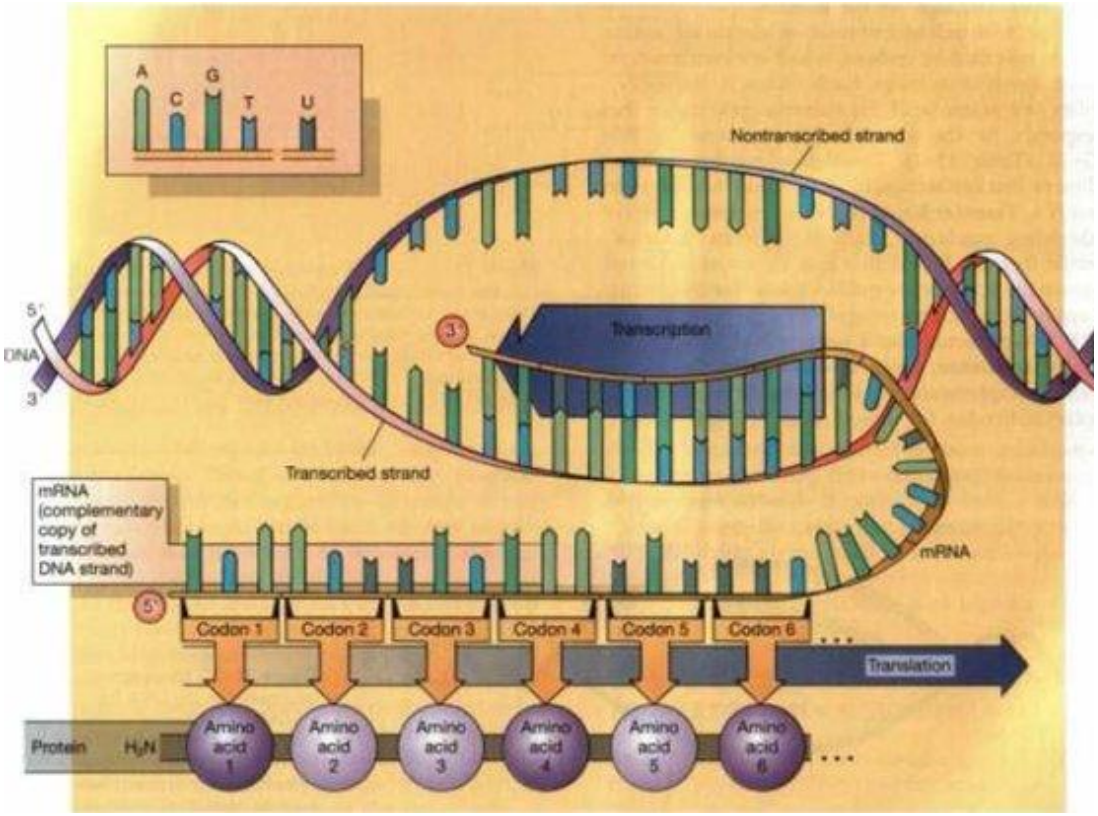
What is gene regulation?

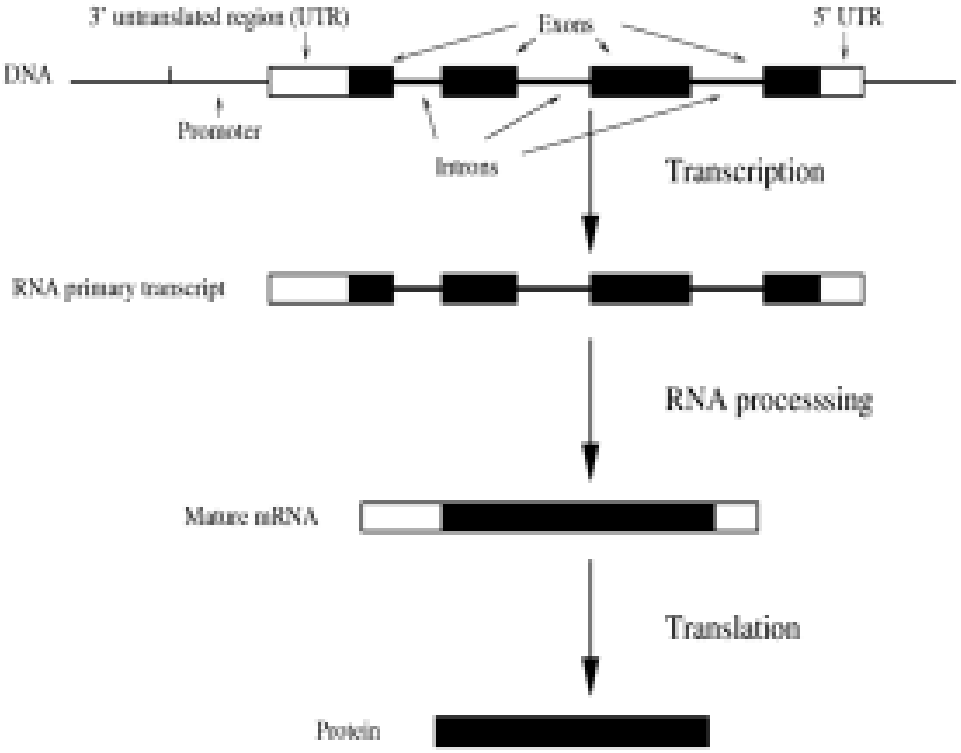
revised version of central dogma



What is the central dogma of molecular biology?

Simplified version





Information is everywhere: the programming of life

http://www.youtube.com/watch?v=00vBqYDBW5s

"Information:

that which can be communicated through symbolic language"

What is genetic epidemiology?

*"… Examining 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**"*

*"It is the discipline investigating genetic and environmental factors that influence the development and distribution of diseases. It **differs from epidemiology** in that explicitly genetic factors and similarities within families are taken into account. On the other hand, it can be **distinguished from medical genetics** by considering populations rather than single patients or families."*

(Ziegler and Van Steen, Brazil 2010)

What is genetic epidemiology?

Hard to define!

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)

What are the 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.

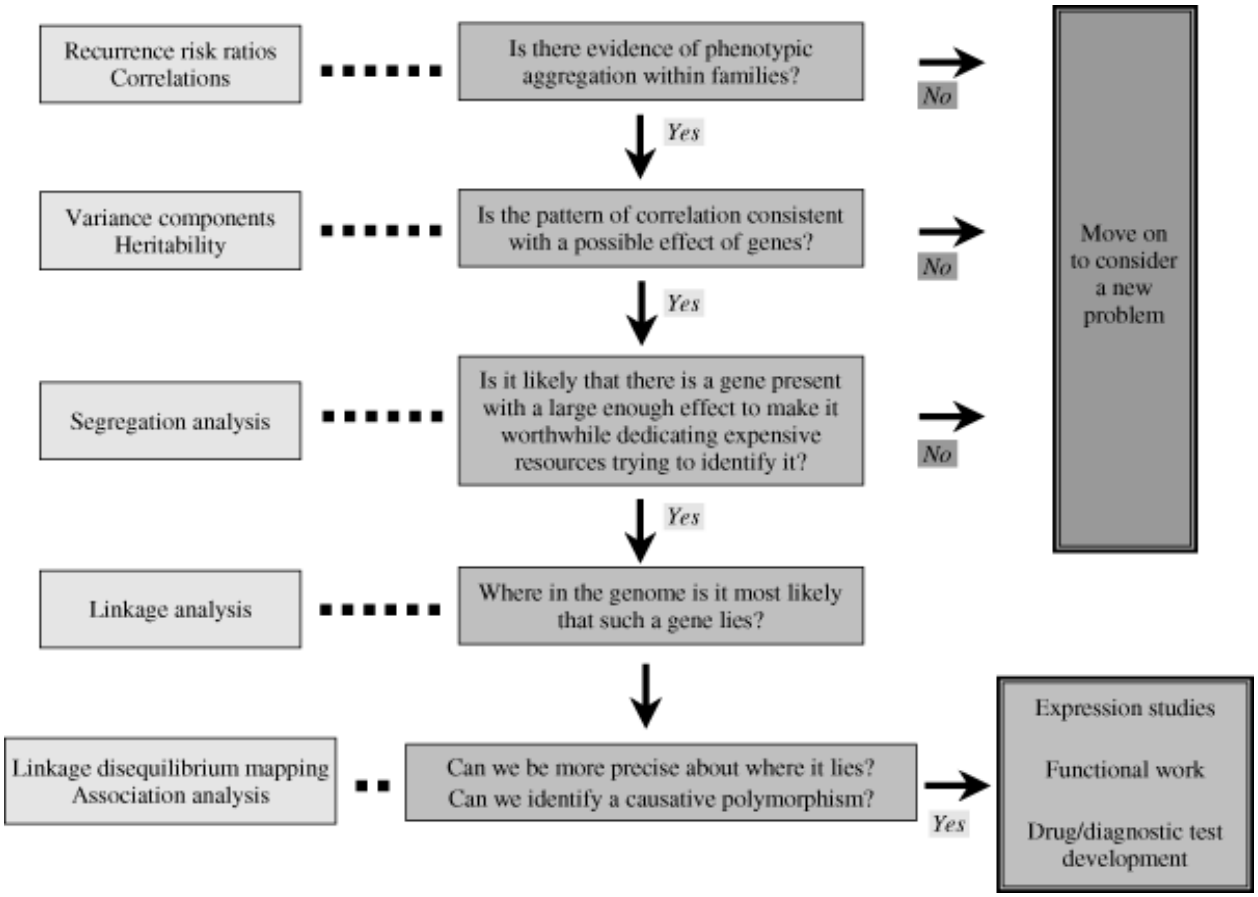
Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, UK (Prof P R Burton MD, M D Tobin PhD); and Centre for Genetic Epidemiology, University of Melbourne, Melbourne, Victoria, Australia (Prof J L Hopper PhD)

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.

What are relevant questions in genetic epidemiology?



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

How to hunt for genes to answer relevant questions ?

- Developing new and better tools to make gene hunts faster, cheaper and practical for any scientist was a primary goal of the **Human Genome Project** (HGP).
- One of these tools is **genetic mapping**, the first step in isolating a gene. Genetic mapping – in the early days - can offer firm evidence that a disease transmitted from parent to child is **linked** to one or more genes. It also provides “clues” about where the gene lies.
- Genetic maps have been used successfully to find the single gene responsible for relatively rare inherited disorders, like cystic fibrosis, but have also been useful as a guide to identify the possible many genes underlying more common disorders, like **asthma**.

How to generate a genetic map?

- Initially, to produce a genetic map, researchers collect blood or tissue samples from **family members** where a certain disease or trait is prevalent.
- Using various laboratory techniques, the scientists isolate DNA from these samples and examine it for the unique patterns seen only in family members who have the disease or trait.
- Before researchers identify the gene responsible for the disease or trait, DNA markers can tell them roughly where the gene is on the chromosome.

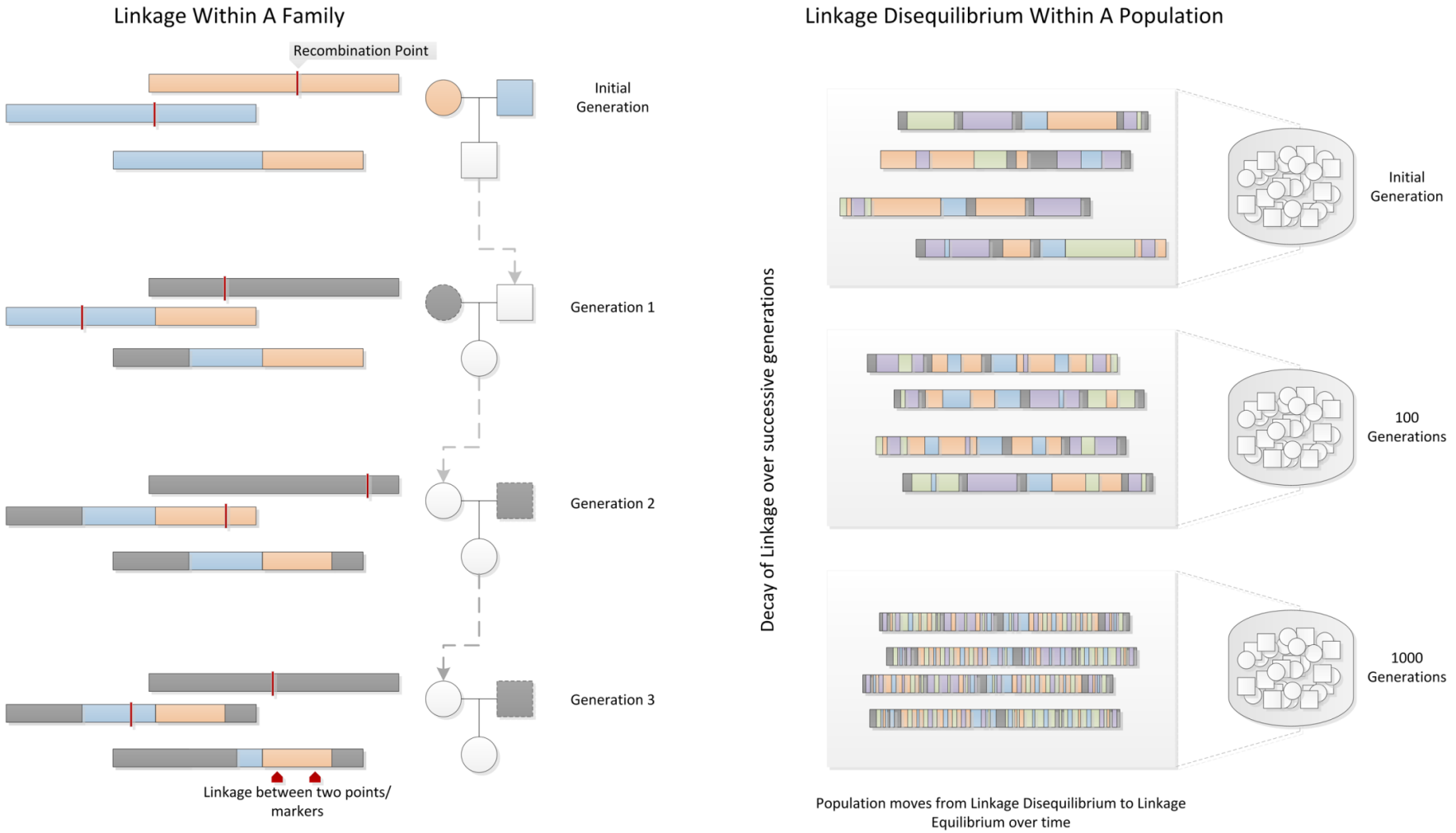
How is this possible?

How to generate a genetic map? (continued)

- This is possible because of recombination, the process we have introduced before.

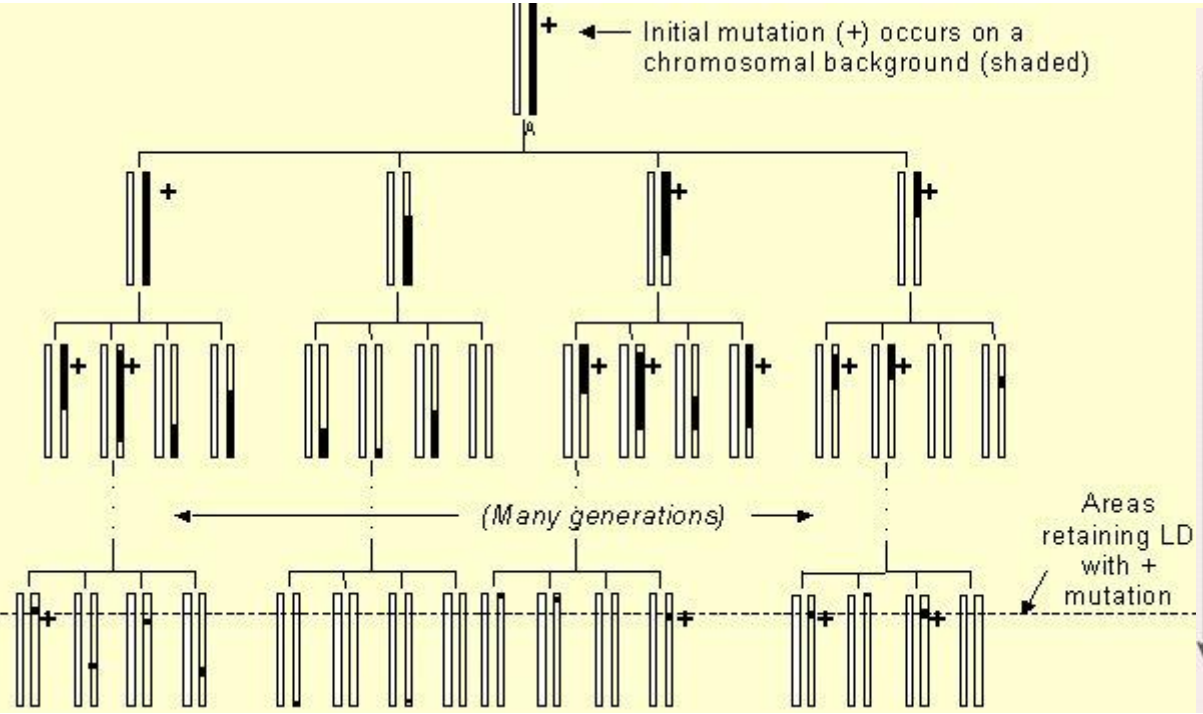
> *As eggs or sperm develop within a person's body, the 23 pairs of chromosomes within those cells exchange - or recombine - genetic material. If a particular gene is close to a DNA marker, the gene and marker will likely stay together during the recombination process, and be passed on together from parent to child. So, if each family member with a particular disease or trait also inherits a particular DNA marker, chances are high that the gene responsible for the disease lies near that marker.*

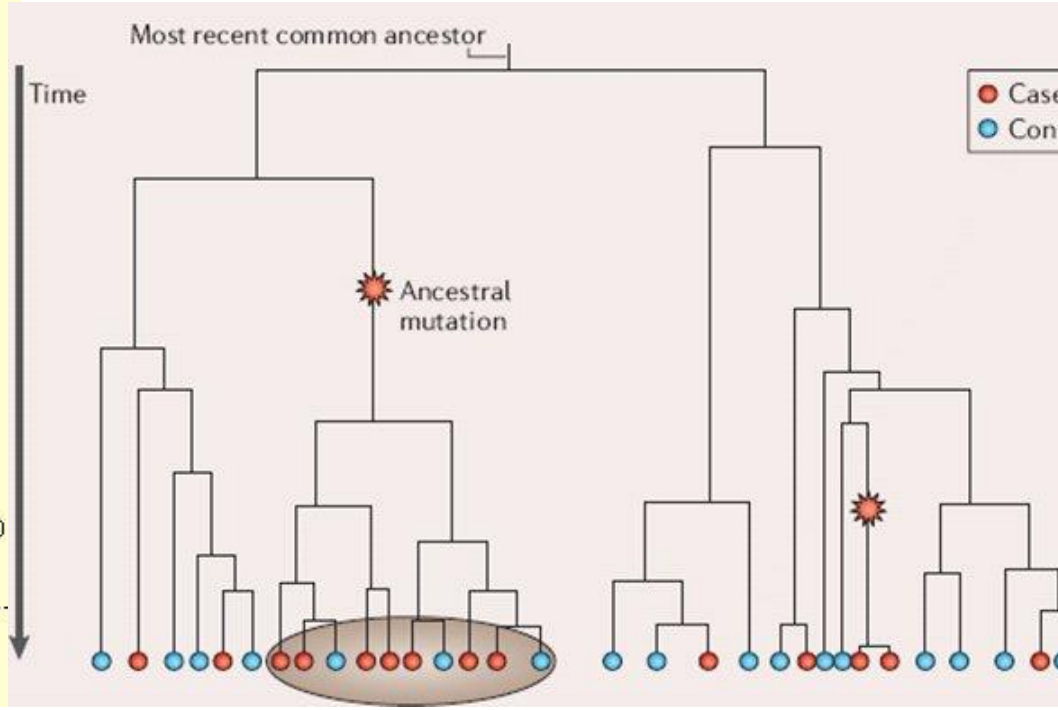
How to generate a genetic map? (What is Linkage Disequilibrium – LD?)



(Bush et al. 2012)

How to generate a genetic map? (What is Linkage Disequilibrium – LD?)





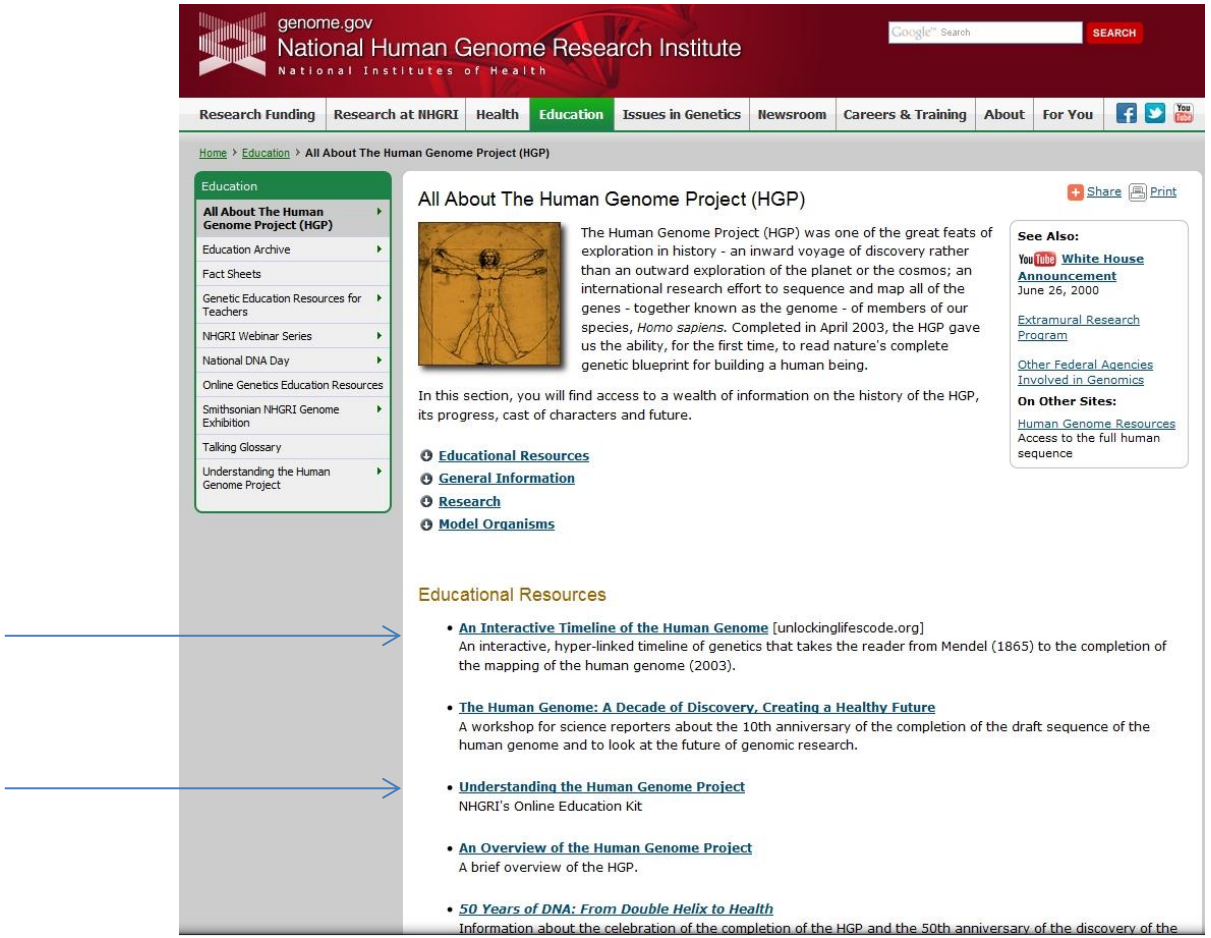
How to generate a genetic map? (continued)

- The more DNA markers there are on a genetic map, the more likely it is that one will be closely linked to a disease gene - and the easier it will be for researchers to zero-in on that gene.

- One of the **first major achievements of the HGP was to develop dense maps of markers spaced evenly across the entire collection of human DNA**.

(http://www.genome.gov/10000715#al-3)

1.c “The Human Genome Project”



Historical overview (interludium)

Gregor Mendel, the father of modern genetics, presents his research on experiments in plant hybridization

Gregor Mendel, a 19th century Augustinian monk, is called the father of modern genetics. He used a monastery garden for crossing pea plant varieties having different heights, colors, pod shapes, seed shapes, and flower positions. Mendel's experiments, between 1856 and 1863, revealed how traits are passed down from parents. For example, when he crossed yellow peas with green peas, all the offspring peas were yellow. But when these offspring reproduced, the next generation was ¾ yellow and ¼ green. Mendel's work, which was presented in 1865, showed that what we now call "genes" determine traits in predictable ways.

1865

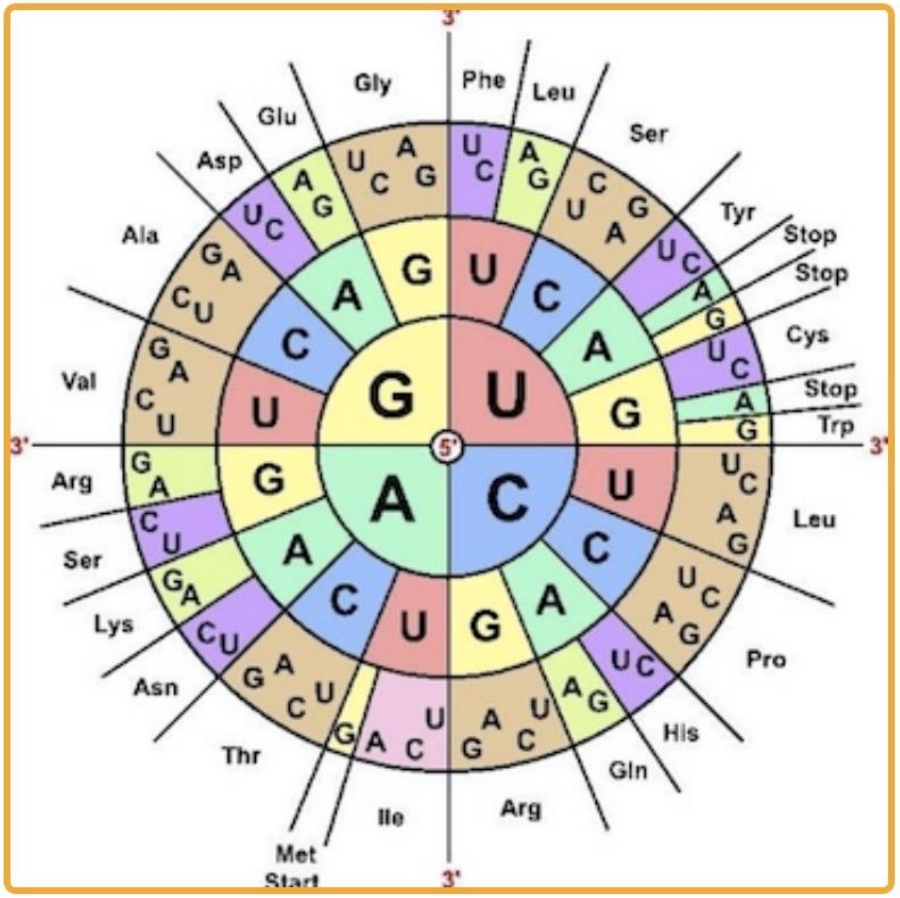
Historical overview

James Watson and Francis Crick discover the double helix structure of DNA

When Francis Crick and James Watson modeled the structure of DNA, they used paper cutouts of the bases (A, C, G, T) and metal scraps from a machine shop. Their model represented DNA as a double helix, with sugars and phosphates forming the outer strands of the helix and the bases pointing into the center. Hydrogen bonds connect the bases, pairing A–T and C–G; and the two strands of the helix are parallel but oriented in opposite directions. Their 1953 paper notes that the model "immediately suggests a possible copying mechanism for the genetic material."

1953

Historical overview

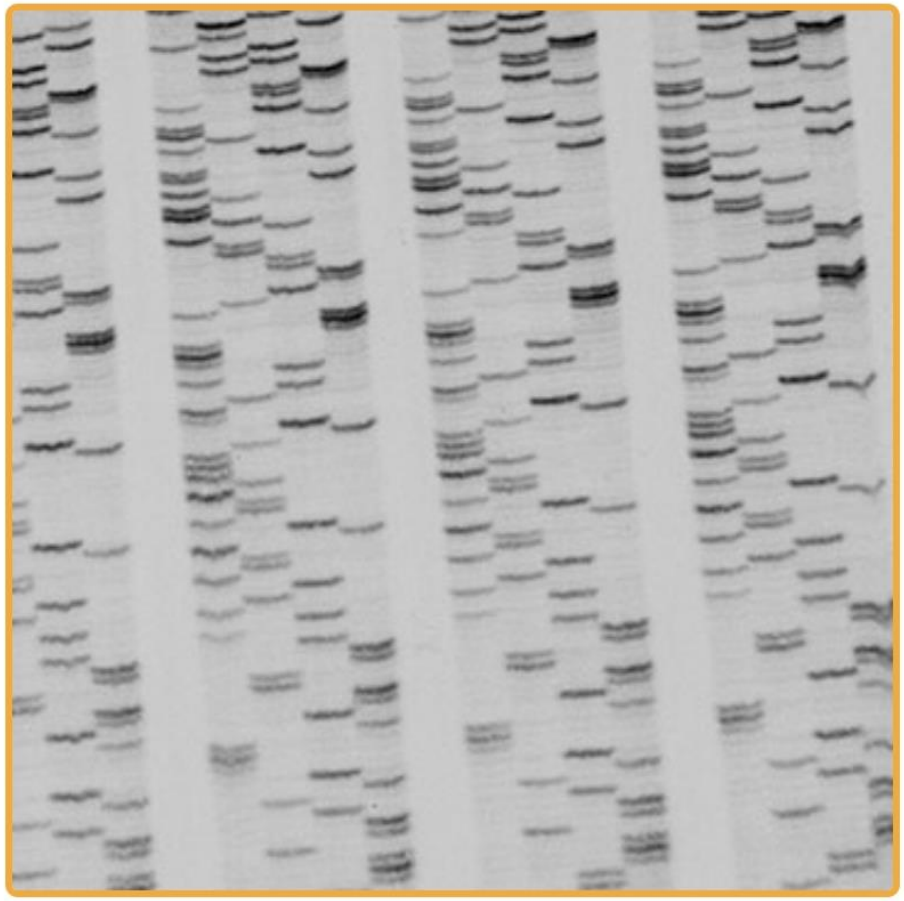


Marshall Nirenberg cracks the genetic code for protein synthesis

In the early 1960s, Marshall Nirenberg and National Institutes of Health colleagues focused on how DNA directs protein synthesis and the role of RNA in these processes. Their 1961 experiment, using a synthetic messenger RNA (mRNA) strand that contained only uracils (U), yielded a protein that contained only phenylalanines. Identifying UUU (three uracil bases in a row) as the RNA code for phenylalanine was their first breakthrough. Within a few years, Nirenberg's team had cracked the 60 mRNA codons for all 20 amino acids. In 1968, Nirenberg shared the Nobel Prize in Physiology or Medicine for his contributions to breaking the genetic code and understanding protein synthesis.

1961

Historical overview

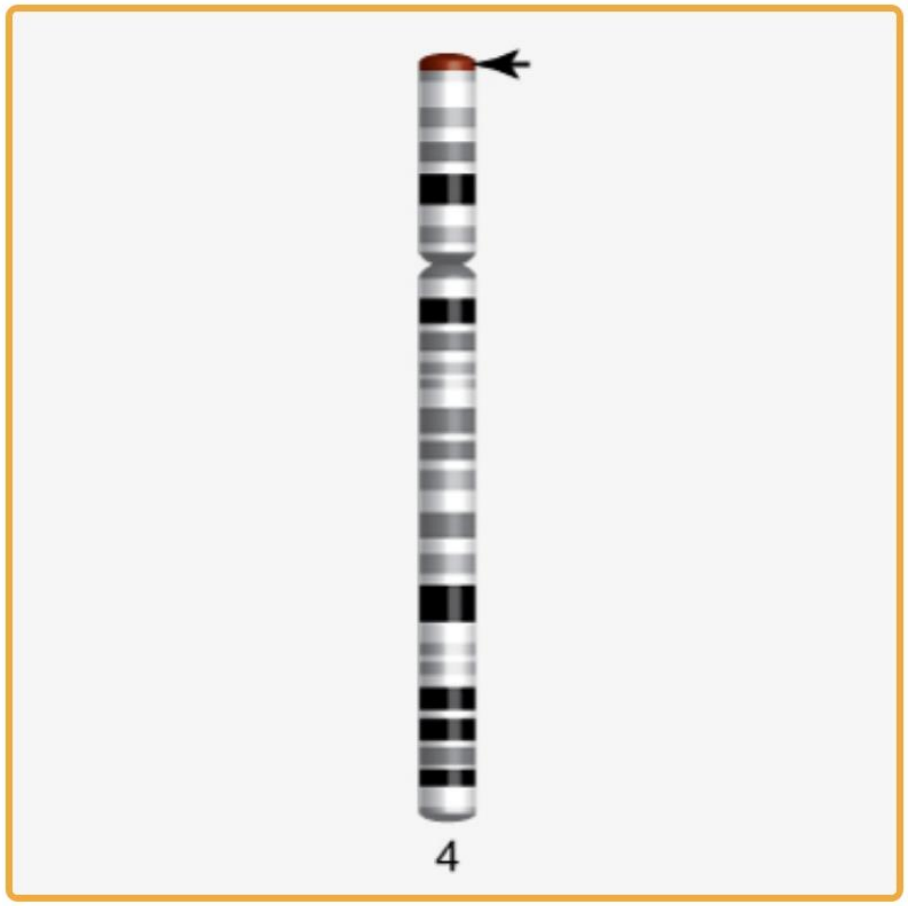


Frederick Sanger develops rapid DNA sequencing technique

In 1977, Frederick Sanger developed the classical "rapid DNA sequencing" technique, now known as the Sanger method, to determine the order of bases in a strand of DNA. Special enzymes are used to synthesize short pieces of DNA, which end when a selected "terminating" base is added to the stretch of DNA being synthesized. Typically, each of these terminating bases is tagged with a radioactive marker, so it can be identified. Then the DNA fragments, of varying lengths, are separated by how rapidly they move through a gel matrix when an electric field is applied – a technique called electrophoresis. Frederick Sanger shared the 1980 Nobel Prize in Chemistry for his contributions to DNA-sequencing methods.

1977

Historical overview

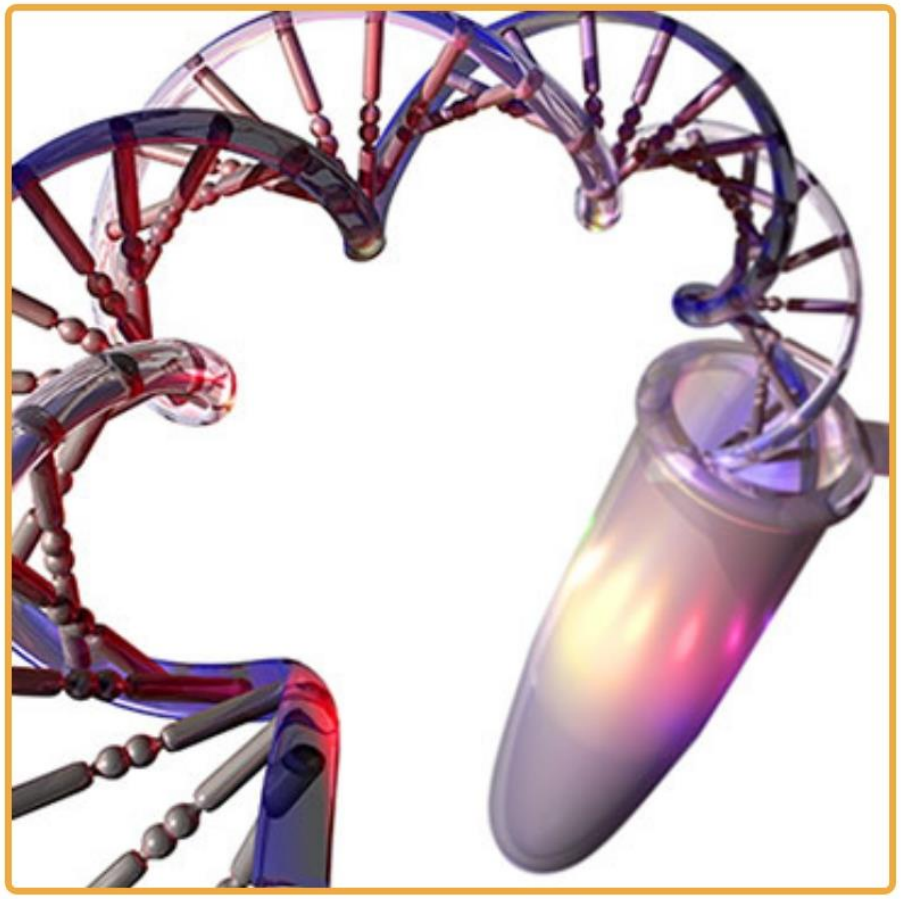


First genetic disease mapped, Huntington's Disease

Huntington's disease (HD) causes the death of specific neurons in the brain, leading to jerky movements, physical rigidity, and dementia. Symptoms usually appear in midlife and worsen progressively. The location of the HD gene, whose mutation causes Huntington's disease, was mapped to chromosome 4 in 1983, making HD the first disease gene to be mapped using DNA polymorphisms – variants in the DNA sequence. The mutation consists of increasing repetitions of "CAG" in the DNA that codes for the protein huntingtin. The number of CAG repeats may increase when passed from parent to child, leading to earlier HD onset in each generation. The gene was finally isolated in 1993.

1983

Historical overview

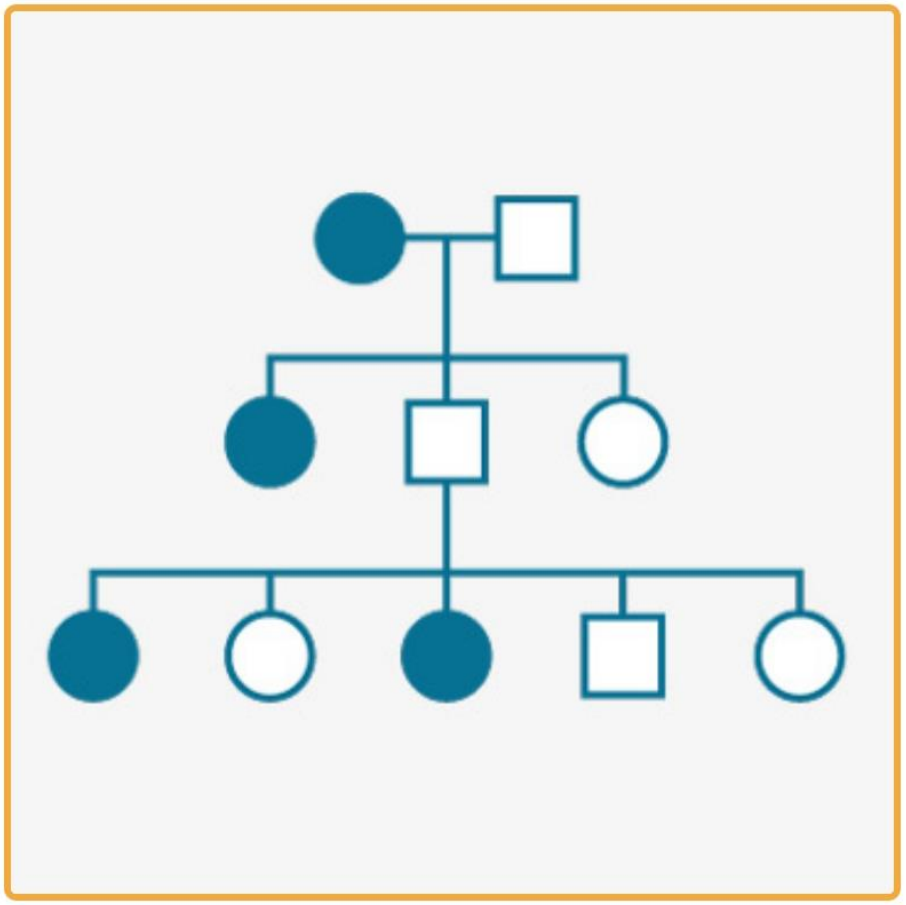


Invention of polymerase chain reaction (PCR) technology for amplifying DNA

Conceived in 1983 by Kary Mullis, the Polymerase Chain Reaction (PCR) is a relatively simple and inexpensive technology used to amplify or make billions of copies of a segment of DNA. One of the most important scientific advances in molecular biology, PCR amplification is used every day to diagnose diseases, identify bacteria and viruses, and match criminals to crime scenes. PCR revolutionized the study of DNA to such an extent that Dr. Mullis was awarded the Nobel Prize in Chemistry in 1993.

1983

Historical overview



First evidence provided for the existence of the BRCA1 gene

BRCA1 (BReast CAncer gene 1) is a "tumor suppressor gene," which normally produces a protein that prevents cells from growing and dividing out of control. However, certain variations of BRCA1 can disrupt its normal function, leading to increased hereditary risk for cancer. The first evidence for existence of the BRCA1 gene was provided in 1990 by the King laboratory at University of California Berkeley. After a heated international race, the gene was finally isolated in 1994. Today, researchers have identified more than 1,000 mutations of the BRCA1 gene, many of them associated with increased risk of cancer, particularly breast and ovarian cancers in women.

1990

Historical overview

The Human Genome Project begins

Beginning in 1984, the U.S. Department of Energy (DOE), National Institutes of Health (NIH), and international groups held meetings about studying the human genome. In 1988, the National Research Council recommended starting a program to map the human genome. Finally, in 1990, NIH and DOE published a plan for the first five years of an expected 15-year project. The project would develop technology for analyzing DNA; map and sequence human and other genomes – including fruit flies and mice; and study related ethical, legal, and social issues.

1990

Historical overview

THE HUMAN GENOME

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[2] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

16 FEBRUARY 2001 VOL 291 SCIENCE www.sciencemag.org

The sequence of the Human Genome – a milestone in modern medicine

- In June 2000 came the announcement that the majority of the human genome had in fact been sequenced, which was followed by the publication of **90 percent of the sequence of the genome's three billion base-pairs** in the journal *Nature*, in February 2001

- Surprises accompanying the sequence publication included:

  - the relatively small **number of human genes**, perhaps as few as **30,000-35,000**;

    Note: 100,000 → 30,000-35,000 → 24,000 → 19,000-20,000

  - the complex architecture of human proteins compared to their homologs - similar genes with the same functions - in, for example, roundworms and fruit flies;

  - the lessons to be taught by repeat sequences of DNA.

Historical overview

2004

March 24, 2004: International Sequencing Consortium Launches Online Resource
The National Human Genome Research Institute announces that the International Sequencing Consortium (ISC) has launched a free, online resource where scientists and the public can get the latest information on the status of sequencing projects for animal, plant and other eukaryotic genomes.

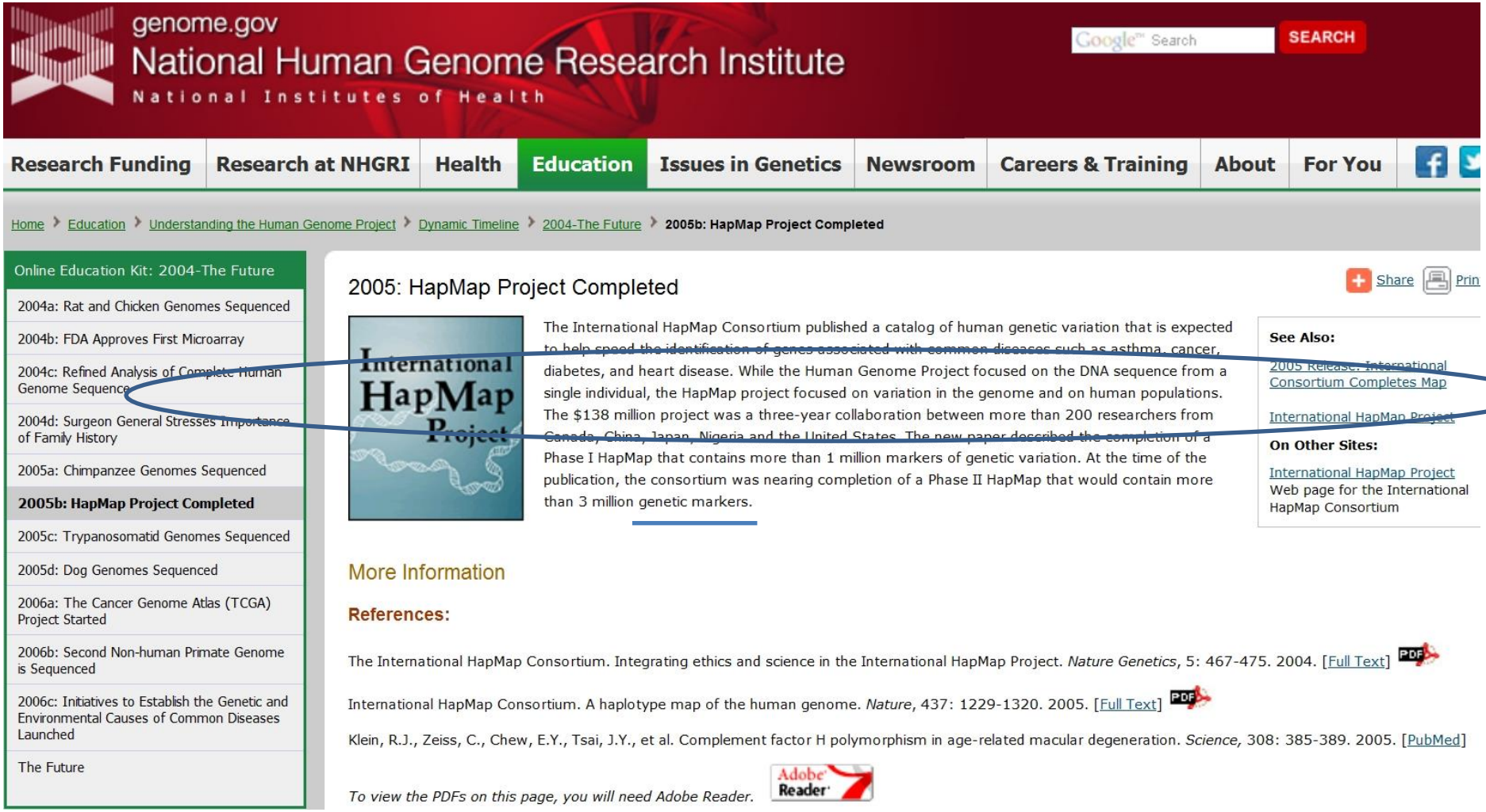
March 31, 2004: Scientists Compare Rat Genome With Human, Mouse
An international research team, supported by the National Institutes of Health (NIH), today announced it has completed a high-quality, draft sequence of the genome of the laboratory rat, and has used that data to explore how the rat's genetic blueprint stacks up against those of mice and humans.

April 21, 2004: NHGRI Scientists Return to the Classroom For Second Annual National DNA Day
On April 30, dozens of researchers and staff from the National Human Genome Research Institute (NHGRI) will head back to high schools in rural and urban communities across the country to share with students some of the exciting research taking place at the National Institutes of Health (NIH).

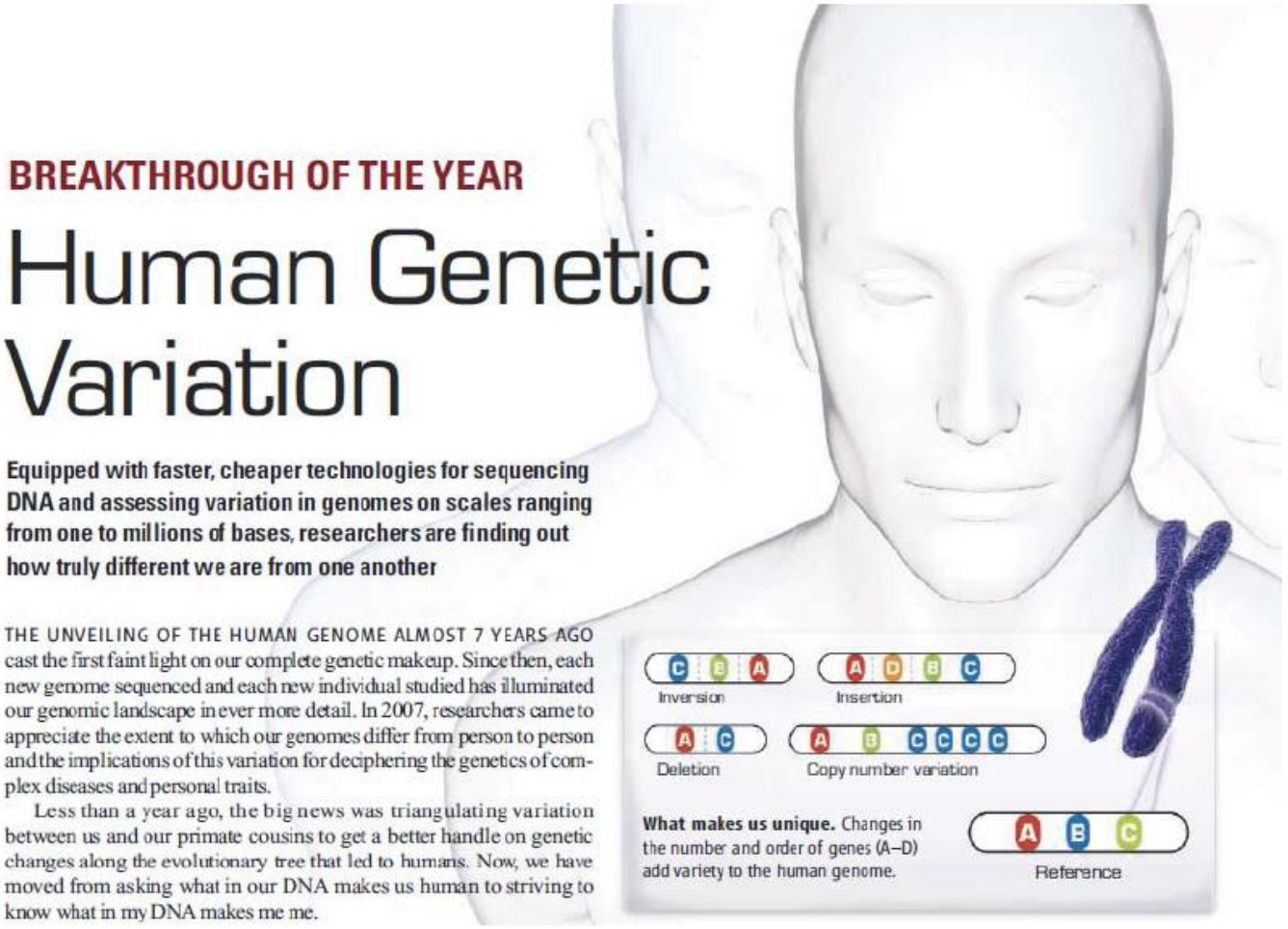
October 14, 2004: NHGRI Seeks Next Generation of Sequencing Technologies
The National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH), today announced it has awarded more than $38 million in grants to spur the development of innovative technologies designed to dramatically reduce the cost of DNA sequencing, a move aimed at broadening the applications of genomic information in medical research and health care.

October 20, 2004: International Human Genome Sequencing Consortium Describes Finished Human Genome Sequence
The International Human Genome Sequencing Consortium, led in the United States by the National Human Genome Research Institute (NHGRI) and the Department of Energy (DOE), today published its scientific description of the finished human genome sequence, reducing the estimated number of human protein-coding genes from 35,000 to only 20,000-25,000, a surprisingly low number for our species.

Historical overview



Historical overview



BREAKTHROUGH OF THE YEAR

Human Genetic Variation

Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another

THE UNVEILING OF THE HUMAN GENOME ALMOST 7 YEARS AGO cast the first faint light on our complete genetic makeup. Since then, each new genome sequenced and each new individual studied has illuminated our genomic landscape in ever more detail. In 2007, researchers came to appreciate the extent to which our genomes differ from person to person and the implications of this variation for deciphering the genetics of complex diseases and personal traits.

Less than a year ago, the big news was triangulating variation between us and our primate cousins to get a better handle on genetic changes along the evolutionary tree that led to humans. Now, we have moved from asking what in our DNA makes us human to striving to know what in my DNA makes me me.

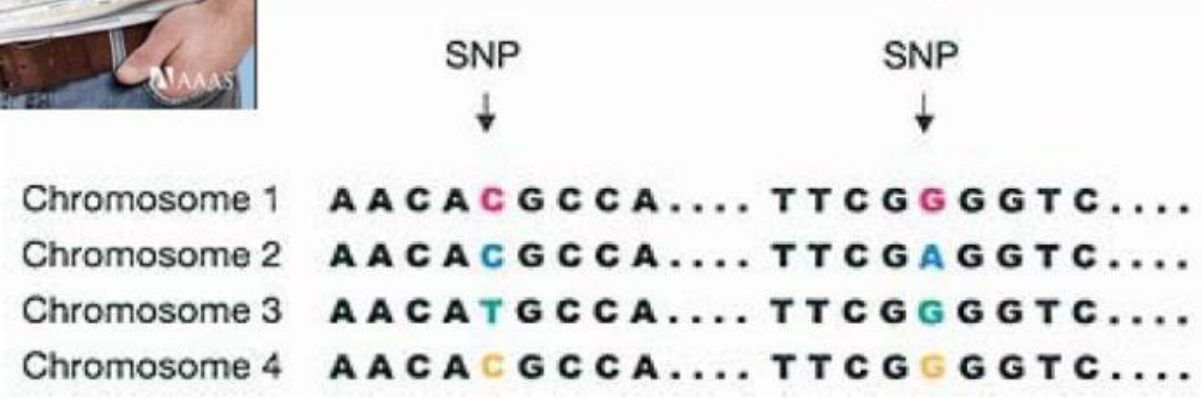
Pennisi 2007 Science 318:1842-3

2007 SCIENTIFIC BREAKTHROUGH OF THE YEAR

Science Magazine, December 21, 2007

"It's all about me!"

Single Nucleotide Polymorphisms (SNPs)



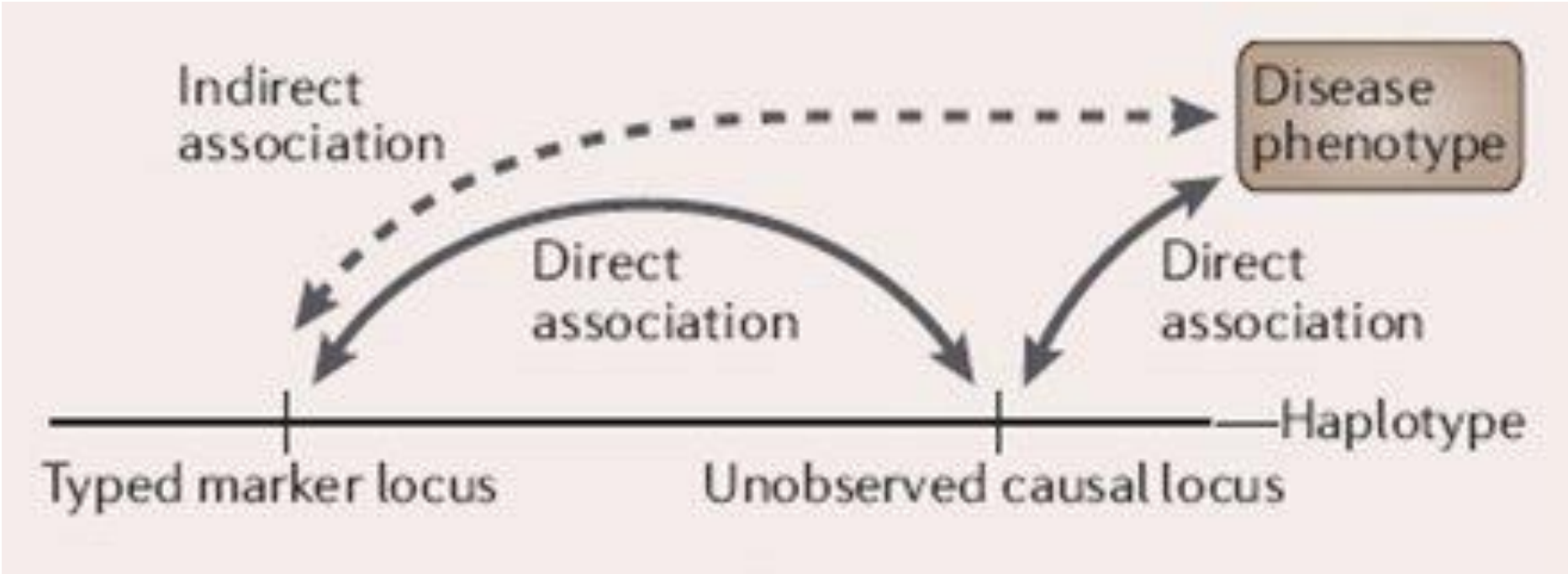
Historical overview: associating genetic variation to disease outcomes

BREAKTHROUGH OF THE YEAR: The Runners-Up

Science **314**, 1850a (2006);
DOI: 10.1126/science.314.5807.1850a

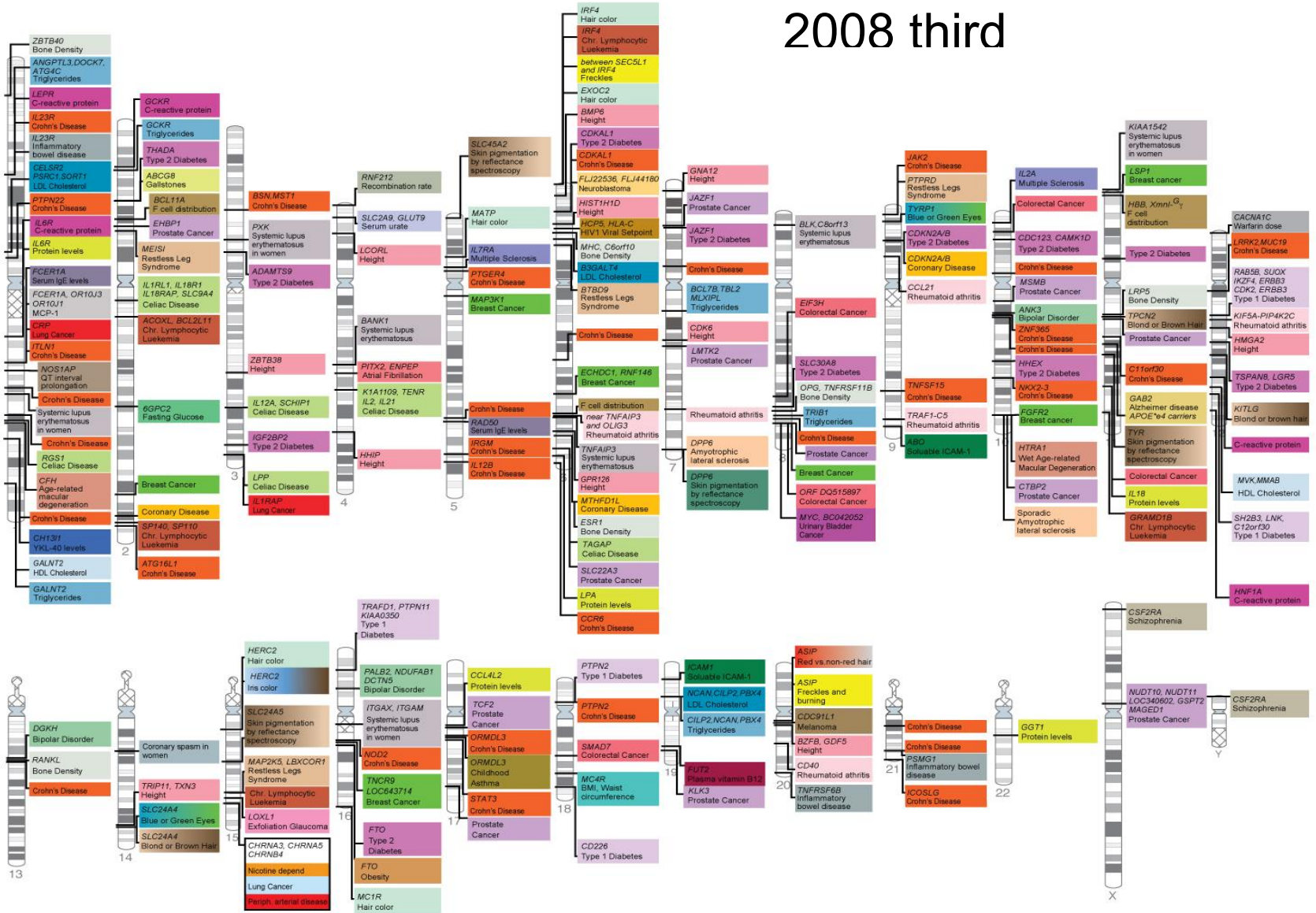
Areas to Watch in 2007

Whole-genome association studies. The trickle of studies comparing the genomes of healthy people to those of the sick is fast becoming a flood. Already, scientists have applied this strategy to macular degeneration, memory, and inflammatory bowel disease, and new projects on schizophrenia, psoriasis, diabetes, and more are heating up. But will the wave of data and new gene possibilities offer real insight into how diseases germinate? And will the genetic associations hold up better than those found the old-fashioned way?

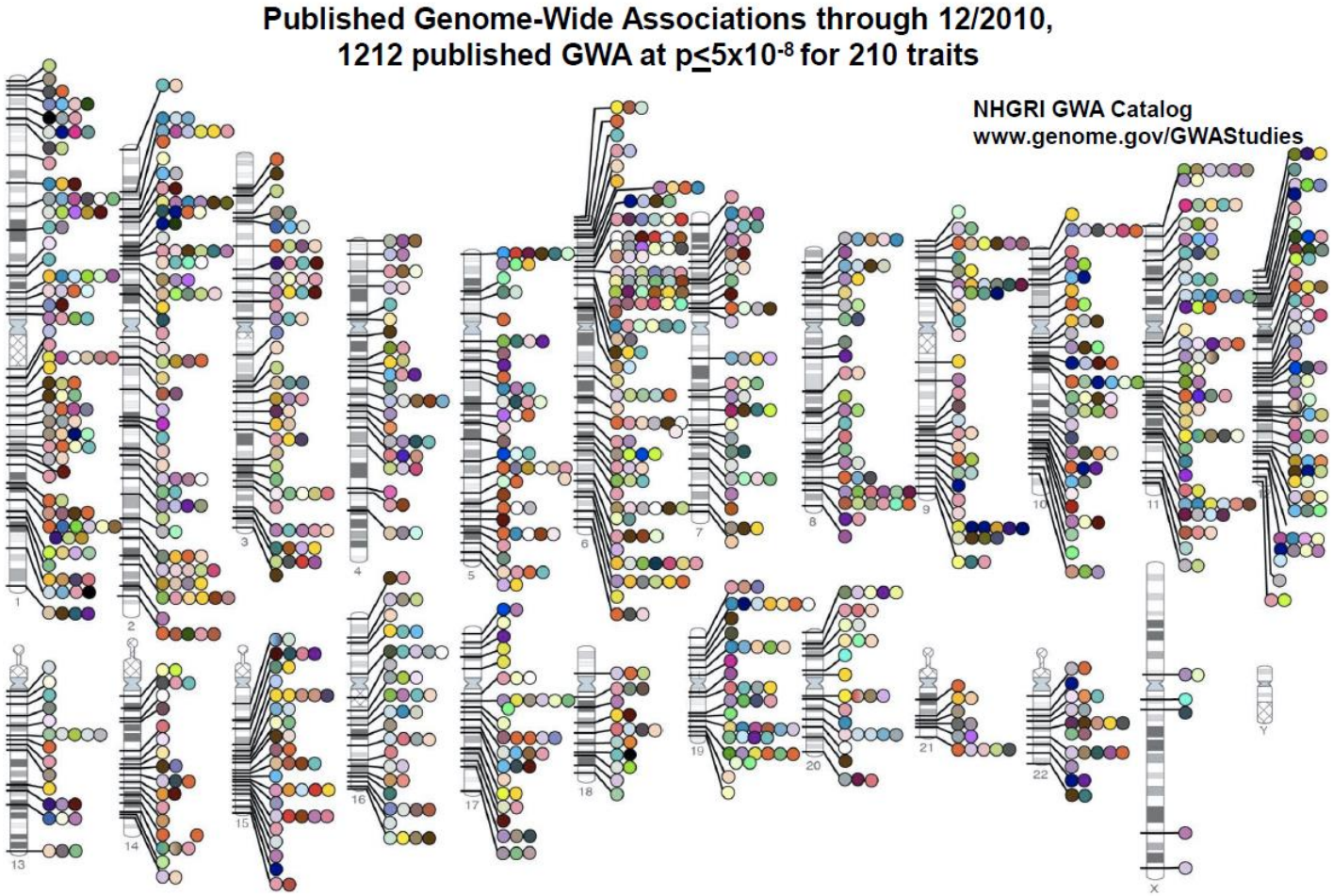


(Balding 2006)

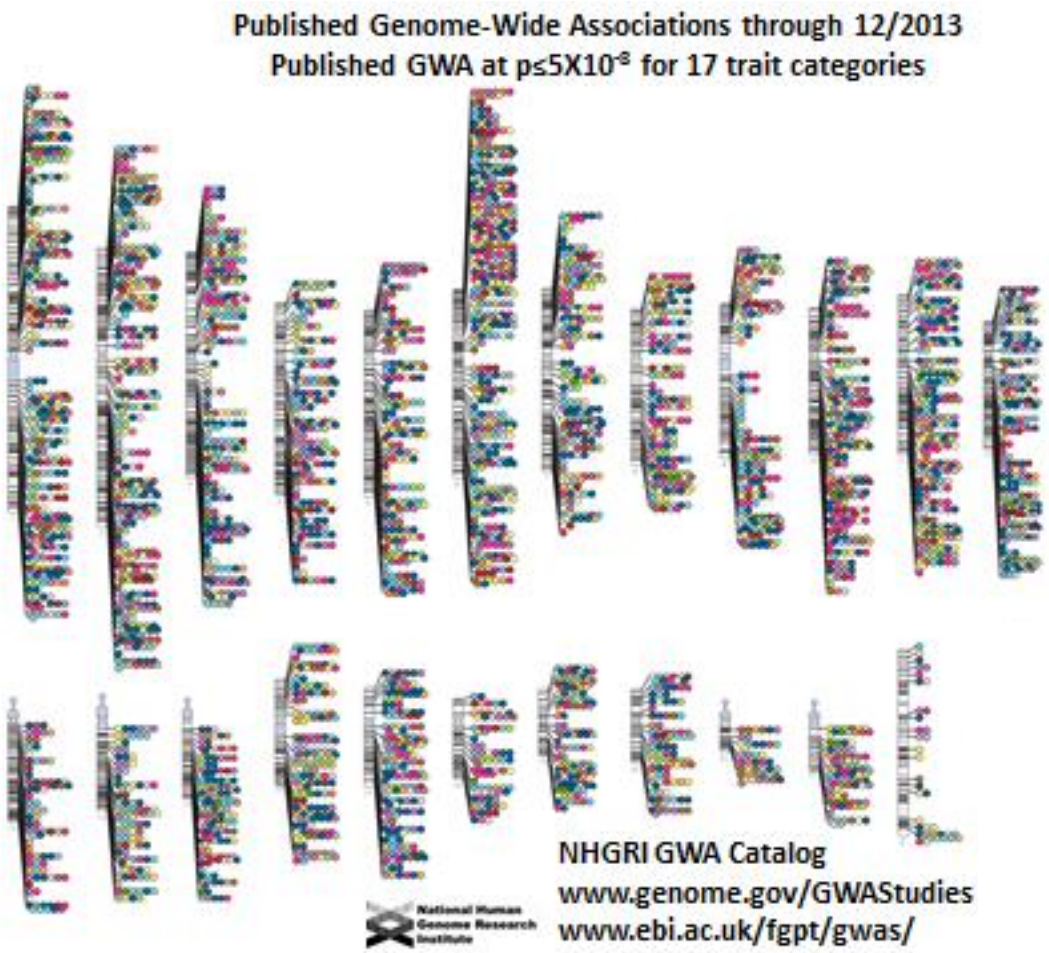
Historical overview: GWAs as a tool to “map” diseases



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



Historical overview: trait categories

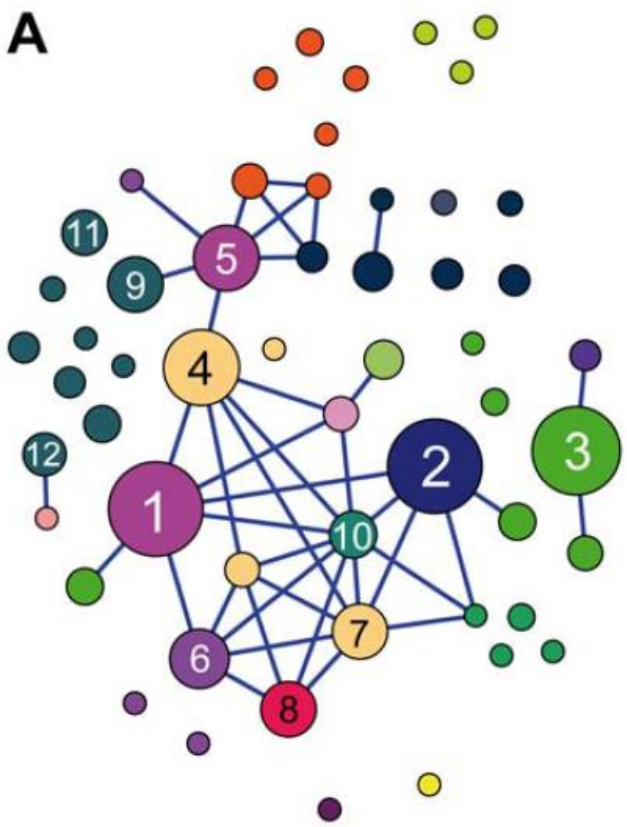


- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurment
- Other measurement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

Historical overview: trait categories and nr of SNPs

Date: 3/3/2020

Historical overview: inter-relationships (networks)

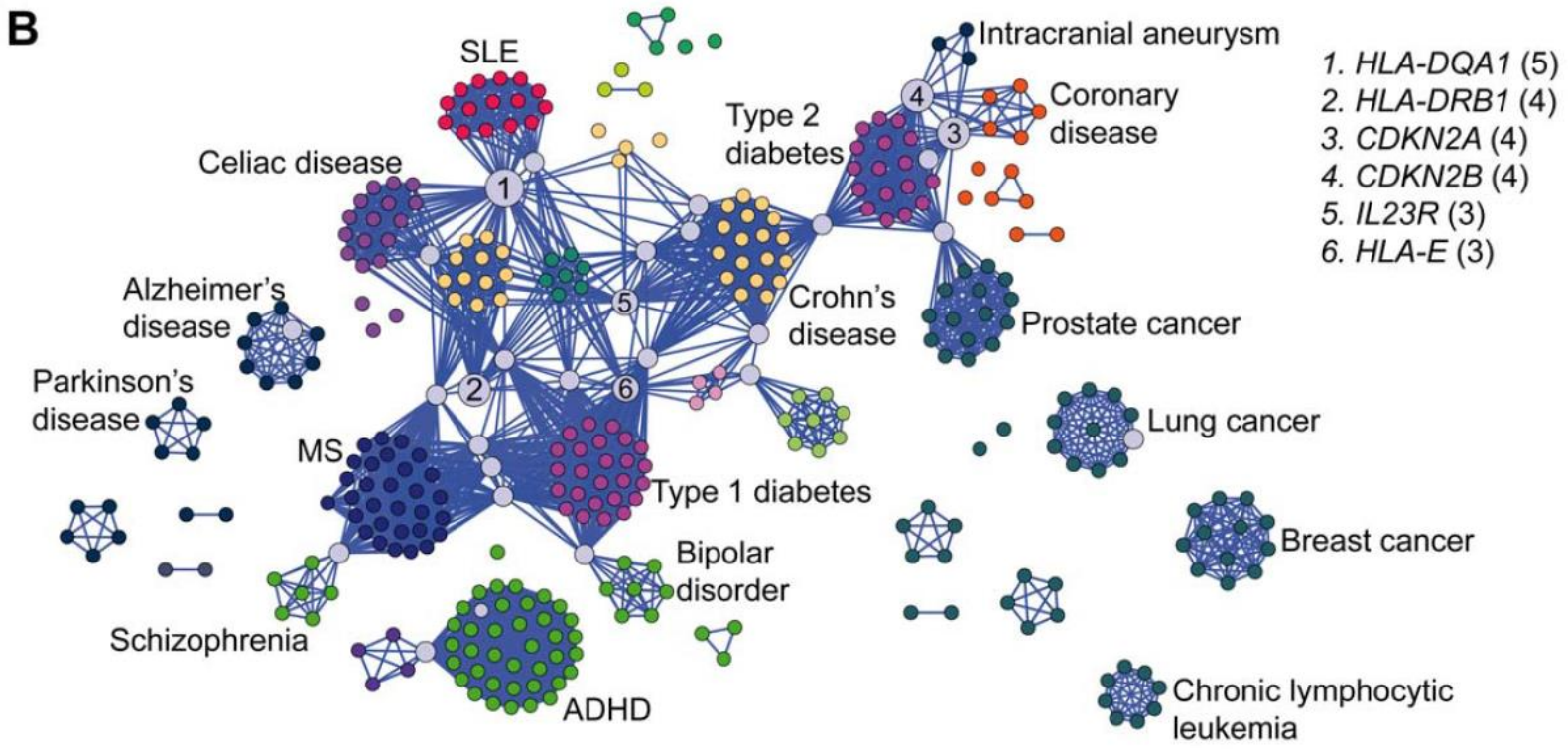


1. Type 1 diabetes (36)
2. Multiple sclerosis (36)
3. ADHD and conduct disorder (33)
4. Crohn's disease (27)
5. Type 2 diabetes (22)
6. Celiac disease (19)
7. Ulcerative colitis(17)
8. Systemic lupus erythematosus (17)
9. Prostate cancer (17)
10. Rheumatoid arthritis (13)
11. Breast cancer (12)
12. Lung cancer (11)

- Cardiovascular diseases (Cv)
- Digestive system diseases
- Endocrine system diseases
- Eye diseases
- Immune system diseases (Is)
- Mental disorders
- Multiple diseases
- Musculoskeletal diseases (Ms)
- Ms, Sc, Is
- Neoplasms
- Nervous system diseases (Ns)
- Ns, Cv
- Ns, Is
- Ns, Ms
- Nutritional and metabolic diseases
- Nm, Es, Is
- Skin and connective tissue disease
- Sc, Is
- Urogenital diseases

(Barrenas et al 2009: complex disease network – nodes are diseases)

Historical overview: inter-relationships (networks)



(Barrenas et al 2009: complex disease GENE network – nodes are genes)

2 The rise of GWAs

(slide Doug Brutlag 2010)

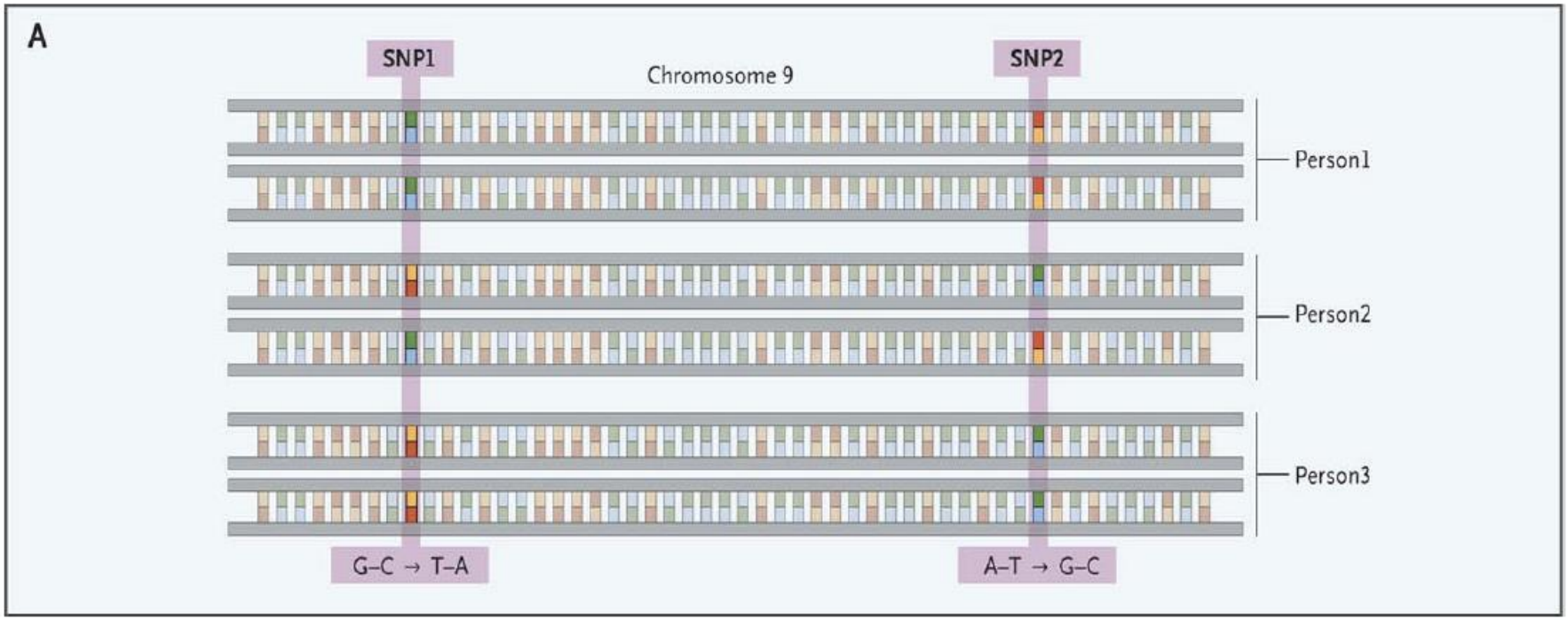
What are GWAs?

- A **genome-wide association study** is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular trait.

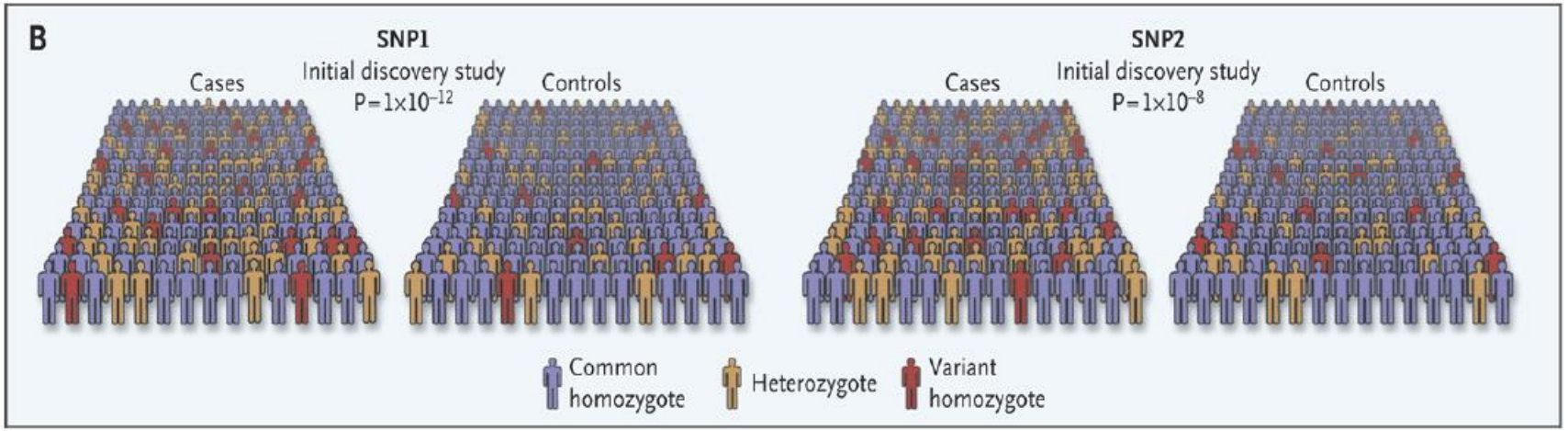
- **Recall**: a **trait** can be defined as a coded phenotype, a particular characteristic such as hair color, BMI, disease, gene expression intensity level, …

Genome-wide association studies: basic principles

The genome-wide association study is typically (but not solely!!!) based on a case‑control design in which single-nucleotide polymorphisms (SNPs) across the human genome are genotyped ... (Panel A: small fragment)



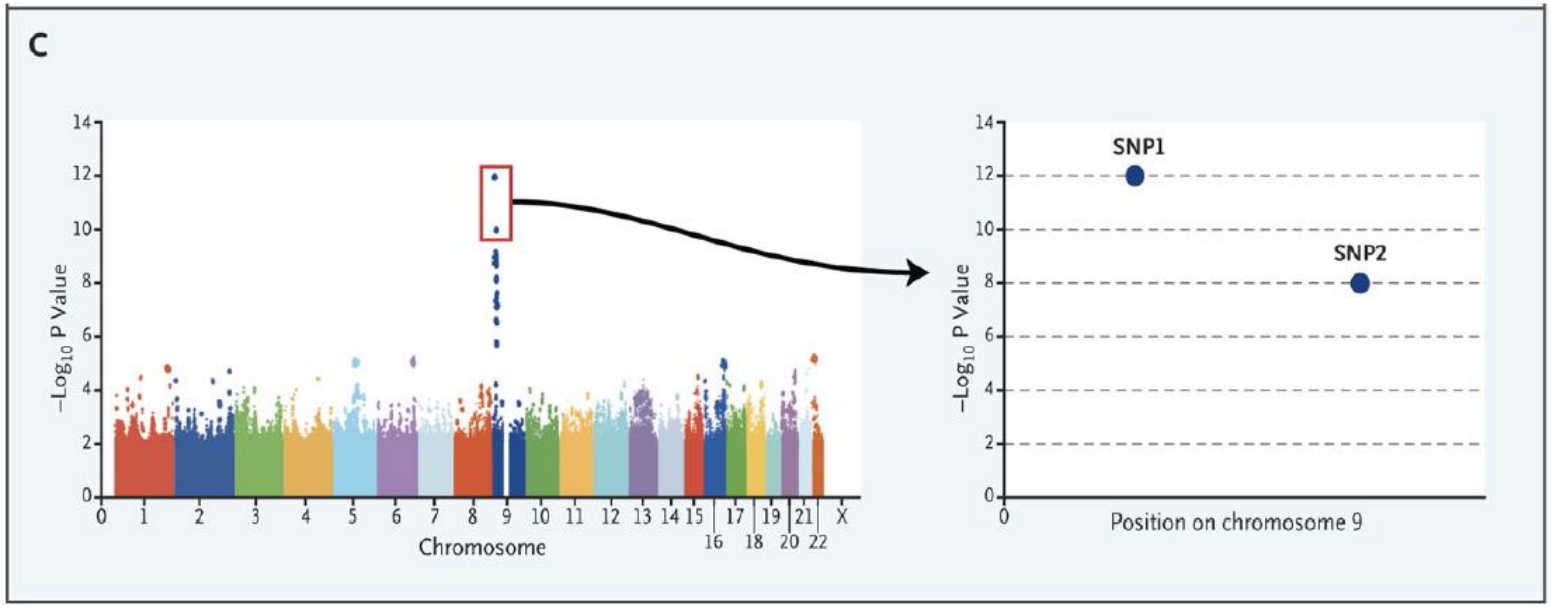
Genome-wide association studies: basic principles



- Panel B, the strength of association between each SNP and disease is calculated on the basis of the prevalence of each SNP in cases and controls. In this example, SNPs 1 and 2 on chromosome 9 are associated with disease, with P values of 10^{-12} and 10^{-8}, respectively

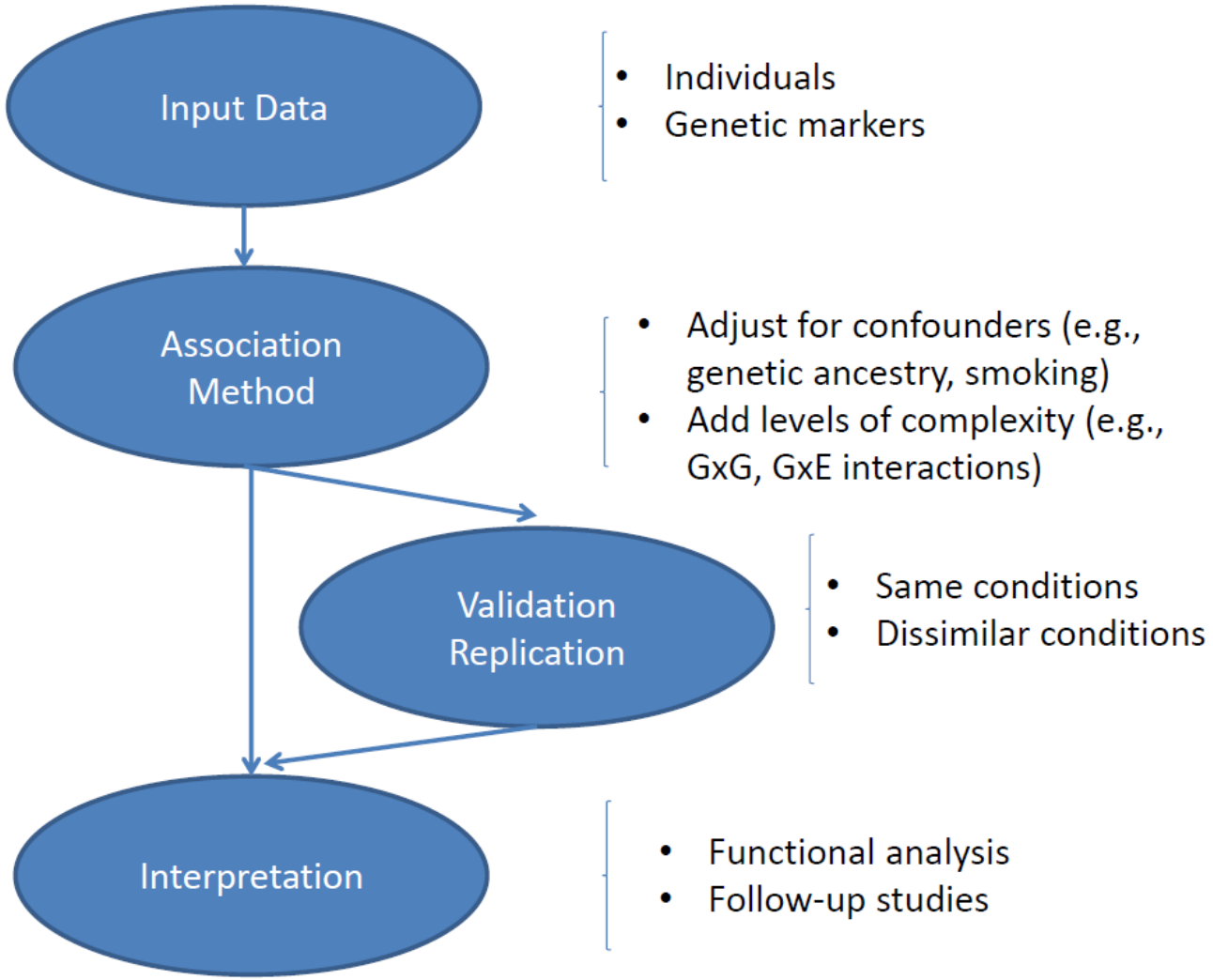
(Manolio 2010)

Genome-wide association studies: basic principles



- The plot in Panel C shows the P values for all genotyped SNPs that have survived a quality-control screen (each chromosome, a different color).
- The results implicate a locus on chromosome 9, marked by SNPs 1 and 2, which are adjacent to each other (graph at right), and other neighboring SNPs. (Manolio 2010)

Genome-wide association studies: key components



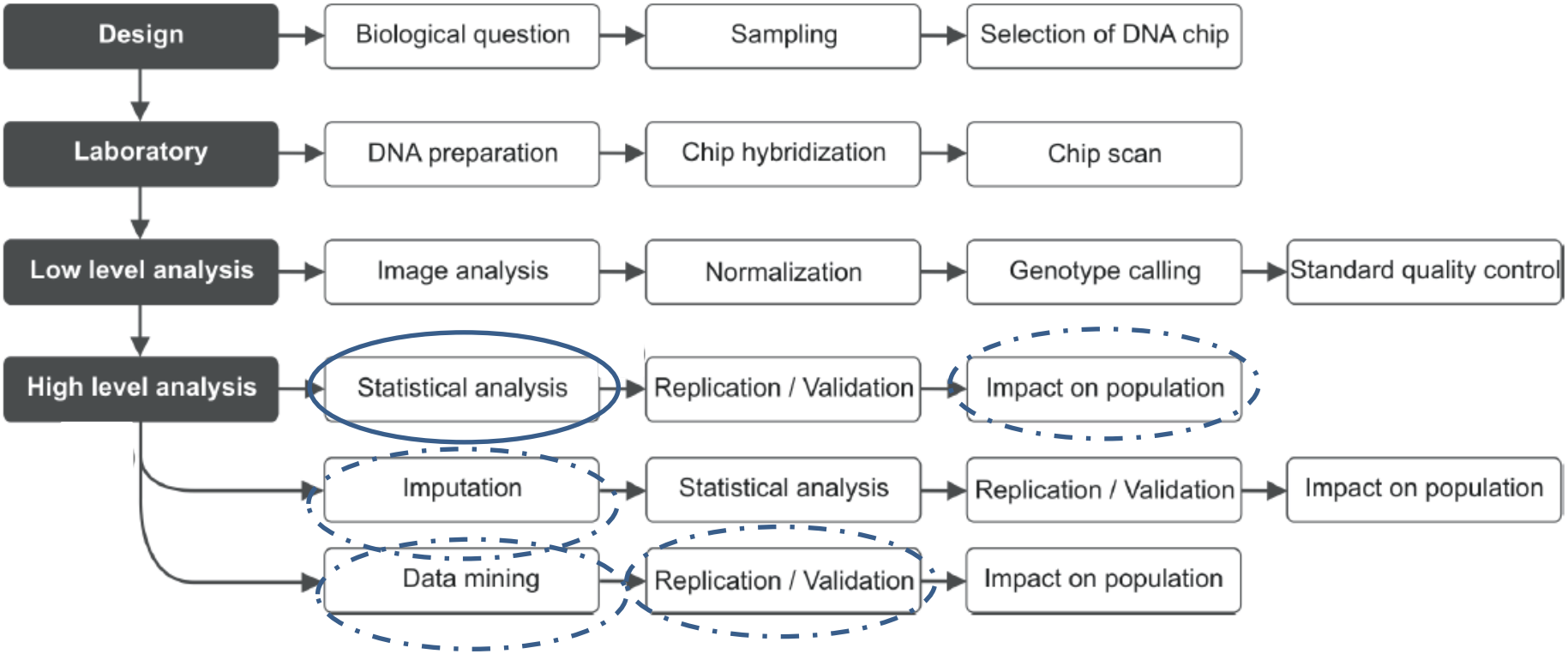
Genome-wide association studies: key components

- To carry out a GWAs, several tools are needed, which include those that deal with data generation and data handling:

  - Computerized data bases with reference human genome sequence
  - Map of human genetic variation
  - Technologies that can quickly and accurately analyze (whole genome) samples for genetic variations that contribute to disease

(http://www.genome.gov/pfv.cfm?pageID=20019523)

Detailed flow of a genome-wide association study



(Ziegler 2009)

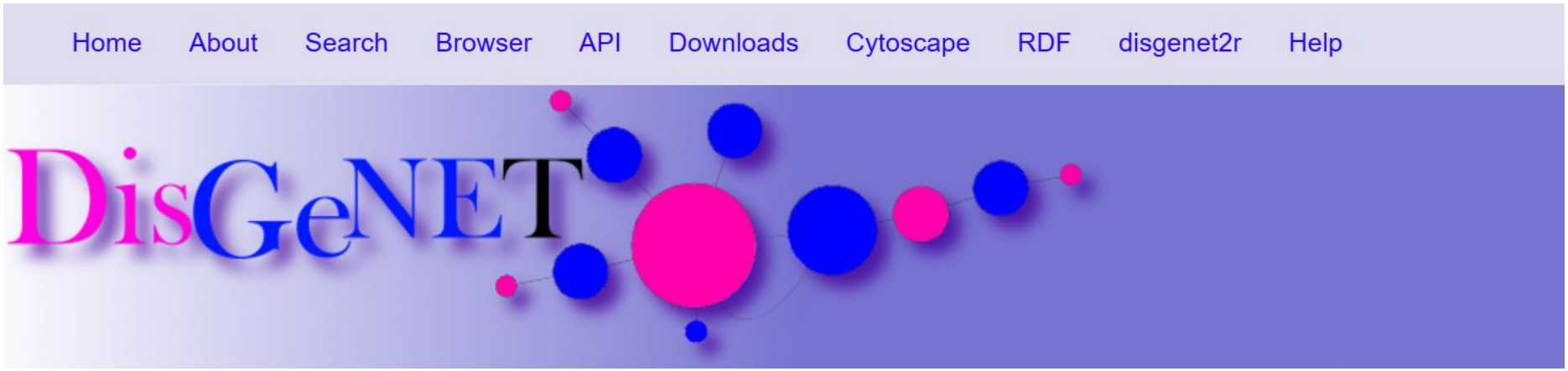
How to access GWAS results?

View the GWAs catalogue (http://www.genome.gov/gwastudies/)

2317 studies (6/10/2014)

(Entries 1-50 of 2317) Page 1 of 47 Next > Last >>

Date Added to Catalog (since 11/25/08)	First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Description	Replication Sample Description	Region	Reported Gene(s)	Mapped Gene(s)	Strongest SNP-Risk Allele	Context	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Platform [SNPs passing QC]	CNV
04/16/14	Chung CM March 03, 2014 *Diabetes Metab Res Rev* Common quantitative trait locus downstream of RETN gene identified by genome-wide association study is associated with risk of type 2 diabetes mellitus in Han Chinese: a Mendelian randomization effect.	Resistin levels	382 Han Chinese ancestry indiviudals	559 Han Chinese ancestry indiviudals	19p13.2	*RETN*	RETN - C19orf59	rs1423096-G		0.78	1×10^{-7}	.322 [0.25-0.40] ug/mL increase	Illumina [NR]	N
10/03/14	Zhang B January 21, 2014 *Int J Cancer* Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians.	Colorectal cancer	1,773 East Asian ancestry cases, 2,642 East Asian ancestry controls	6,902 East Asian ancestry cases, 7,862 East Asian ancestry controls	18q21.1	*SMAD7*	SMAD7	rs7229639-A	intron	0.145	3×10^{-11}	1.22 [1.15-1.29]	Affymetrix & Illumina [1,695,815] (imputed)	N
10/06/14	Xie T January 17, 2014 *Neurobiol Aging* A genome-wide association study combining pathway analysis for typical sporadic	Amyotrophic lateral sclerosis (sporadic)	250 Han Chinese ancestry cases, 250 Han Chinese ancestry controls	NA	View full set of 175 SNPs								Illumina [859,311] (pooled)	N
					NA	*RAB9P1*	NA	kgp22272527-?		NR	8×10^{-11}	NR		
					NA	*MYO18B*	NA	kgp8087771-?		0.2	2×10^{-10}	3.0327 [2.212039-4.157817]		
					12q24.33	*GPR133*	GPR133	rs11061269-?	intron	0.08	8×10^{-10}	3.7761 [2.49-5.74]		
					21q22.3	*TMPRSS2*	TMPRSS2 -	rs9977018-?		0.05	2×10^{-9}	NR		



DisGeNET is a discovery platform containing one of the largest publicly available collections of genes and variants associated to human diseases (Piñero *et al.*, 2019; Piñero *et al.*, 2016; Piñero *et al.*, 2015). DisGeNET integrates data from expert curated repositories, GWAS catalogues, animal models and the scientific literature. DisGeNET data are homogeneously annotated with controlled vocabularies and community-driven ontologies. Additionally, several original metrics are provided to assist the prioritization of genotype–phenotype relationships.

The current version of DisGeNET (v6.0) contains 628,685 gene-disease associations (GDAs), between 17,549 genes and 24,166 diseases, disorders, traits, and clinical or abnormal human phenotypes, and 210,498 variant-disease associations (VDAs), between 117,337 variants and 10,358 diseases, traits, and phenotypes.

DisGeNet !

The information in DisGeNET can be accessed in several ways:

- The web interface, through the Search and Browse functionalities
- The Resource Description Framework (DisGeNET-RDF) representation via the SPARQL endpoint, and the Faceted Browser
- The DisGeNET Cytoscape App
- Scripts in the most commonly used programming languages
- The disgenet2r package.
- The SQLite database
- Tab separated files.

DisGeNET is a versatile platform that can be used for different research purposes including the *investigation of the molecular underpinnings of human diseases and their comorbidities, the analysis of the properties of disease genes, the generation of hypothesis on drug therapeutic action and drug adverse effects, the validation of computationally predicted disease genes and the evaluation of text-mining methods performance*.

Rise of bioinformatics determines rise of GWAs (1)

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 10 2007, pages 1294–1296
doi:10.1093/bioinformatics/btm108

Genetics and population analysis

GenABEL: an R library for genome-wide association analysis

Yurii S. Aulchenko[1,*], Stephan Ripke[2], Aaron Isaacs[1] and Cornelia M. van Duijn[1]

[1]Department of Epidemiology and Biostatistics, Erasmus MC Rotterdam, Postbus 2040, 3000 CA Rotterdam, The Netherlands and [2]Statistical Genetics Group, Max-Planck-Institute of Psychiatry, Kraepelinstr. 10, D-80804 Munich, Germany

Received on December 3, 2006; revised on February 14, 2007; accepted on March 13, 2007

Advance Access publication March 23, 2007

Associate Editor: Martin Bishop

ABSTRACT

Here we describe an R library for genome-wide association (GWA) analysis. It implements effective storage and handling of GWA data, fast procedures for genetic data quality control, testing of association of single nucleotide polymorphisms with binary or quantitative traits, visualization of results and also provides easy interfaces to standard statistical and graphical procedures implemented in base R and special R libraries for genetic analysis. We evaluated GenABEL using one simulated and two real data sets. We conclude that GenABEL enables the analysis of GWA data on desktop computers.

Availability: http://cran.r-project.org

Contact: i.aoultchenko@erasmusmc.nl

With these objectives in mind, we developed the GenABEL software, implemented as an R library. R is a free, open source language and environment for statistical analysis (http://www.r-project.org/). Building upon existing statistical analysis facilities allowed for rapid development of the package.

2 IMPLEMENTATION

2.1 Objective (1)

GWA data storage using standard R data types is ineffective. A SNP genotype for a single person may take four values (AA, AB, BB and missing). Two bits, therefore, are required to store these data. However, the standard R data types occupy 32 bits, leading to an overhead of 1500%, compared to the theoretical optimum. Use of the raw R data format, occupying

Downloaded from http://bioinformatics.oxfo

Rise of bioinformatics determines rise of GWAs (2)

BIOINFORMATICS

Vol. 26 ISMB 2010, pages i208–i216
doi:10.1093/bioinformatics/btq191

Multi-population GWA mapping via multi-task regularized regression

Kriti Puniyani, Seyoung Kim and Eric P. Xing*
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

Motivation: Population heterogeneity through admixing of different founder populations can produce spurious associations in genome-wide association studies that are linked to the population structure rather than the phenotype. Since samples from the same population generally co-evolve, different populations may or may not share the same genetic underpinnings for the seemingly common phenotype. Our goal is to develop a unified framework for detecting causal genetic markers through a joint association analysis of multiple populations.

Results: Based on a multi-task regression principle, we present a multi-population group lasso algorithm using L_1/L_2-regularized regression for joint association analysis of multiple populations that are stratified either via population survey or computational estimation. Our algorithm combines information from genetic markers across populations, to identify causal markers. It also implicitly accounts for correlations between the genetic markers, thus enabling better control over false positive rates. Joint analysis across populations enables the detection of weak associations common to all populations with greater power than in a separate analysis of each population. At the same time, the regression-based framework allows causal alleles that are unique to a subset of the populations to be correctly identified. We demonstrate the effectiveness of our method on HapMap-simulated and lactase persistence datasets, where we significantly outperform state of the art methods, with greater power for detecting weak associations and reduced spurious associations.

Availability: Software will be available at http://www.sailing.cs.cmu.edu/

the geographical distribution of the individuals. For example, it has been shown that such heterogeneity is present in the HapMap data (The International HapMap Consortium, 2005) across European, Asian and African populations; and heterogeneity at a finer scale within European ancestry has been found in many genomic regions in the UK samples of Wellcome trust case control consortium (WTCCC) dataset (Wellcome Trust Case Control Consortium, 2007). Although the standard assumption in existing approaches for association mapping is that the effects of causal mutations are likely to be common across multiple populations, the individuals in the same population or geographical region tend to co-evolve, and are likely to possess a population-specific causal allele for the same phenotype. For example, Tishkoff *et al.* (2006) reported that the lactase-persistence phenotype is caused by different mutations in Africans and Europeans. In addition, the same genetic variation has been observed to be correlated with gene-expression levels with different association strengths across different HapMap populations. Our goal is to be able to leverage information across multiple populations, to find causal markers in a multi-population association study.

1.1 Highlights of this article

We propose a novel multi-task-regression-based technique that performs a joint GWA mapping on individuals from multiple populations, rather than separate analysis of each population, to detect associated genome variations. The joint inference is achieved by using a multi-population group lasso (MPGL), with an L_1/L_2

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest

Rise of bioinformatics determines rise of GWAs (3)

BIOINFORMATICS APPLICATIONS NOTE

Vol. 24 no. 1 2008, pages 140–142
doi:10.1093/bioinformatics/btm549

Genetics and population analysis

GWAsimulator: a rapid whole-genome simulation program

Chun Li[1,*] and Mingyao Li[2]

[1]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37232 and [2]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

Received on July 20, 2007; revised on October 10, 2007; accepted on October 29, 2007

Advance Access publication November 15, 2007

Associate Editor: Martin Bishop

ABSTRACT

Summary: GWAsimulator implements a rapid moving-window algorithm to simulate genotype data for case-control or population samples from genomic SNP chips. For case-control data, the program generates cases and controls according to a user-specified multi-locus disease model, and can simulate specific regions if desired. The program uses phased genotype data as input and has the flexibility of simulating genotypes for different populations and different genomic SNP chips. When the HapMap phased data are used, the simulated data have similar local LD patterns as the HapMap data. As genome-wide association (GWA) studies become increasingly popular and new GWA data analysis methods are being developed, we anticipate that GWAsimulator will be an important tool for evaluating performance of new GWA analysis methods.

Availability: The C++ source code, executables for Linux, Windows and MacOS, manual, example data sets and analysis program are available at http://biostat.mc.vanderbilt.edu/GWAsimulator

Contact: chun.li@vanderbilt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

2 METHODS

The program can generate unrelated case-control (sampled retrospectively conditional on affection status) or population (sampled randomly) data of genome-wide SNP genotypes with patterns of LD similar to the input data.

2.1 Phased input data and control file

The program requires phased data as input. If the HapMap data are used, the number of phased autosomes and X chromosomes are 120 and 90 for both CEU and YRI, 90 and 68 for CHB, and 90 and 67 for JPT. Additional parameters needed by the program should be provided in a control file, including disease model (see Section 2.2), window size (see Section 2.3), whether to output the simulated data (see Section 2.4), and the number of subjects to be simulated.

2.2 Determination of disease model

For simulations of case-control data, a disease model is needed. The program allows the user to specify disease model parameters, including disease prevalence, the number of disease loci, and for each disease locus, its location, risk allele and genotypic relative risk. If the user wants to simulate specific regions, the start and end positions need

Downloaded from http://bioinformatics.oxfordjournals.org

Rise of bioinformatics determines rise of GWAs (4)

BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 5 2009, pages 662–663
doi:10.1093/bioinformatics/btp017

Genome analysis

AssociationViewer: a scalable and integrated software tool for visualization of large-scale variation data in genomic context

Olivier Martin[1,†], Armand Valsesia[1,2,†], Amalio Telenti[3], Ioannis Xenarios[1] and Brian J. Stevenson[1,2,*]

[1]Swiss Institute of Bioinformatics, [2]Ludwig Institute for Cancer Research, 1015 Lausanne and [3]Institute of Microbiology, University Hospital, University of Lausanne, 1011 Lausanne, Switzerland

Received on September 16, 2008; revised on December 16, 2008; accepted on January 5, 2009

Advance Access publication January 25, 2009

Associate Editor: John Quackenbush

ABSTRACT

Summary: We present a tool designed for visualization of large-scale genetic and genomic data exemplified by results from genome-wide association studies. This software provides an integrated framework to facilitate the interpretation of SNP association studies in genomic context. Gene annotations can be retrieved from Ensembl, linkage disequilibrium data downloaded from HapMap and custom data imported in BED or WIG format. AssociationViewer integrates functionalities that enable the aggregation or intersection of data tracks. It implements an efficient cache system and allows the display of several, very large-scale genomic datasets.

Availability: The Java code for AssociationViewer is distributed under the GNU General Public Licence and has been tested on Microsoft Windows XP, MacOSX and GNU/Linux operating systems. It is available from the SourceForge repository. This also includes Java webstart, documentation and example datafiles.

Contact: brian.stevenson@licr.org

Supplementary information: Supplementary data are available at http://sourceforge.net/projects/associationview/ online.

represented in BED or WIG format and implements aggregation (union) or intersection of data tracks.

2 PROGRAM OVERVIEW

2.1 Cache and memory management

With increasing data volumes, efficient resource management is essential. One approach is to store the data in a cache with fast indexing mechanisms to retrieve the data, and to keep in memory only the information that is visualized. We implemented such a system in AssociationViewer. For comparison, loading a single dataset with 500 K SNPs in WGAViewer needs about 224 MB of RAM, whereas loading 10 different datasets (a total of 10 M data points) and displaying all genes on chromosome 1 needs only 50 MB in AssociationViewer.

2.2 Data import and export

A typical GWA dataset consists of a list of SNPs with *P*-values derived from an association analysis. In AssociationViewer, such

Bioconductor

Method

Open Access

Bioconductor: open software development for computational biology and bioinformatics

Robert C Gentleman[1], Vincent J Carey[2], Douglas M Bates[3], Ben Bolstad[4], Marcel Dettling[5], Sandrine Dudoit[4], Byron Ellis[6], Laurent Gautier[7], Yongchao Ge[8], Jeff Gentry[1], Kurt Hornik[9], Torsten Hothorn[10], Wolfgang Huber[11], Stefano Iacus[12], Rafael Irizarry[13], Friedrich Leisch[9], Cheng Li[1], Martin Maechler[5], Anthony J Rossini[14], Gunther Sawitzki[15], Colin Smith[16], Gordon Smyth[17], Luke Tierney[18], Jean YH Yang[19] and Jianhua Zhang[1]

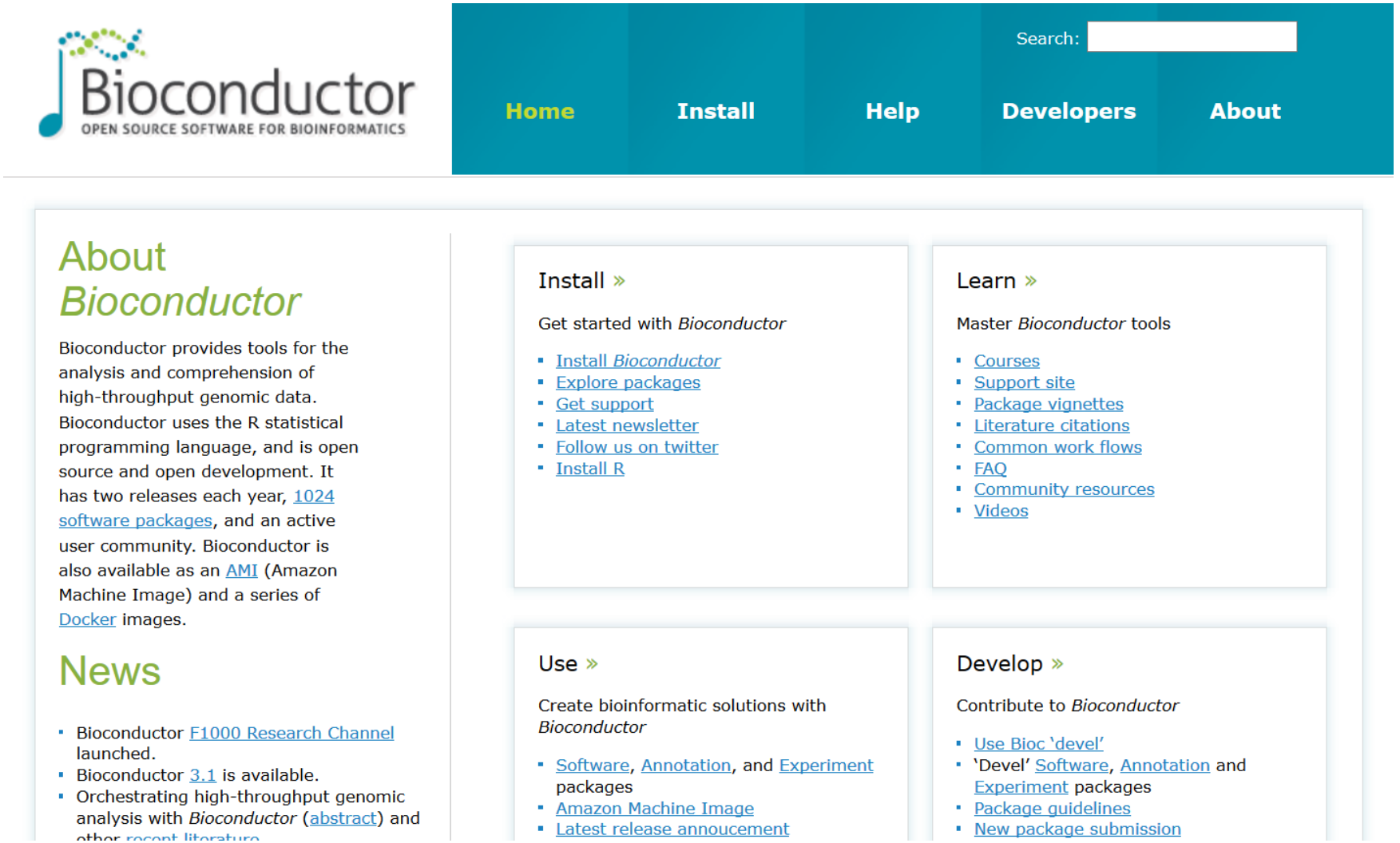
Published: 15 September 2004

Genome ***Biology*** 2004, **5**:R80

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/10/R80

Received: 19 April 2004
Revised: 1 July 2004
Accepted: 3 August 2004

© 2004 Gentleman *et al.*; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



(http://www.bioconductor.org/)

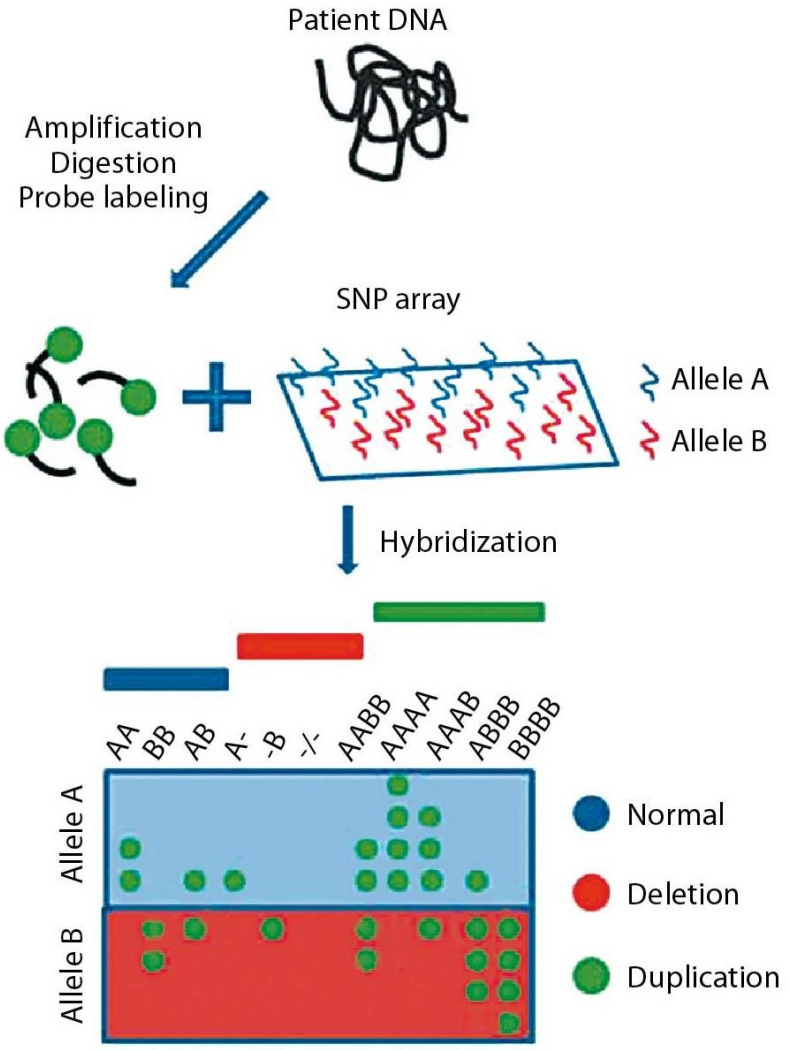
3 Study Design

Components of a study design for GWA studies

- The design of a genetic association study may refer to
  - study scale:
    - Genetic (e.g., hypothesis-drive, panel of candidate genes)
    - Genomic (e.g., hypothesis-free, genome-wide)
  - marker design:
    - Which markers are most informative in GWAs? Common variants-SNPs and/or Rare Variants (MAF<1%)
    - Which platform is the most promising? Least error-prone? Marker-distribution over the genome?
  - subject design

3.a Marker Level

- Costs may play a role, but a balance is needed between costs and chip/sequencing platform performance
- Coverage also plays a role (e.g., exomes only or a uniform spread).
- When choosing **Next Generation Sequencing platforms, also rare variants** can be included in the analysis, in contrast to the older **SNP-arrays** (see right panel).



From common variants towards including rare variants

- Hypothesis 1 for GWAs: Common Disease – Common Variant (CDCV):

  - This hypothesis argues that **genetic variations with appreciable frequency** in the population at large, but **relatively low penetrance** (i.e. the probability that a carrier of the relevant variants will express the disease), are the major contributors to genetic susceptibility to common diseases (Lander, 1996; Chakravarti, 1999; Weiss & Clark, 2002; Becker, 2004).
  - The hypothesis speculates that the gene variation underlying susceptibility to common heritable diseases existed within the **founding population of contemporary humans** → explains the success of GWAs?

From common variants towards including rare variants

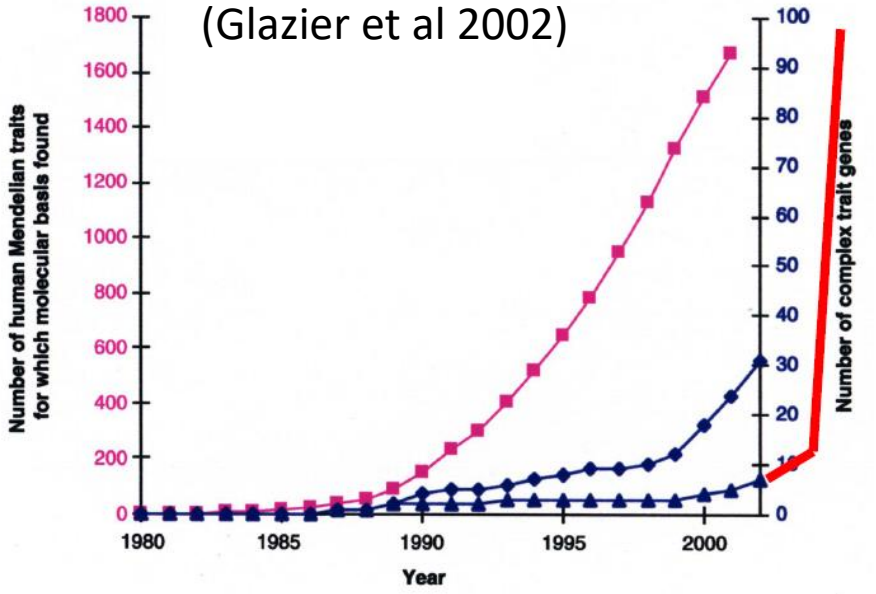
- Hypothesis 2 for GWAs: Common Disease – Rare Variant (CDRV):

  - This hypothesis argues that **rare DNA sequence variations**, each with **relatively high** (moderate to high) **penetrance**, are the major contributors to genetic susceptibility to common diseases.
  - Some argumentations behind this hypothesis include that by reaching an appreciable frequency for common variations, these variations are not as likely to have been subjected to negative selection. Rare variations, on the other hand, may be **rare because they are being selected against due to their deleterious nature**.

There is room for both hypothesis in current research !
(Schork et al. 2009)

Identified # of traits for which a molecular basis exists: importance of SNPs

(Glazier et al 2002)



PINK : Human Mendelian traits

BLUE middle line : All complex traits

BLUE bottom line + red extension: Human complex traits

Complex disease (definition):

The term complex trait/disease refers to any phenotype that

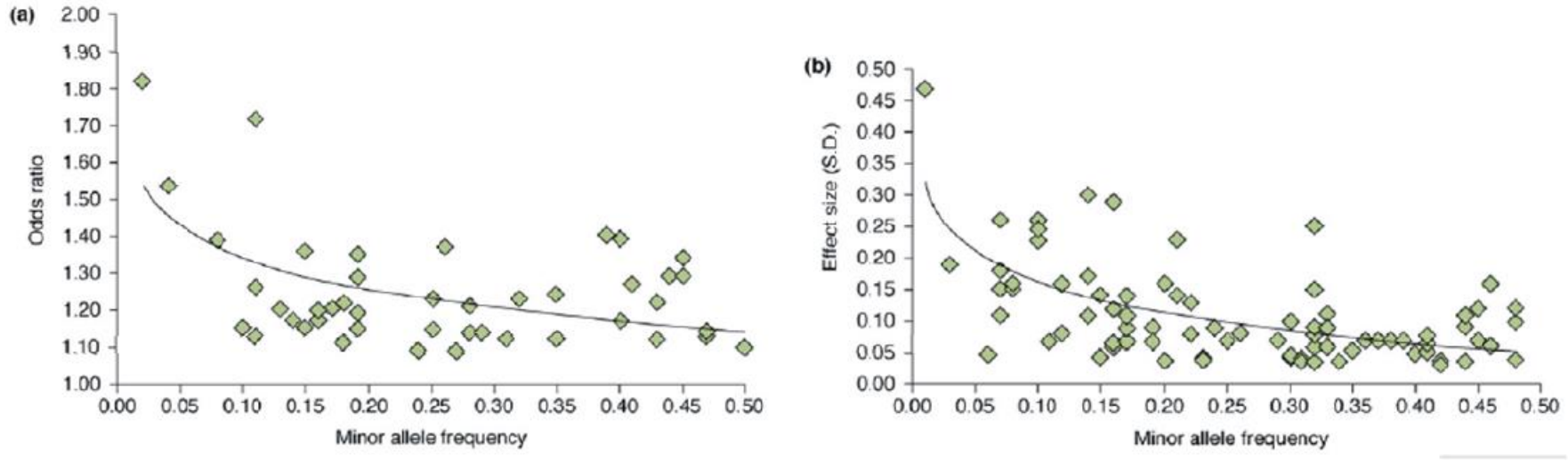
does NOT exhibit classic Mendelian inheritance attributable to a single gene;

although they may exhibit familial tendencies (familial clustering, concordance among relatives).

Distribution of SNP "effects"

Dichotomous Traits

Quantitative Traits



Arking & Chakravarti 2009 Trends Genet

Food for thought:

- The higher the MAF, the lower the effect size
- Rare variants analysis is in its infancy in 2009 ….

3.b Subject Level

Aim	Selection scheme
Increased effect size	Extreme sampling: Severely affected cases vs. extremely normal controls
Genes causing early onset	Affected, early onset vs. normal, elderly
Genes with large / moderate effect size	Cases with positive family history vs. controls with negative family history
Specific GxE interaction	Affected vs. normal subjects with heavy environmental exposure
Longevity genes	Elderly survivors serve as cases vs. young serve as controls
Control for covariates with strong effect	Affected with favorable covariates vs. normal with unfavorable covariate

Morton & Collins 1998 Proc Natl Acad Sci USA 95:11389

Popular design 1: cases and controls

Avoiding bias – checking assumptions:

1. Cases and controls drawn from same population
2. Cases representative for all cases in the population
3. All data collected similarly in cases and controls

Advantages:

1. Simple
2. Cheap
3. Large number of cases and controls available
4. Optimal for studying rare diseases

Disadvantages:

1. Population stratification
2. Prone to batch effects and other biases
3. Case definition / severity
4. Overestimation of risk for common diseases

Popular design 2: family-based

Avoiding bias – checking assumptions:

1. Families representative for population of interest
2. Same genetic background in both parents

Advantages:

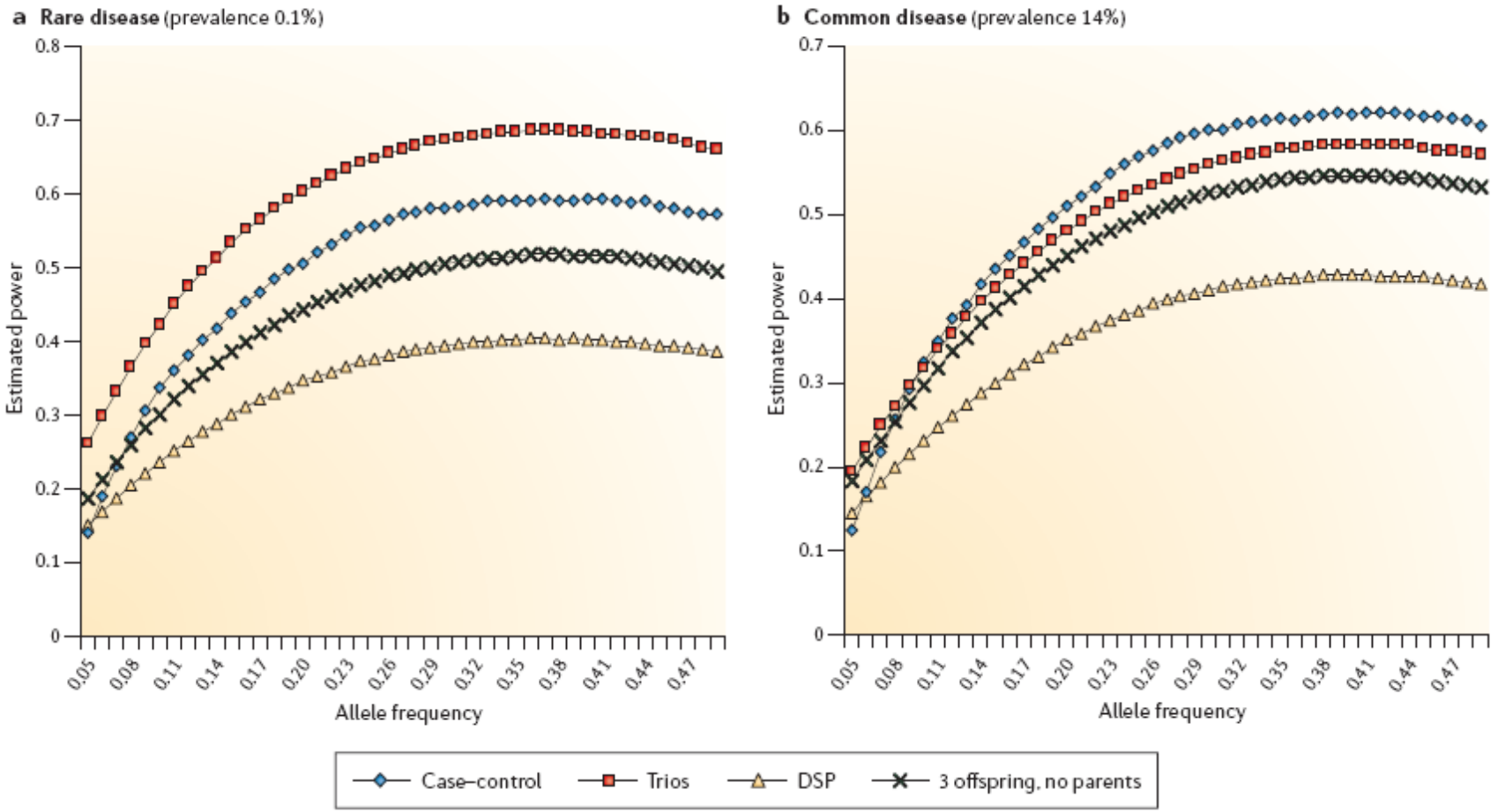
1. Controls immune to population stratification (no association without linkage, no "spurious" (false positive) association)
2. Checks for Mendelian inheritance possible (fewer genotyping errors)
3. Parental phenotyping not required (late onset diseases)
4. Simple logistics for diseases in children
5. Allows investigating imprinting ("bad allele" from father or mother?)

Disadvantages

1. Cost inefficient
2. Sensitive to genotyping errors
3. Lower power when compared with case-control studies

Some more power considerations

- Rare versus common diseases (Lange and Laird 2006)



4 Pre-analysis steps

4.a Quality control

Standard file format for GWA studies

Standard data format: tped = transposed ped format file

FamID	PID	FID	MID	SEX	AFF	$SNP1_1$	$SNP1_2$	$SNP2_1$	$SNP2_2$
1	1	0	0	1	1	A	A	G	T
2	1	0	0	1	1	A	C	T	G
3	1	0	0	1	1	C	C	G	G
4	1	0	0	1	2	A	C	T	T
5	1	0	0	1	2	C	C	G	T
6	1	0	0	1	2	C	C	T	T

ped file

Chr	SNP name	Genetic distance	Chromosomal position
1	SNP1	0	123456
1	SNP2	0	123654

map file

Standard file format for GWA studies (continued)

Chr	SNP	Gen. dist.	Pos	PID 1		PID 2		PID 3		PID 4		PID 5		PID 6	
1	SNP1	0	123456	A	A	A	C	C	C	A	C	C	C	C	C
1	SNP2	0	123654	G	T	G	T	G	G	T	T	G	T	T	T

tped file

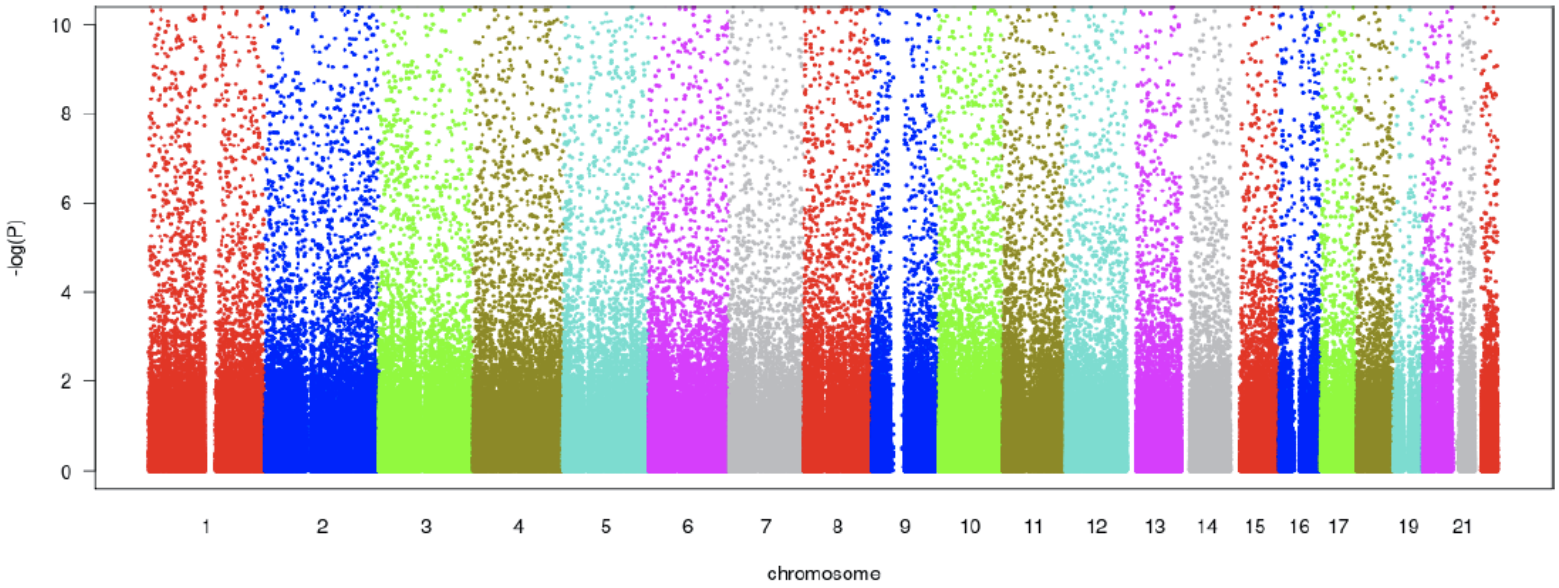
tfam file: First 6 columns of standard ped file

FamID	PID	FID	MID	SEX	AFF
1	1	0	0	1	1
2	1	0	0	1	1
3	1	0	0	1	1
4	1	0	0	1	2
5	1	0	0	1	2
6	1	0	0	1	2

tfam file

Why is quality control (QC) important?

BEFORE QC → true signals are lost in false positive signals

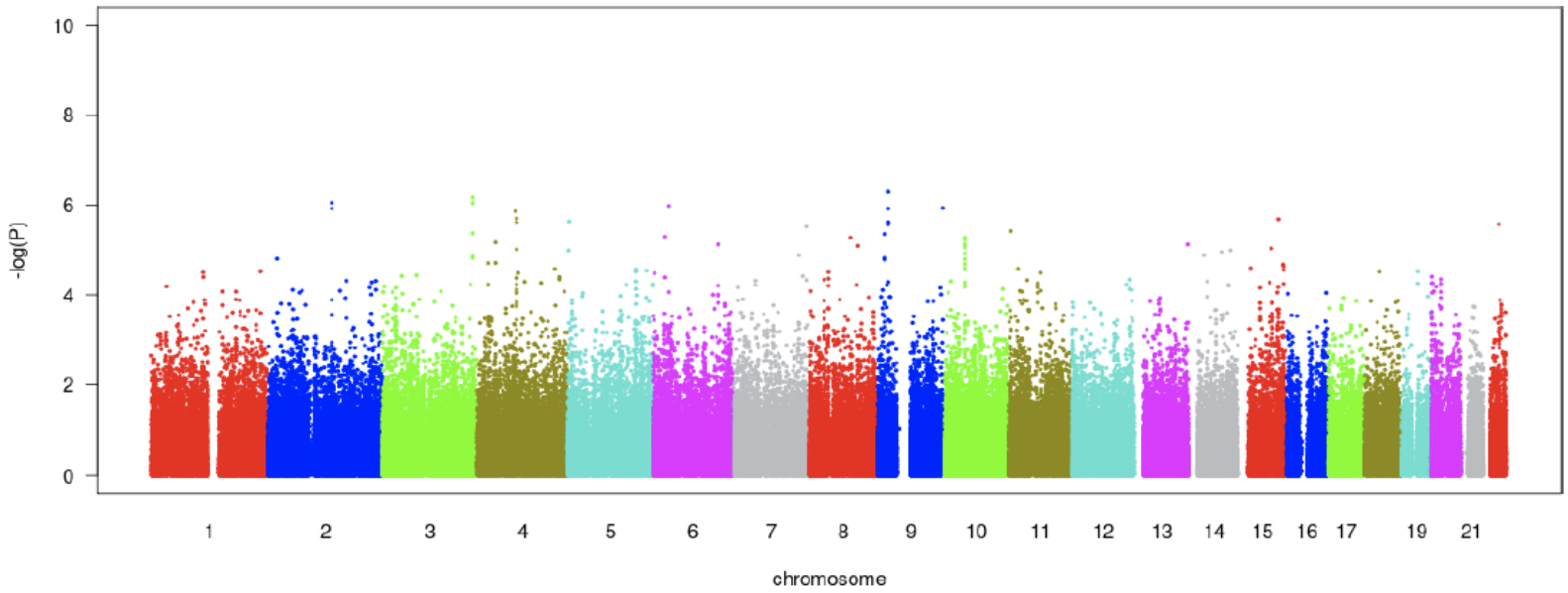


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(Ziegler and Van Steen 2010)

Why is quality control important?

AFTER QC → skyline of Manhattan (→ name of plot: Manhattan plot):



Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

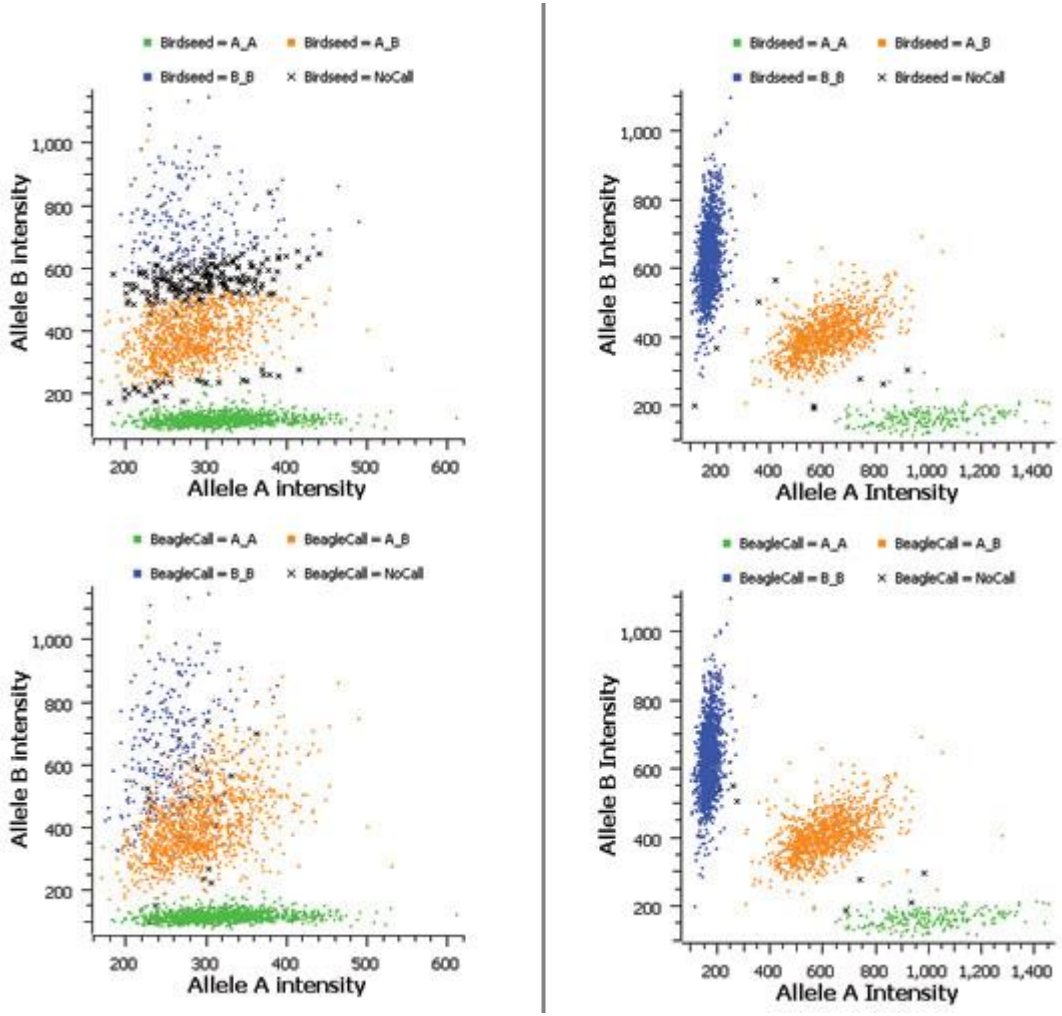
SNPs passing standard quality control: 270,701

(Ziegler and Van Steen 2010)

What is the standard quality control?

- Quality control can be performed on different levels:
  - Subject or sample level
  - Marker level (in this course: SNP level)
  - X-chromosomal SNP level (in this course not considered)
- Consensus on how to best QC data has led to the so-called “Travemünde criteria” (obtained in the town Travemünde) – see later

Marker level QC thresholds may be genotype calling algorithm dependent



Allele signal intensity genotype calling cluster plots for two different SNPs from the same study population.

Upper panels: Birdseed genotypes

Lower panels: BEAGLECALL genotypes.

The plots on the left show a SNP with poor resolution of A_B and B_B genotype clusters and the increased clarity of genotype calls that comes from using BEAGLECALL

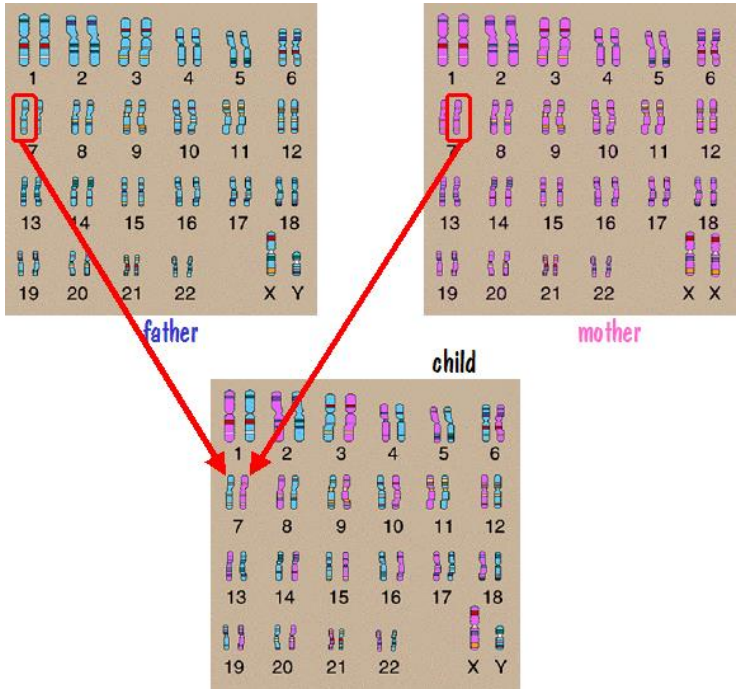
(Golden Helix Blog)

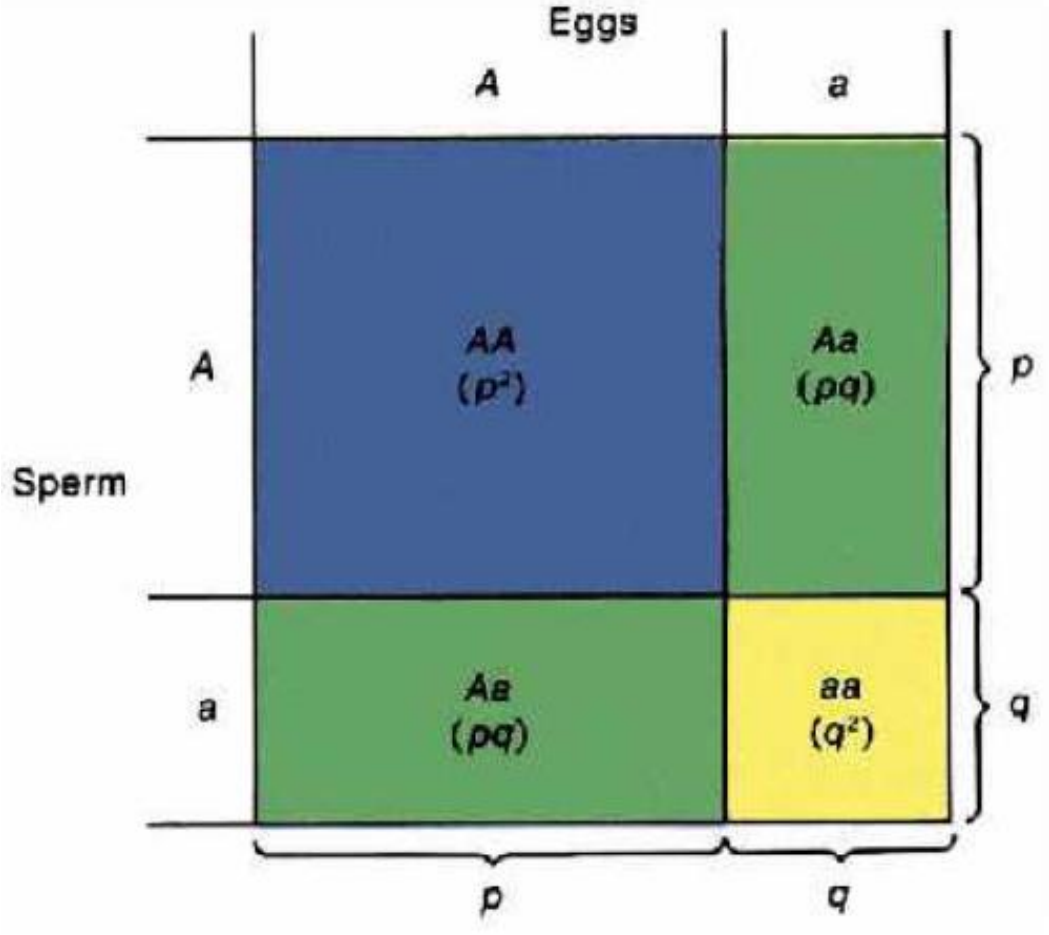
Quality control at the marker level

- **Minor allele frequency (MAF):**
  - Genotype calling algorithms perform poorly for SNPs with low MAF
  - Power is low for detecting associations to genetic markers with low MAF (with standard large-sample statistics)
- **Missing frequency (MiF)**
  - 1 minus call rate
  - MiF needs to be investigated separately in cases and controls because differential missingness may bias association results
- **Hardy-Weinberg equilibrium (HWE)**
  - SNPs excluded if substantially more or fewer subjects heterozygous at a SNP than expected (excess heterozygosity or heterozygote deficiency)

What is Hardy-Weinberg Equilibrium (HWE)?

Consider diallelic SNP with alleles A and a





What is Hardy-Weinberg Equilibrium (HWE)?

Consider diallelic SNP with alleles A_1 and A_2

- Genotype frequencies

  $P(A_1A_1) = p_{11}$, $P(A_1A_2) = p_{12}$, $P(A_2A_2) = p_{22}$
- Allele frequencies $P(A_1) = p = p_{11} + \frac{1}{2}p_{12}$, $P(A_2) = q = p_{22} + \frac{1}{2}p_{12}$

If

- $P(A_1A_1) = p_{11} = p^2$
- $P(A_1A_2) = p_{12} = 2pq$
- $P(A_2A_2) = p_{22} = q^2$

the population is said to be in HWE at the SNP

(Ziegler and Van Steen 2010)

Distorting factors to HWE causing evolution to occur

1. **Non-random mating**

2. **Mutation** - by definition mutations change allele frequencies causing evolution

3. **Migration** - if new alleles are brought in by immigrants or old alleles are taken out by emigrants then the frequencies of alleles will change causing evolution

4. **Genetic drift** - random events due to small population size (bottleneck caused by storm and leading to reduced variation, migration events leading to founder effects)

5. **Natural selection** – some genotypes give higher reproductive success **(Darwin)**

The Travemünde criteria

Level	Filter criterion	Standard value for filter
Sample level	Call fraction	≥ 97%
	Cryptic relatedness	Study specific
	Ethnic origin	Study specific; visual inspection of principal components
	Heterozygosity	Mean ± 3 std.dev. over all samples
	Heterozygosity by gender	Mean ± 3 std.dev. within gender group
SNP level	MAF	≥ 1%
	MiF	≤ 2% in any study group, e.g., in both cases and controls
	MiF by gender	≤ 2% in any gender
	HWE	$p < 10^{-4}$

(Ziegler 2009)

The Travemünde criteria

Level	Filter criterion	Standard value for filter
SNP level	Difference between control groups	$p > 10^{-4}$ in trend test
	Gender differences among controls	$p > 10^{-4}$ in trend test
X-Chr SNPs	Missingness by gender	No standards available
	Proportion of male heterozygote calls	No standards available
	Absolute difference in call fractions for males and females	No standards available
	Gender-specific heterozygosity	No standard value available

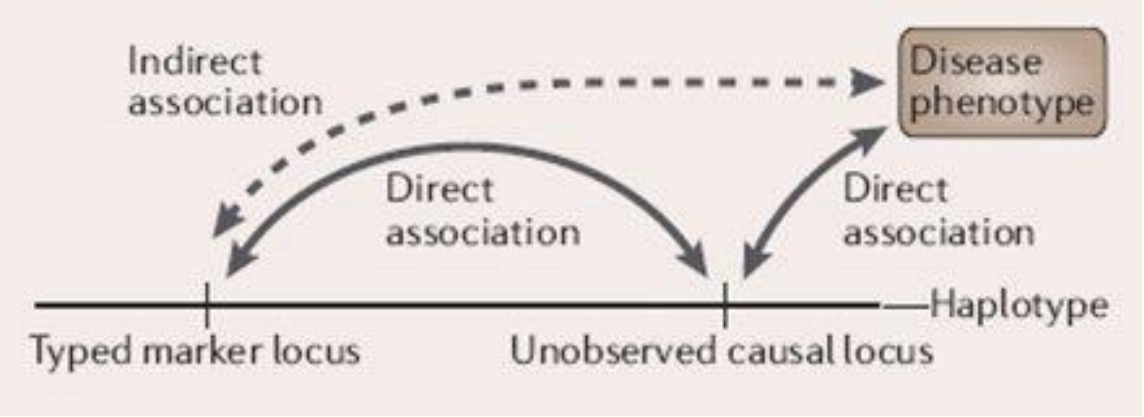
(Ziegler 2009)

4.b Linkage disequilibrium

- **Linkage Disequilibrium** (LD) is a measure of co-segregation of alleles in a population – linkage + allelic association

  Two alleles at different loci that occur together on the same chromosome (or gamete) more often than would be predicted by random chance.

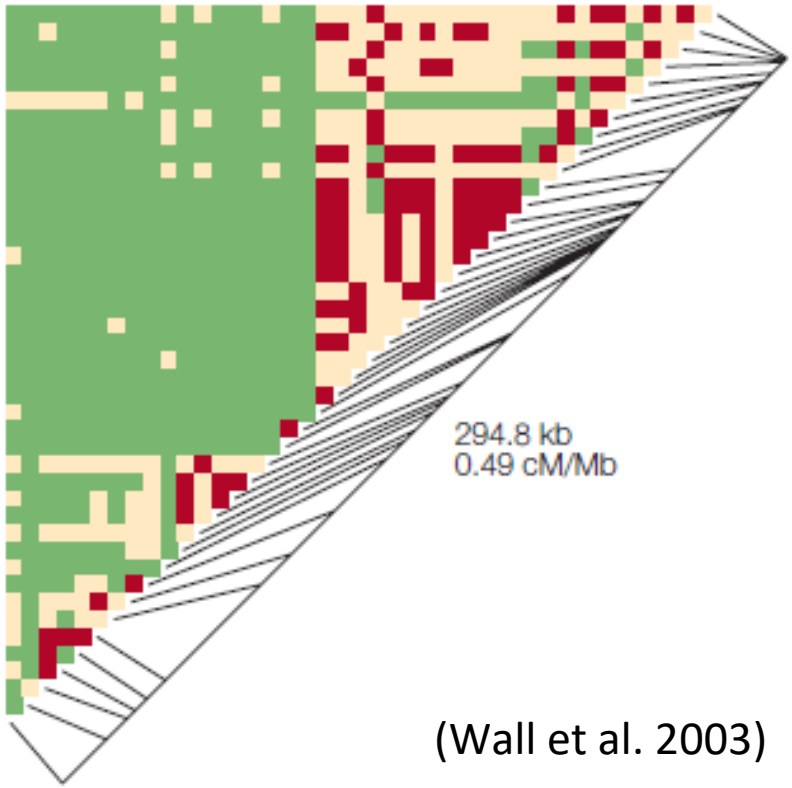
- It is a very important concept for GWAs, since it gives the rational for performing genetic association studies



Different ways to determine LD blocks

- LD-block computation methods

  - often do not allow intermediate regions of low LD between strongly associated SNP pairs:
    - small blocks,
    - high between-block correlations (Kim et al 2018)

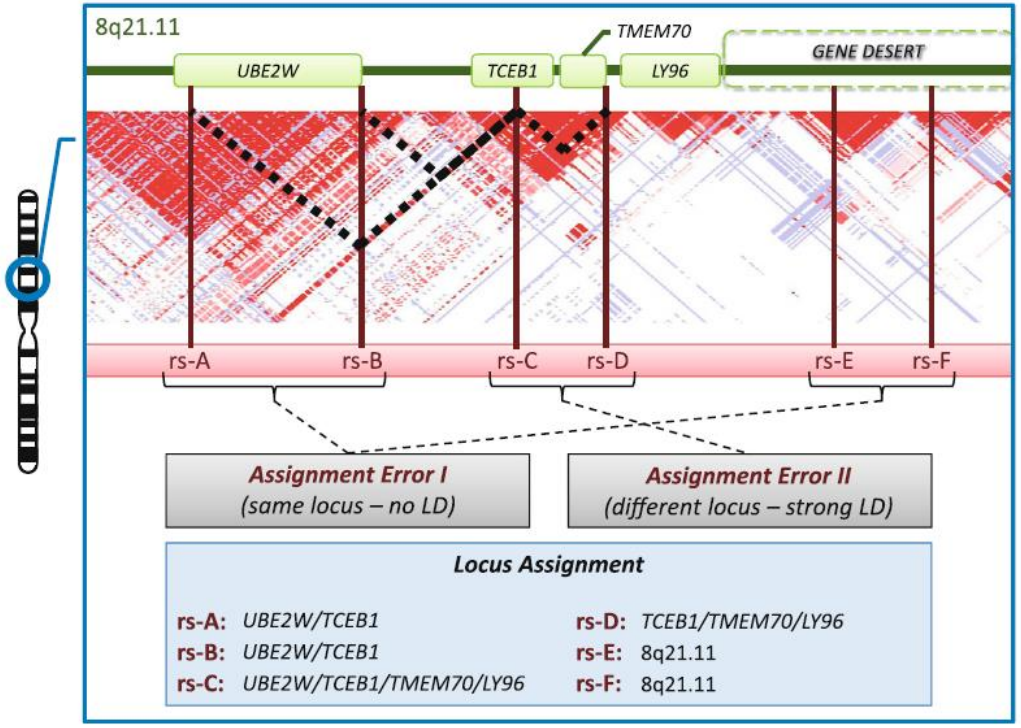


(Wall et al. 2003)

Big-LD and MB-MDR ←→ classical enrichment (Junior et al.)

- **LD based locus assignment and its error sources**
  - whole genomic region captu
    with the marker originally re
  - loci: genes located within th

(Arnold et al 2012)



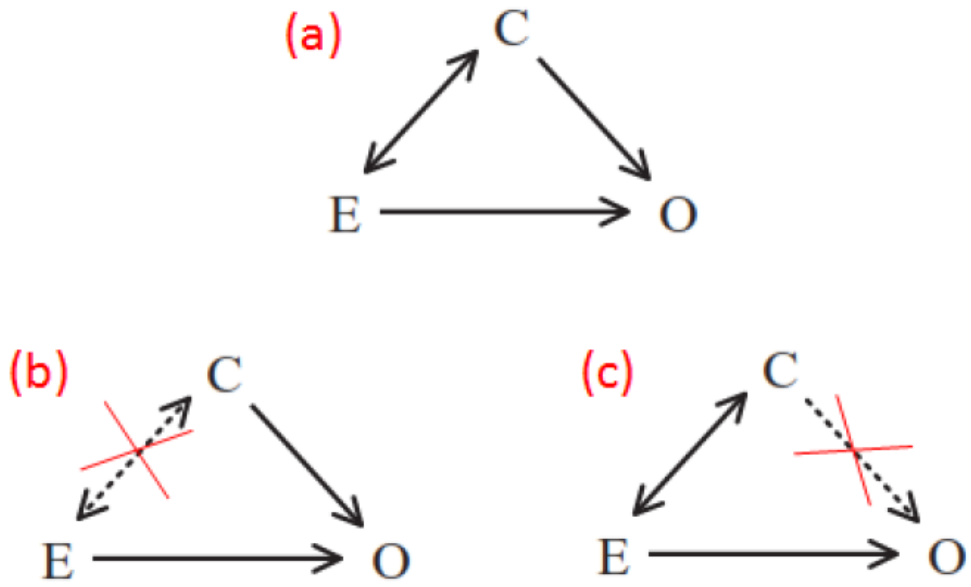
4.c Confounding by shared genetic ancestry – “population stratification”

If successful, the random allocation of subjects to the exposure which characterise RCTs ensures a balanced distribution of known and unknown confounding factors between exposed and non-exposed subjects. This is equivalent of removing the association between the exposure and all potential confounders (Figure 1b), and therefore, the possibility of confounding itself. In this case, the effect of the exposure on the outcome can be directly estimated by simply comparing outcomes between exposed and unexposed subjects (1).

Regression uses mathematical modelling to estimate the effect of confounders on the outcome, and to "remove" this effect statistically. This is equivalent of removing (or, more realistically, reducing) the association between confounder and outcome, thus eliminating the second necessary condition for confounding (Figure 1c).

Two necessary — albeit not sufficient — conditions for an extraneous factor ("confounder") to produce such a bias are (Figure 1a):

1. the confounder is a risk factor for the outcome;
2. the confounder is associated with the exposure, i.e. its distribution is different among individuals with different exposure status.

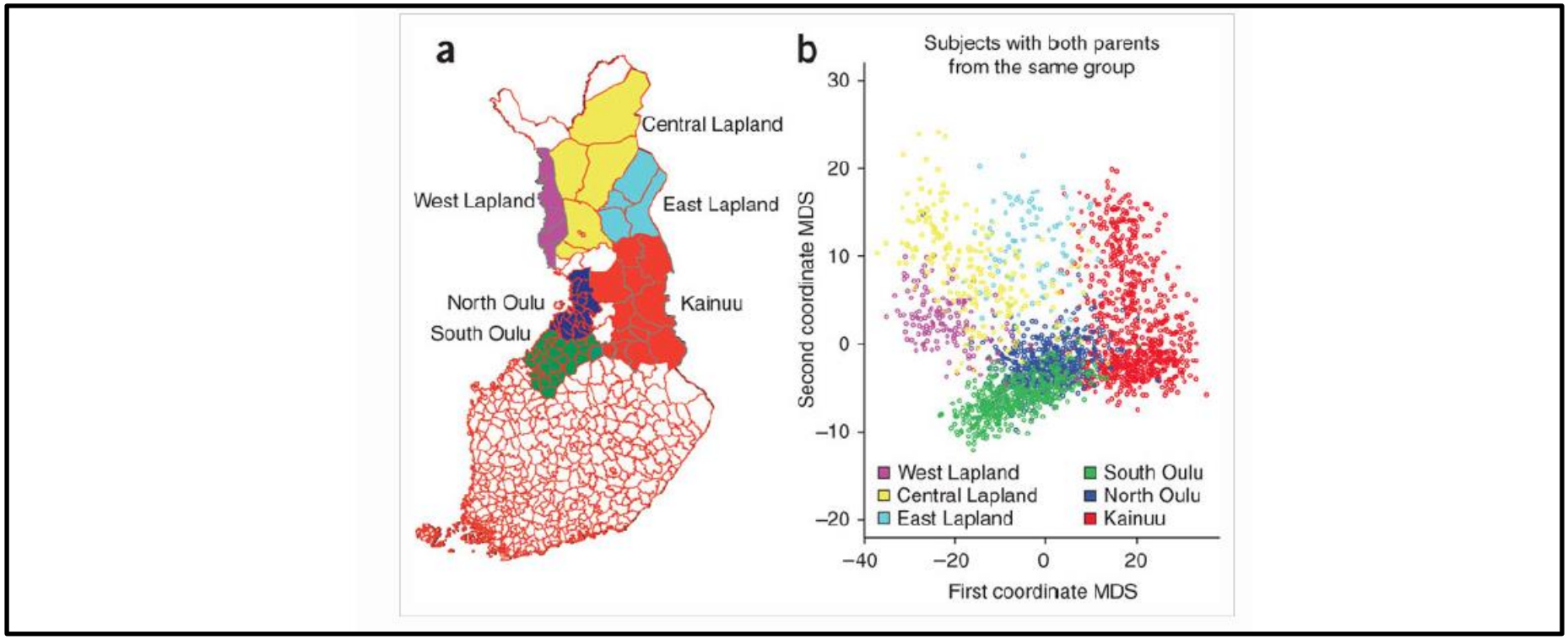


(Cois 2014)

Figure 1: Schematic illustration of confounding control. Arrows represent causal effects, double arrows associations of any nature. E = exposure, C = confounder, O = outcome.

Confounding by shared genetic ancestry: heterogeneity in populations

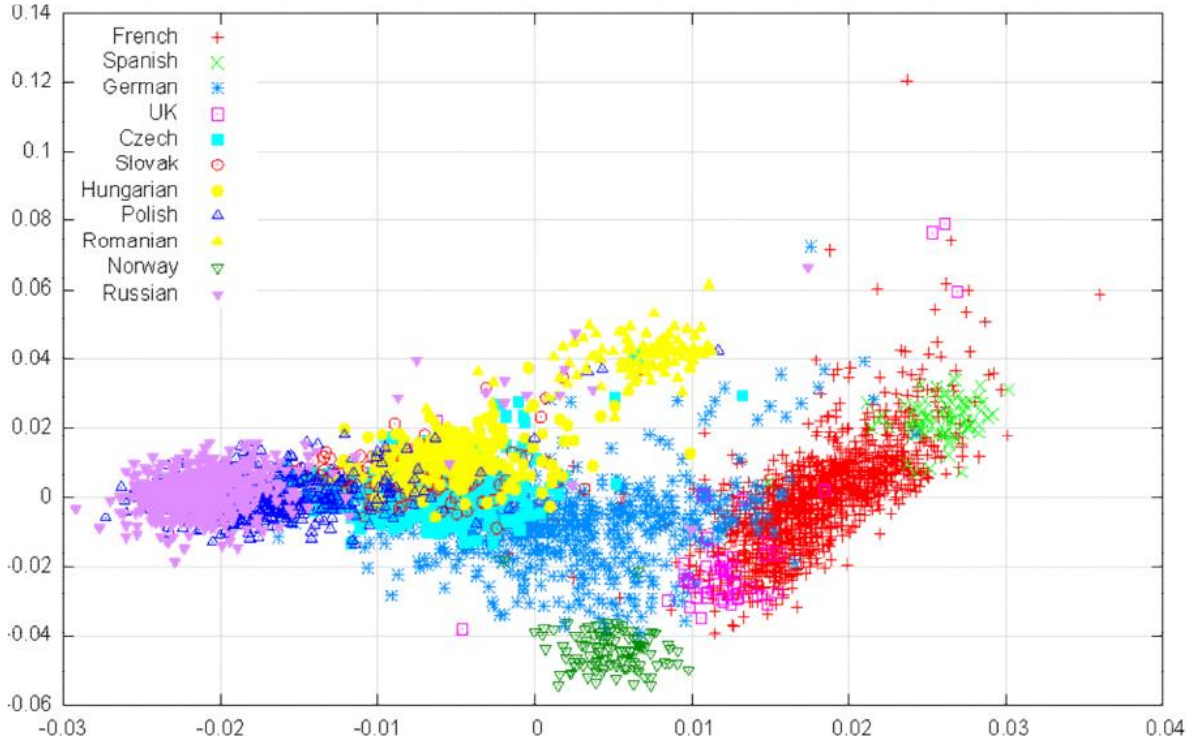
- There can be population structure in all populations, even those that appear to be relatively “homogeneous”



(Sabatti et al. 2009)

Confounding by shared genetic ancestry : creating a PC space

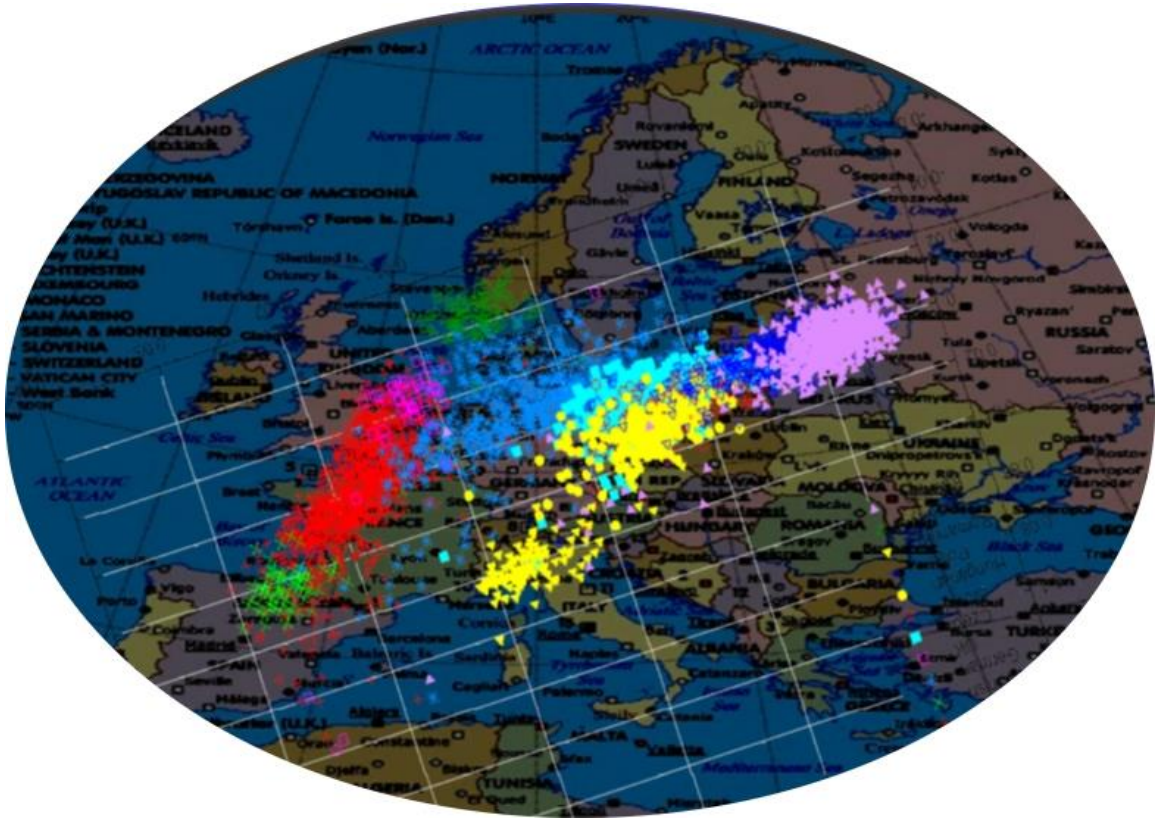
- In European data, the first 2 principal components “nicely” reflect the N-S and E-W axes !



Y-axis: PC2 (6% of variance); X-axis: PC1 (26% of variance)

Confounding by shared genetic ancestry : creating a PC space

- In European data, the first 2 principal components “nicely” reflect the N-S and E-W axes !



Y-axis: PC2 (6% of variance); X-axis: PC1 (26% of variance)

The versatile use of PCs in genetic epidemiology

Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 17

A Family-Based Association Test for Repeatedly Measured Quantitative Traits Adjusting for Unknown Environmental and/or Polygenic Effects

Christoph Lange, *Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA*
Kristel van Steen, *Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave, Boston, MA 02115*
Toby Andrew, *Twin Research & Genetic Epidemiology Unit, St Thomas' Hospital, London SE1 7EH, UK*
Helen Lyon, *Brigham and Women's Hospital, Harvard Medical School, Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115, USA*
Dawn L. DeMeo, *Brigham and Women*
Benjamin Raby, *Brigham and Women's Hospital, Harvard Medical School, Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115, USA*
Amy Murphy, *Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA*
Edwin K. Silverman, *Brigham and Women*
Alex MacGregor, *Twin Research & Genetic Epidemiology Unit, St Thomas' Hospital, London SE1 7EH, UK*
Scott T. Weiss, *Brigham and Women's Hospital, Harvard Medical School, Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115, USA*
Nan M. Laird, *Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA*

Not everything in life is linear!

OXFORD

Briefings in Bioinformatics, 00(00), 2018, 1–17

doi: 10.1093/bib/bby081
Advance Access Publication Date: 14 September 2018
Review Article

Principals about principal components in statistical genetics

Fentaw Abegaz, Kridsadakorn Chaichoompu, Emmanuelle Génin, David W. Fardo, Inke R. König, Jestinah M. Mahachie John and Kristel Van Steen

Corresponding author: Fentaw Abegaz, GIGA-R, Medical Genomics–BIO3, University of Liege, Liege, Belgium. Tel.: +32 43669965; E-mail: y.fabegaz@ulg.ac.be

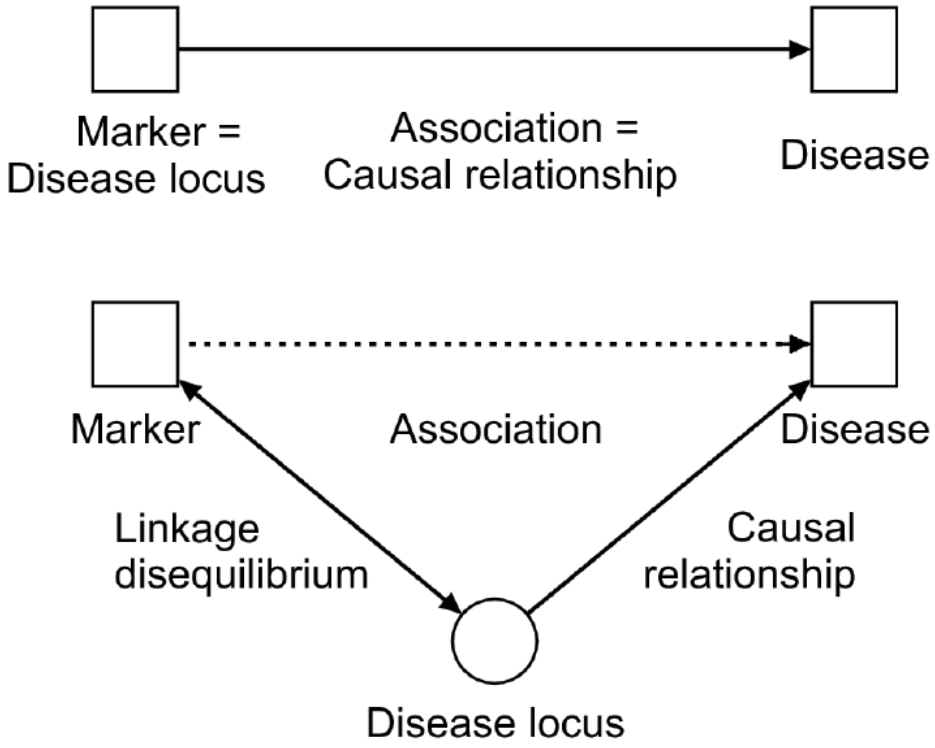
Abstract

Principal components (PCs) are widely used in statistics and refer to a relatively small number of uncorrelated variables derived from an initial pool of variables, while explaining as much of the total variance as possible. Also in statistical genetics, principal component analysis (PCA) is a popular technique. To achieve optimal results, a thorough understanding about the different implementations of PCA is required and their impact on study results, compared to alternative approaches. In this review, we focus on the possibilities, limitations and role of PCs in ancestry prediction, genome-wide association studies, rare variants analyses, imputation strategies, meta-analysis and epistasis detection. We also describe several variations of classic PCA that deserve increased attention in statistical genetics applications.

Key words: principal component analysis; population stratification; statistical genetics; exploration and prediction

5 Analysis Steps

5.a Testing for Genetic Associations (focus on SNPs)



(Ziegler and Van Steen 2010)

What if my Y is binary? Testing for association between case/control status and a SNP

- Fill in the table below and perform a chi-squared test for independence between rows and columns → **genotype test → 2 df**

	AA	**Aa**	**aa**
Cases			
Controls			

Sum of entries = cases+controls

- Fill in the table below and perform a *chi-squared test for independence* between rows and columns → **allelic test (ONLY valid under HWE) → 1df**

	A	**a**
Cases		
Controls		

Sum of entries is 2 x (cases + controls)

Toy example of chi-square test of independence

	Blue	Green	Pink	
Boys	100(72)	150(108)	20(120)	300
Girls	20(48)	30(72)	180(80)	200
	120	180	200	N = 500

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(100-72)^2}{72} + \frac{(20-48)^2}{48} + \frac{(150-108)^2}{108} + \frac{(30-72)^2}{72} + \frac{(20-120)^2}{120} + \frac{(180-80)^2}{80}$$

What is the df ?

The impact of different encoding schemes for SNPs

	Coding scheme for statistical modeling/testing					
Indiv. genotype	**X1**	**X1**	**X2**	**X1**	**X1**	**X1**
	Additive coding	**Genotype coding (general mode of inheritance)**		**Dominant coding (for a)**	**Recessive coding (for a)**	**Advantage Heterozygous**
AA	**0**	**0**	**0**	**0**	**0**	**0**
Aa	**1**	**1**	**0**	**1**	**0**	**1**
aa	**2**	**0**	**1**	**1**	**1**	**0**

Tests in GWAS using the regression framework

- **Example 1:**

$$Y = \beta_0 + \beta_1 SNP + \varepsilon$$

  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 \neq 0$
  - $\mathrm{df_F} = \mathrm{n} - 2$ (this links to df in variance estimation)
  - $\mathrm{df_R} = \mathrm{n} - 1$ (this links to df in variance estimation)

The variance of a discrete random variable is:

$$\sigma_X^2 = \sum_{All\,x} (x - \mu_X)^2 p(x)$$

It can be shown that for testing $\beta_1 = 0$ versus $\beta_1 \neq 0$

- $F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} : \frac{SSE(F)}{df_F} = \frac{b_1^2}{s^2(b_1)} = (t^*)^2$

Tests in GWAS using the regression framework

- **Example 2:**

$$Y = \beta_0 + \beta_1 SNP + \beta_2 PC_1 + \beta_3 PC_2 + \varepsilon$$

  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 \neq 0$
  - $\mathrm{df_F} = \mathrm{n} - 4$
  - $\mathrm{df_R} = \mathrm{n} - 3$

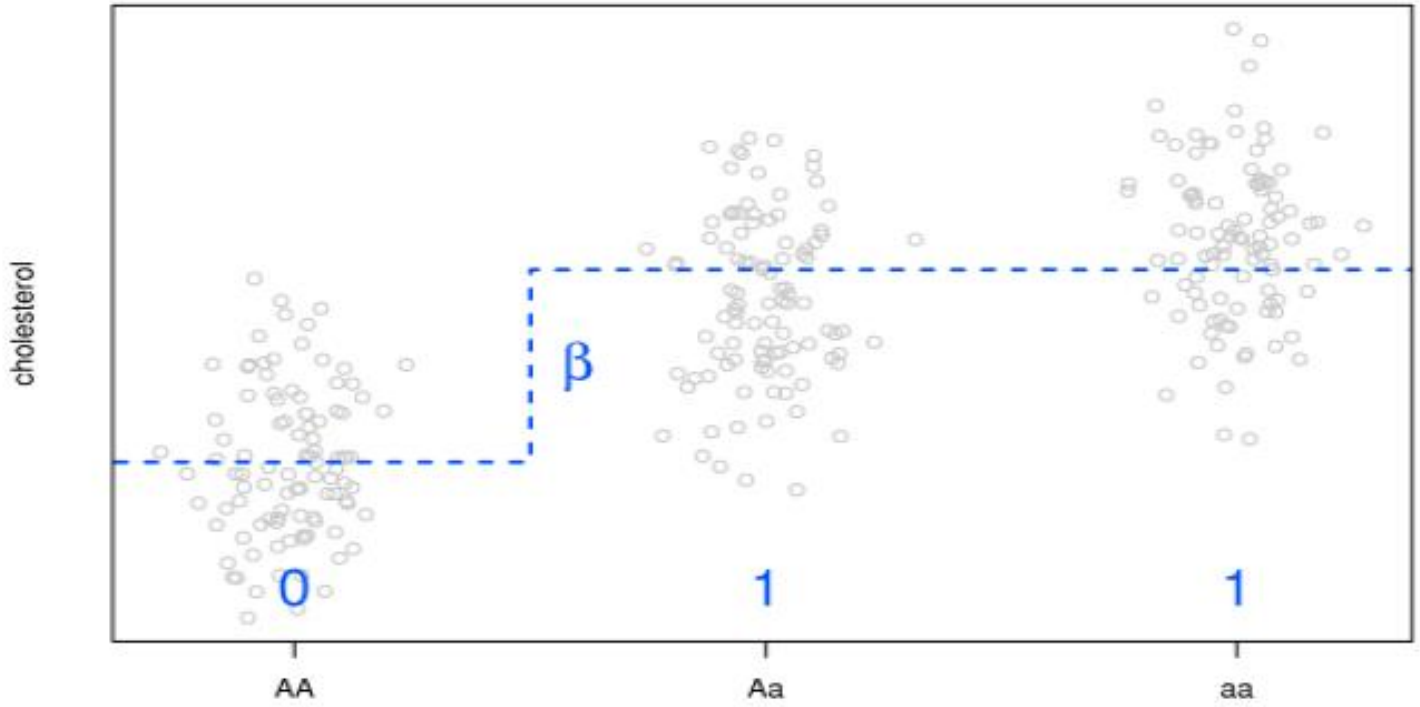
How many dfs would the corresponding F-test have?

How many dfs would a corresponding $t(^2)$ test have?

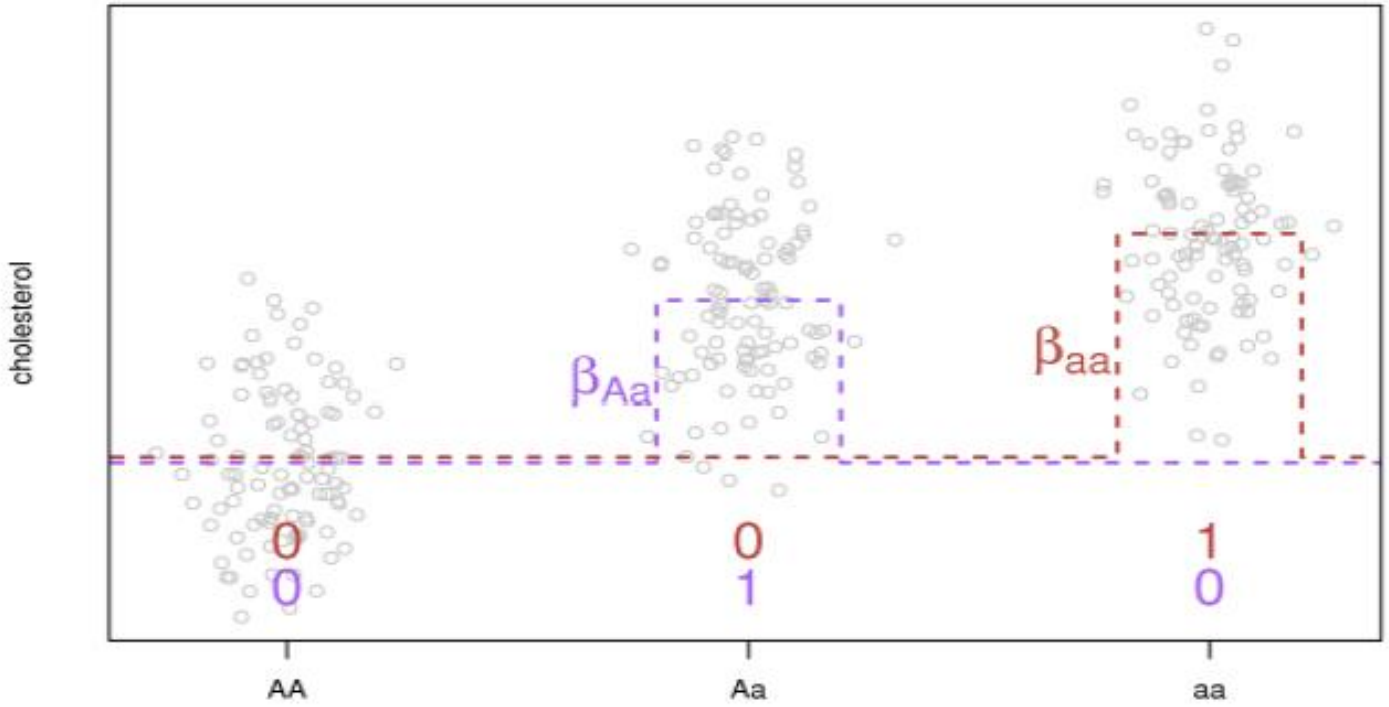
The impact of different encoding schemes for SNPs

	Coding scheme for statistical modeling/testing					
Indiv. genotype	**X1**	**X1**	**X2**	**X1**	**X1**	**X1**
	Additive coding	**Genotype coding (general mode of inheritance)**		**Dominant coding (for a)**	**Recessive coding (for a)**	**Advantage Heterozygous**
AA	**0**	**0**	**0**	**0**	**0**	**0**
Aa	**1**	**1**	**0**	**1**	**0**	**1**
aa	**2**	**0**	**1**	**1**	**1**	**0**

Which encoding scheme provides a good fit to the data?

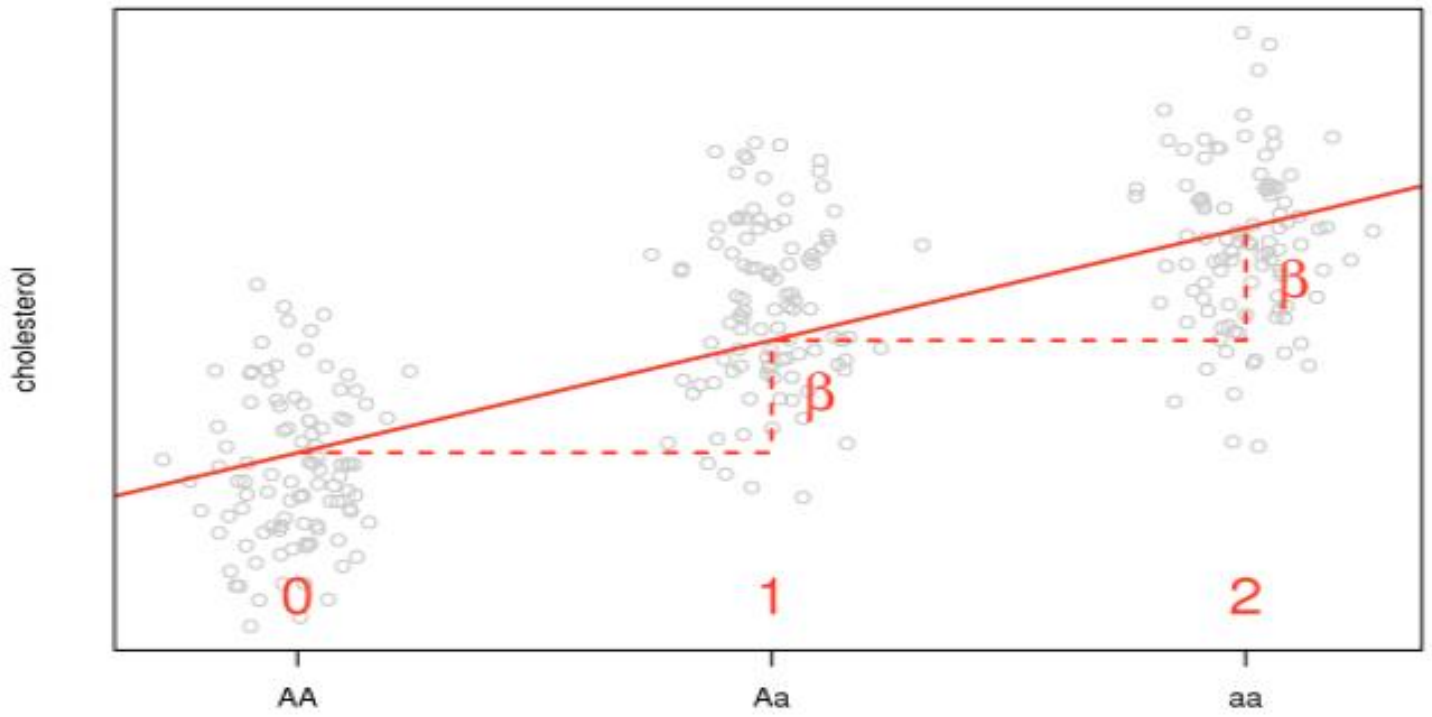


Which encoding scheme provides a good fit to the data?



Robust vs overkill ?

Which encoding scheme provides a good fit to the data?



Most commonly used

Model diagnostics

- There are 4 principal assumptions which justify the use of **linear regression** models for purposes of prediction:
  - **linearity** of the relationship between dependent and independent variables
  - independence of the errors (no serial correlation)
  - homoscedasticity (constant variance) of the errors
    - versus time (when time matters)
    - versus the predictions (or versus any independent variable)
  - normality of the error distribution. (http://www.duke.edu/~rnau/testing.htm)

- To check **model assumptions**: go to **quick-R** and regression diagnostics (http://www.statmethods.net/stats/rdiagnostics.html)

QQ plots for model diagnostics – Q for Quantile

- Quantiles are points in your data below which a certain proportion of your data fall.

  What is the 0.5 quantile for normally distributed data?

- Here we generate a random sample of size 200 from a normal distribution and find the quantiles for 0.01 to 0.99 using the quantile function:

```
quantile(rnorm(200),probs = seq(0.01,0.99,0.01))
```

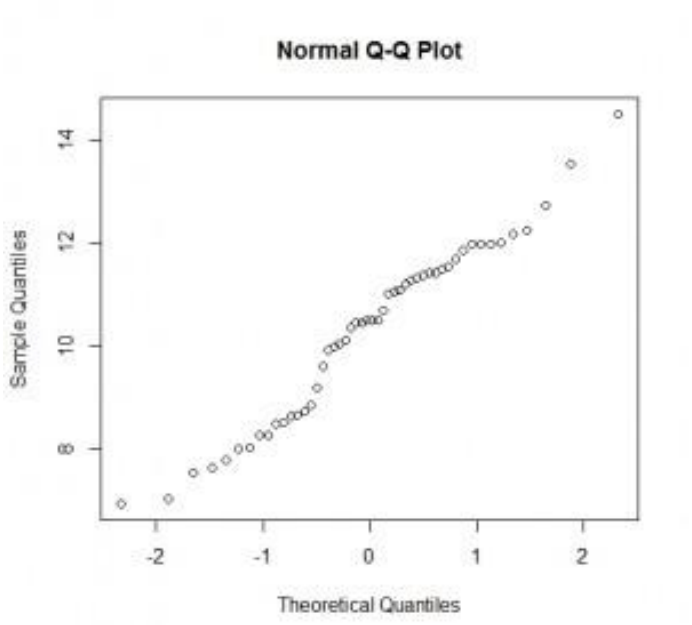
- Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution.
  The number of quantiles is selected to match the size of your sample data.

  The quantile function in R offers 9 different quantile algorithms!

  See help(quantile)

QQ plots for model diagnostics – Q for Quantile

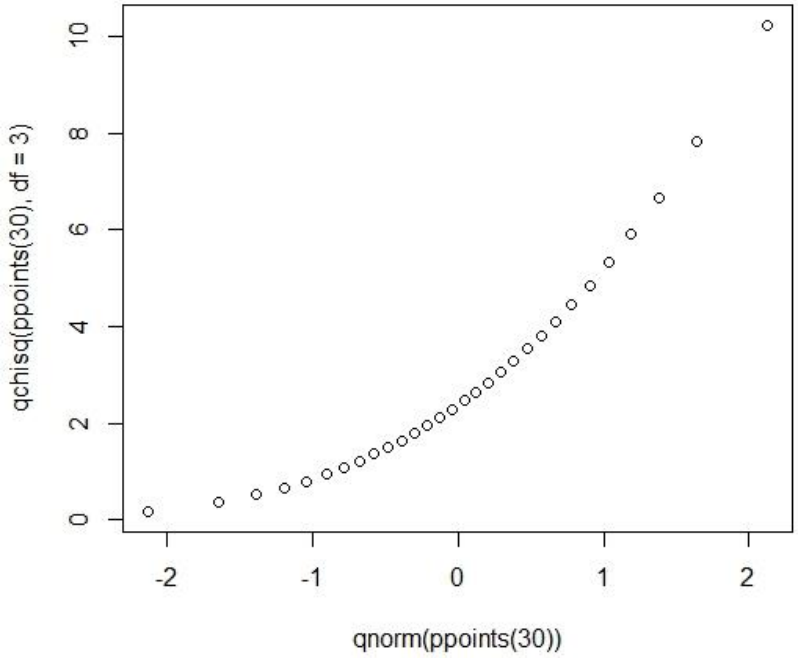
- A Q-Q plot is a scatterplot created by plotting **two sets of quantiles** against one another.
- If both sets of quantiles come from the same distribution, we should see the points forming a line that's roughly straight.
- Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



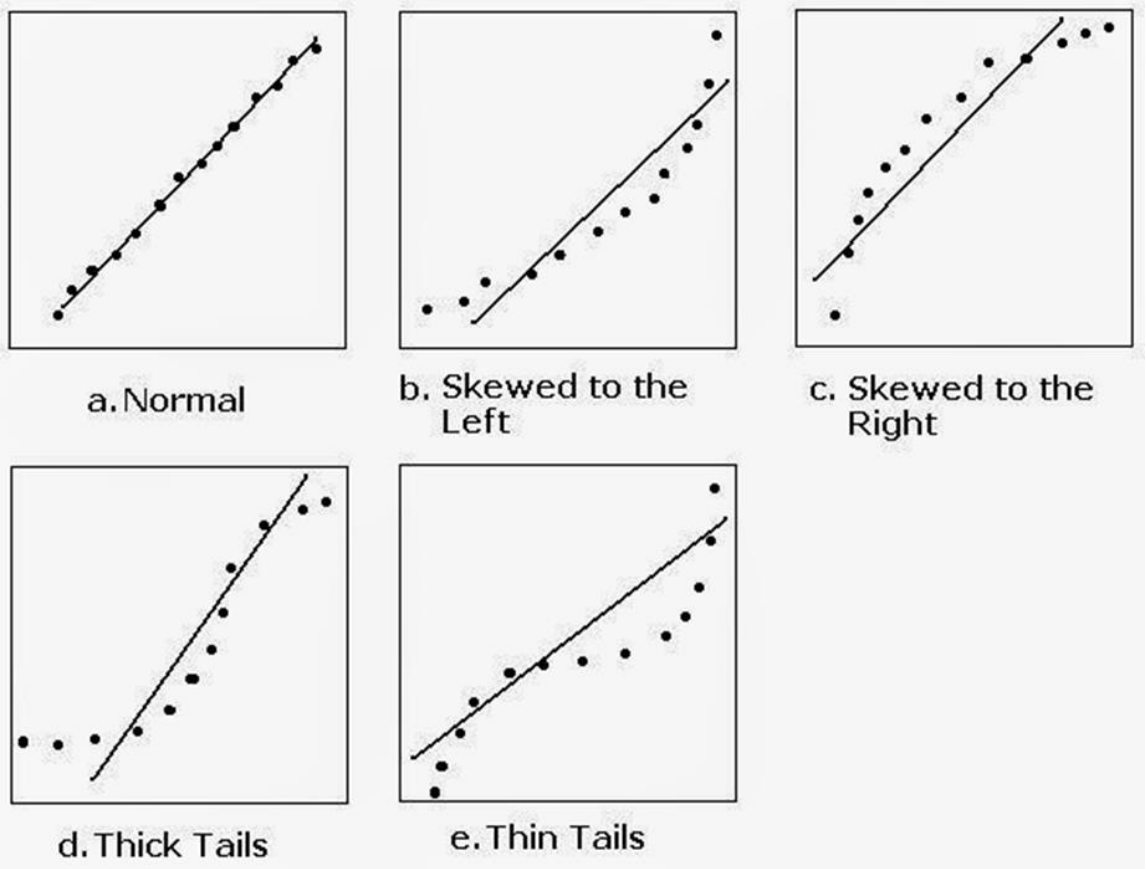
Examples of QQ plots: no straight line

- QQ plot of a distribution that's skewed right; a Chi-square distribution with 3 degrees of freedom against a Normal distribution

qqplot(qnorm(ppoints(30)), qchisq(ppoints(30),df=3))



Examples of QQ plots: some frequent scenarios

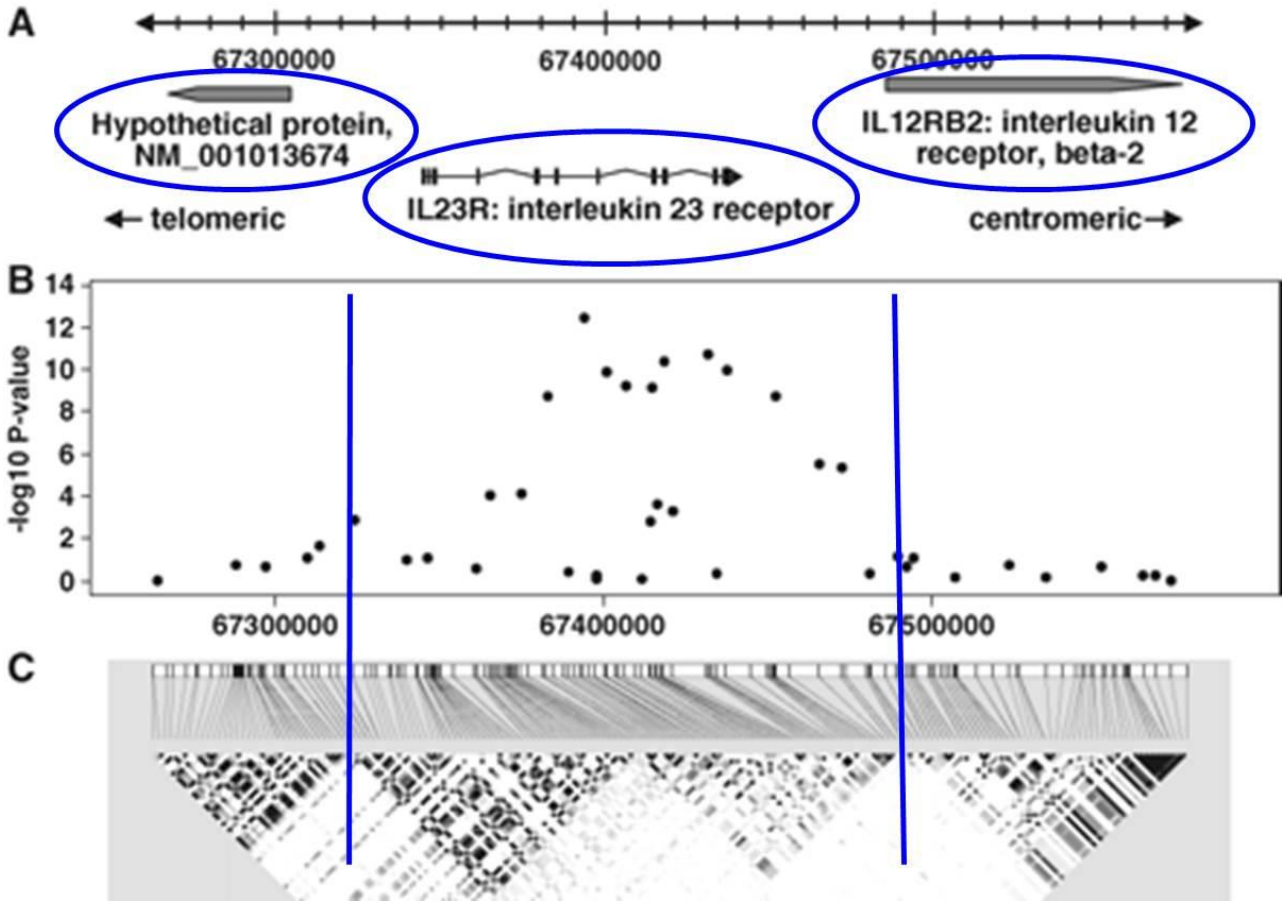


5.b Causation

Establishing causation: causal variants for human complex traits

- Wet lab efforts
  - Gene knock-out experiments
    - The findings of animal experiments may not be directly applicable to the human situation because of genetic, anatomic, and physiologic differences
- Dry lab efforts
  - As opposed to association studies that benefit from LD, the main challenge in identifying causal variants at associated loci analytically lies in distinguishing among the many closely correlated variants due to LD.

Finding the “relevant” loci – naïve approach



(Duerr et al 2006)

Genome-wide Causation Studies of Complex Diseases

Rong Jiao[1], Xiangning Chen[2], Eric Boerwinkle[3] & Momiao Xiong[1*]

[1]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA

[2] Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, Nevada, USA

[3]Epidemiology, Human Genetics & Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA

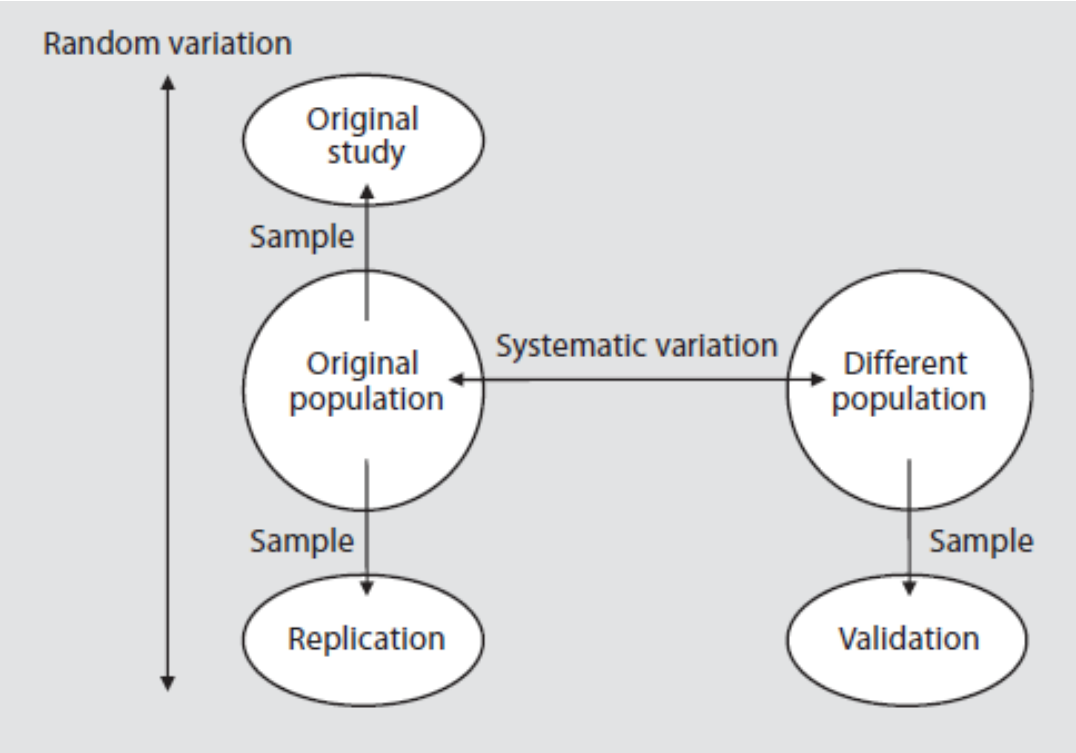
Key words: Causal inference, GWAS, GWCS, additive noise models, linkage disequilibrium, prediction

(https://arxiv.org/ftp/arxiv/papers/1907/1907.07789.pdf)

6 Post Association Analysis Steps

6.a Replication and Validation

The difference



(Igl et al. 2009)

Guidelines for replication studies

- Replication studies should be of sufficient size to demonstrate the effect
- Replication studies should conducted in independent datasets
- Replication should involve the same phenotype
- Replication should be conducted in a similar population
- The same SNP should be tested
- The replicated signal should be in the same direction
- Joint analysis should lead to a lower *p*-value than the original report
- Well-designed negative studies are valuable

Note that SNPs are most likely to replicate when they
- show modest to strong statistical significance,
- have common minor allele frequency,
- exhibit modest to strong **genetic effect size** (~strength of association)

6.b GWA Interpretation and follow-up

REVIEW

The Post-GWAS Era: From Association to Function

Michael D. Gallagher[1,2] and Alice S. Chen-Plotkin[1,*]

During the past 12 years, genome-wide association studies (GWASs) have uncovered thousands of genetic variants that influence risk for complex human traits and diseases. Yet functional studies aimed at delineating the causal genetic variants and biological mechanisms underlying the observed statistical associations with disease risk have lagged. In this review, we highlight key advances in the field of functional genomics that may facilitate the derivation of biological meaning post-GWAS. We highlight the evidence suggesting that causal variants underlying disease risk often function through regulatory effects on the expression of target genes and that these expression effects might be modest and cell-type specific. We moreover discuss specific studies as proof-of-principle examples for current statistical, bioinformatic, and empirical bench-based approaches to downstream elucidation of GWAS-identified disease risk loci.

(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5986732/)

Some criteria to assess the functional relevance of a variant

Criteria	Strong support for functional significance	Moderate support for functional significance	Evidence against functional significance
Nucleotide sequence	Variant disrupts a known functional or structural motif	Variant is a missense change or disrupts a putative functional motif; changes to protein structure might occur	Variant disrupts a non-coding region with no known functional or structural motif
Evolutionary conservation	Consistent evidence from multiple approaches for conservation across species and multigene families	Evidence for conservation across species or multigene families	Nucleotide or amino-acid residue not conserved
Population genetics	In the absence of laboratory error, strong deviations from expected population frequencies in cases and/or controls in a particular ethnicity	In the absence of laboratory error, moderate to small deviations from expected population frequencies in cases and/or controls; effects are not well characterized by ethnicity	Population genetics data indicates no deviations from expected proportions
Experimental evidence	Consistent effects from multiple lines of experimental evidence; effect in human context is established; effect in target tissue is known	Some (possibly inconsistent) evidence for function from experimental data; effect in human context or target tissue is unclear	Experimental evidence consistently indicates no functional effect
Exposures (for example, genotype–environment interaction studies)	Variant is known to affect the metabolism of the exposure in the relevant target tissue	Variant might affect metabolism of the exposure or one of its components; effect in target tissue might not be known	Variant does not affect metabolism of exposure of interest
Epidemiological evidence	Consistent and reproducible reports of moderate-to-large magnitude associations	Reports of association exist; replication studies are not available	Prior studies show no effect of variant

(Rebbeck et al 2004)

Finding the “relevant” loci - via “functional genomics”

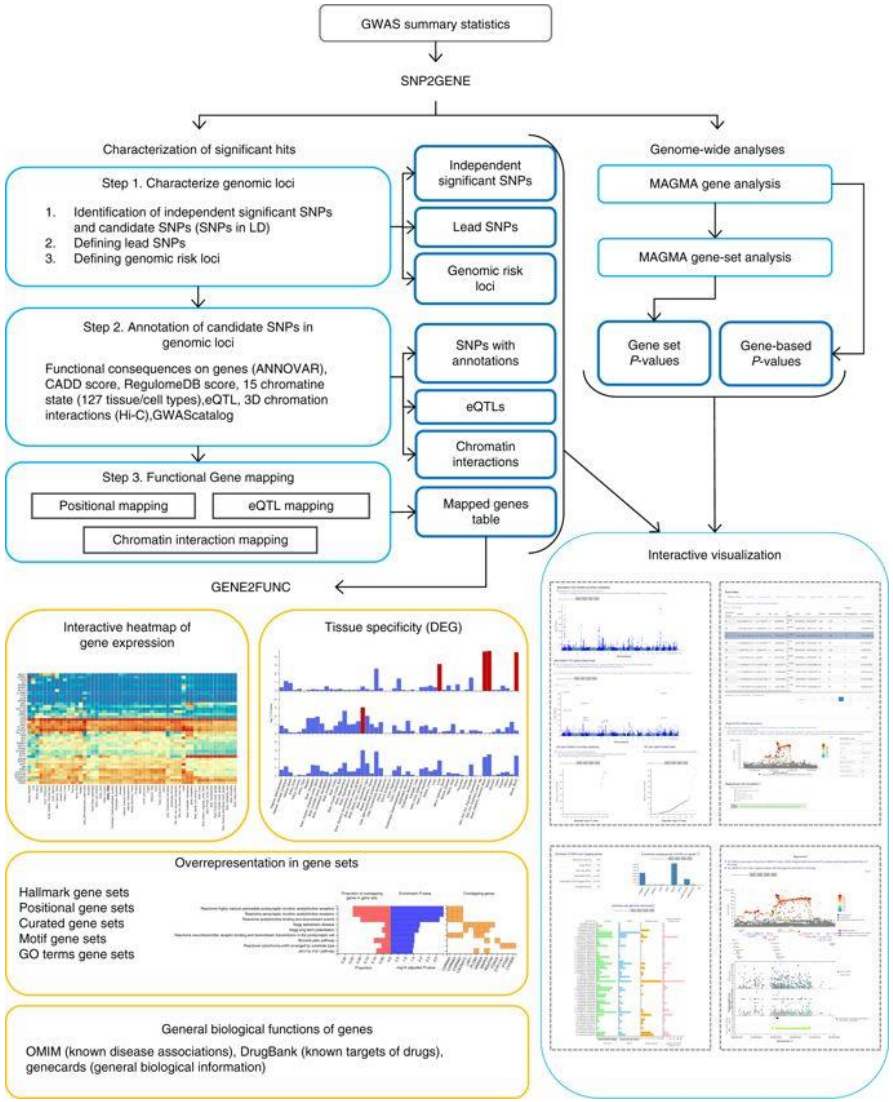
(https://data.broadinstitute.org/mpg/depict/)

Finding the “relevant” loci - via “functional genomics”

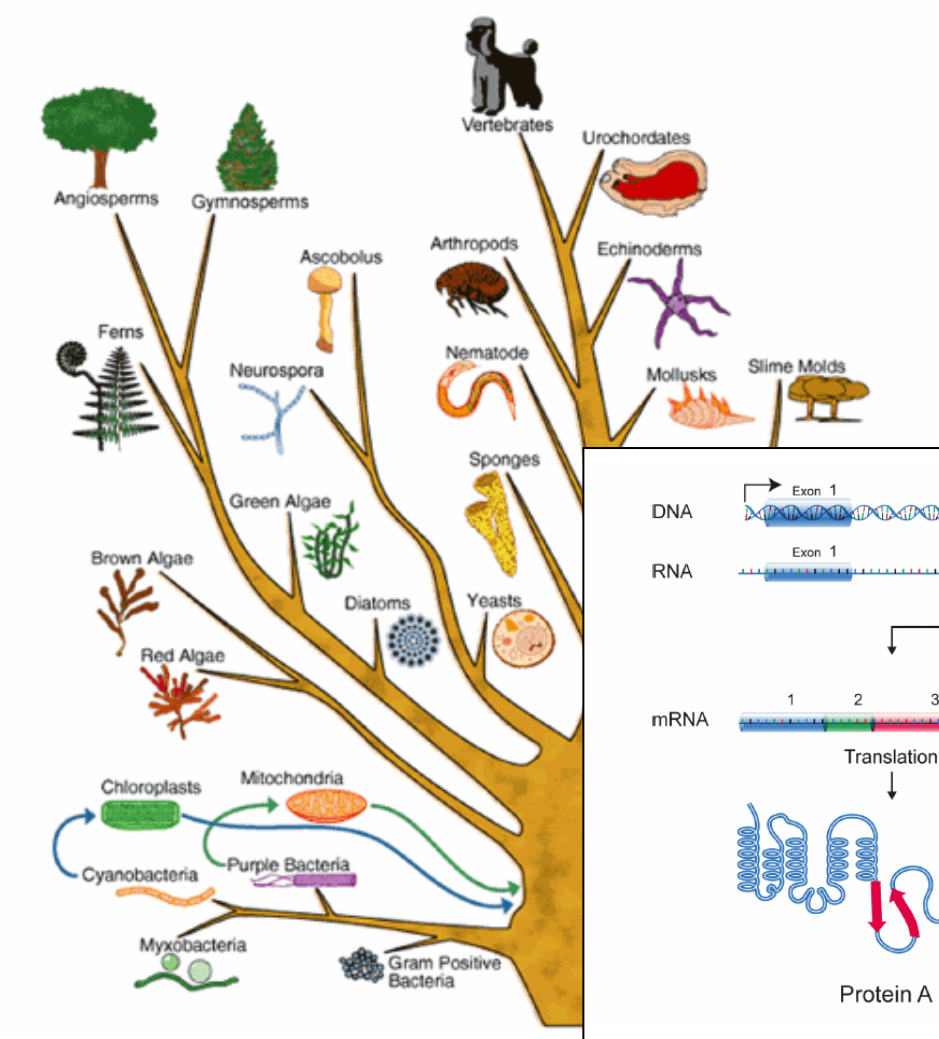
FUMA on GWAS summary statistics. SNP2GENE prioritizes functional SNPs and genes, outputs tables (blue boxes), and creates Manhattan, quantile–quantile (QQ) and interactive regional plots (box at right bottom).

GENE2FUNC provides four outputs; a gene expression heatmap, enrichment of differentially expressed gene (DEG) sets in a certain tissue compared to all other tissue types, overrepresentation of gene sets, and links to external biological information of input genes.

(https://fuma.ctglab.nl/)

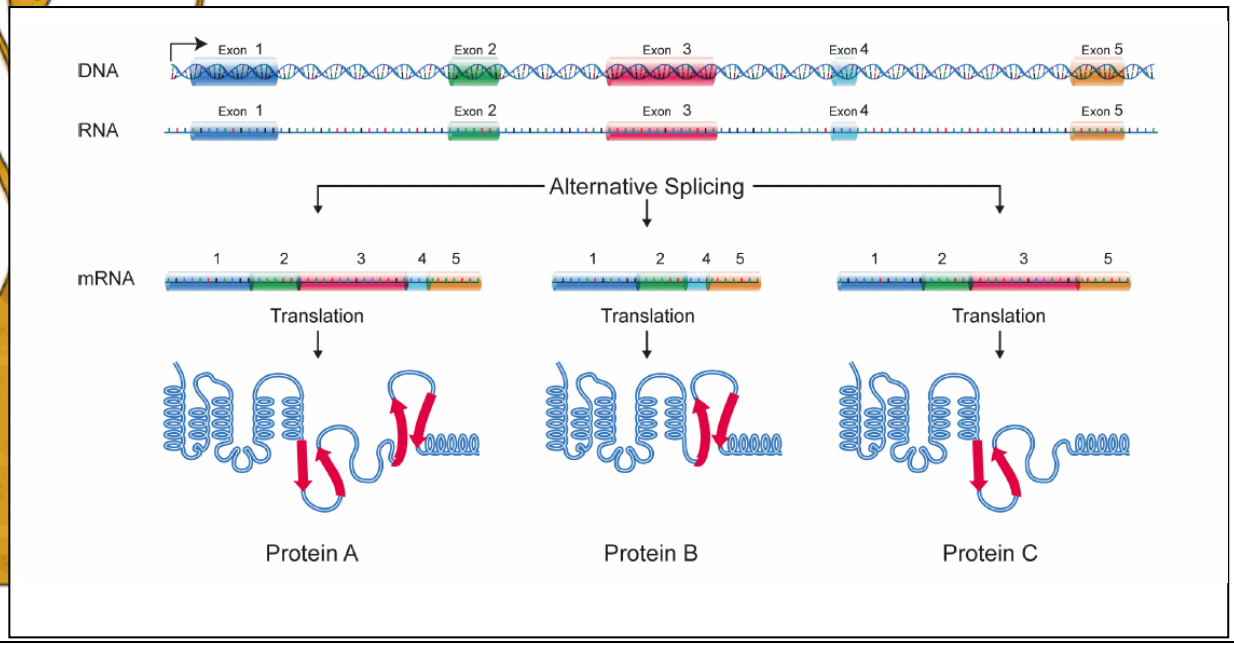


Natural to look at gene expression (in its full complexity!)



Q: "How can something as complicated as a human have only 25 percent more genes than the tiny roundworm *C. elegans*?"

Part of the answer seems to involve **alternative splicing:**



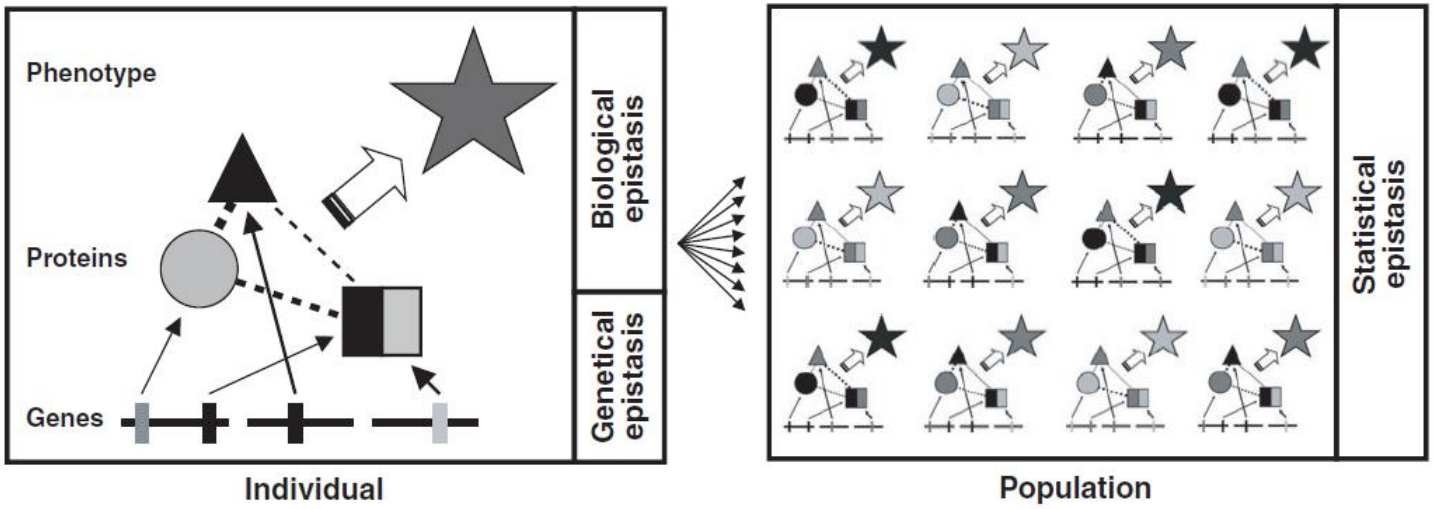
7 Epistasis analysis

The **interactome** refers to the entire complement of interactions between DNA, RNA, proteins and metabolites within a cell. These interactions are influenced by genetic alterations and environmental stimuli. As a consequence, the interactome should be examined or considered in ***particular contexts***.

Formal definition of epistasis

(Moore 2005; Moore and Williams 2005)

- The original definition (**driven by biology**) refers to a variant or allele at one locus preventing the variant at another locus from manifesting its effect (William Bateson 1861-1926).
- A later definition of epistasis (**driven by statistics**) is expressed in terms of deviations from a model of additive multiple effects (Ronald Fisher 1890-1962).



(Logistic) Regression

- Alternatively, we can assume additive effects of each allele at each locus, leading to a single interaction term (instead of 4 next!)

		Locus H	
Locus G	2	1	0
2	$\beta_0 + 2\beta_G + 2\beta_H + 4\beta$	$\beta_0 + 2\beta_G + \beta_H + 2\beta$	$\beta_0 + 2\beta_G$
1	$\beta_0 + \beta_G + 2\beta_H + 2\beta$	$\beta_0 + \beta_G + \beta_H + \beta$	$\beta_0 + \beta_G$
0	$\beta_0 + 2\beta_H$	$\beta_0 + \beta_H$	β_0

- This corresponds in statistical analysis packages to the model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_G X_1 + \beta_H X_2 + \beta X_1 X_2$$

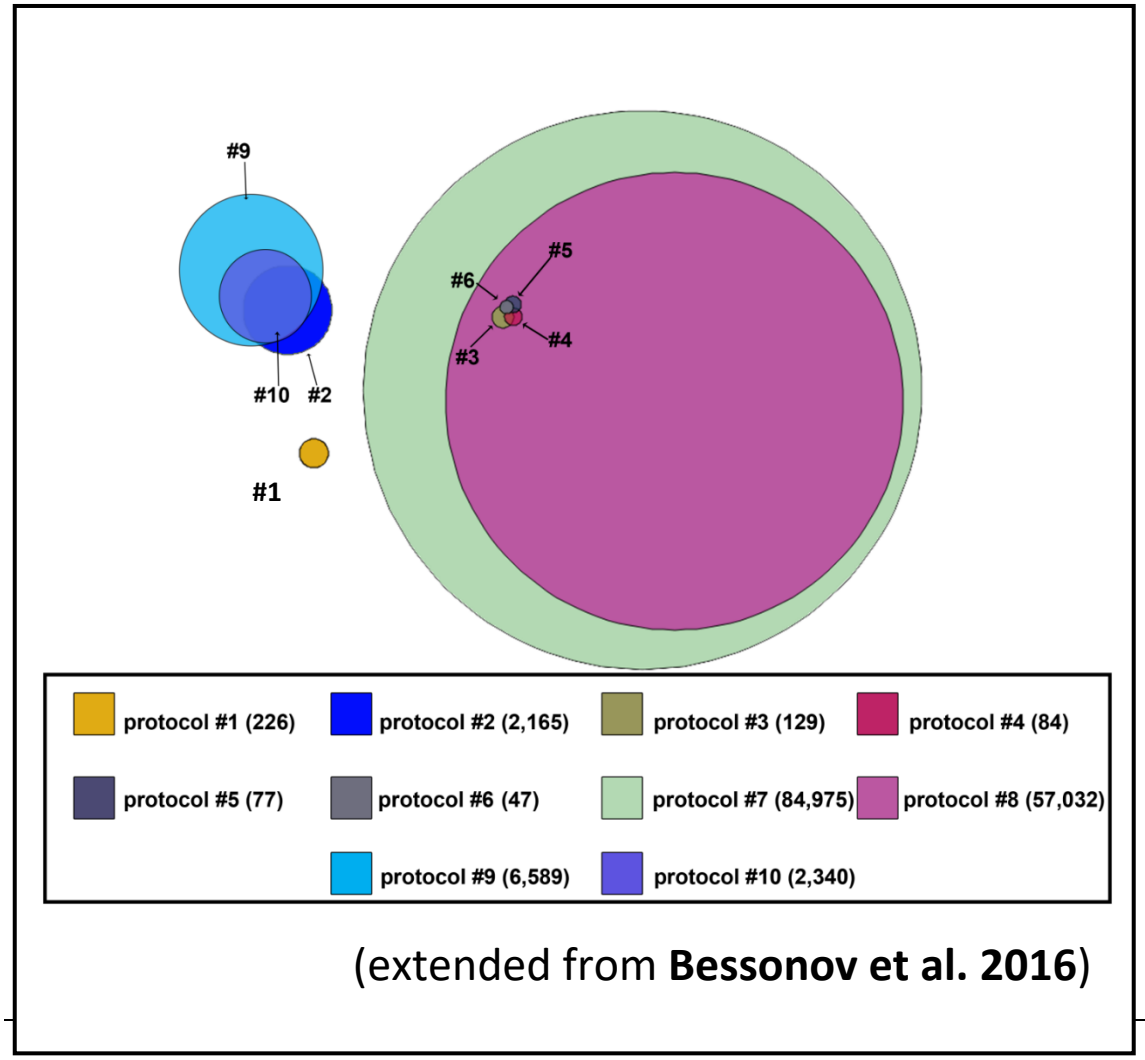
and **dosage encoding for X1 and X2**.

(Logistic) Regression

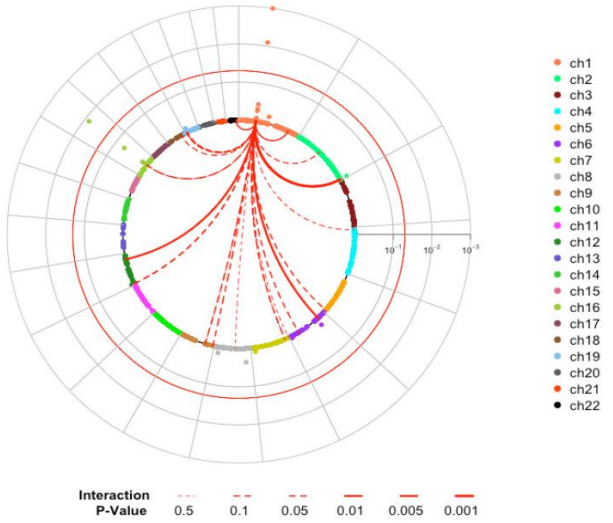
- Most general saturated (9 parameter) genotype model allows all 9 penetrances to take different values
- Log odds is modelled in terms of a baseline effect (β_0), main effects of locus *G* (β_{G1}, β_{G2}), main effects of locus *H* (β_{H1}, β_{H2}), 4 int. terms
- This corresponds in statistical analysis packages to **encoding X1, X2 (0,1,2) as a "factor"**

	Locus H		
Locus G	2	1	0
2	$\beta_0+\beta_{G2}+\beta_{H2}+\beta_{22}$	$\beta_0+\beta_{G2}+\beta_{H1}+\beta_{21}$	$\beta_0+\beta_{G2}$
1	$\beta_0+\beta_{G1}+\beta_{H2}+\beta_{12}$	$\beta_0+\beta_{G1}+\beta_{H1}+\beta_{11}$	$\beta_0+\beta_{G1}$
0	$\beta_0+\beta_{H2}$	$\beta_0+\beta_{H1}$	β_0

Importance of SNP encoding scheme (Ankylosing Spondylitis; WTCCC2 - ~2000 cases + 5000 controls)



(extended from **Bessonov et al. 2016**)

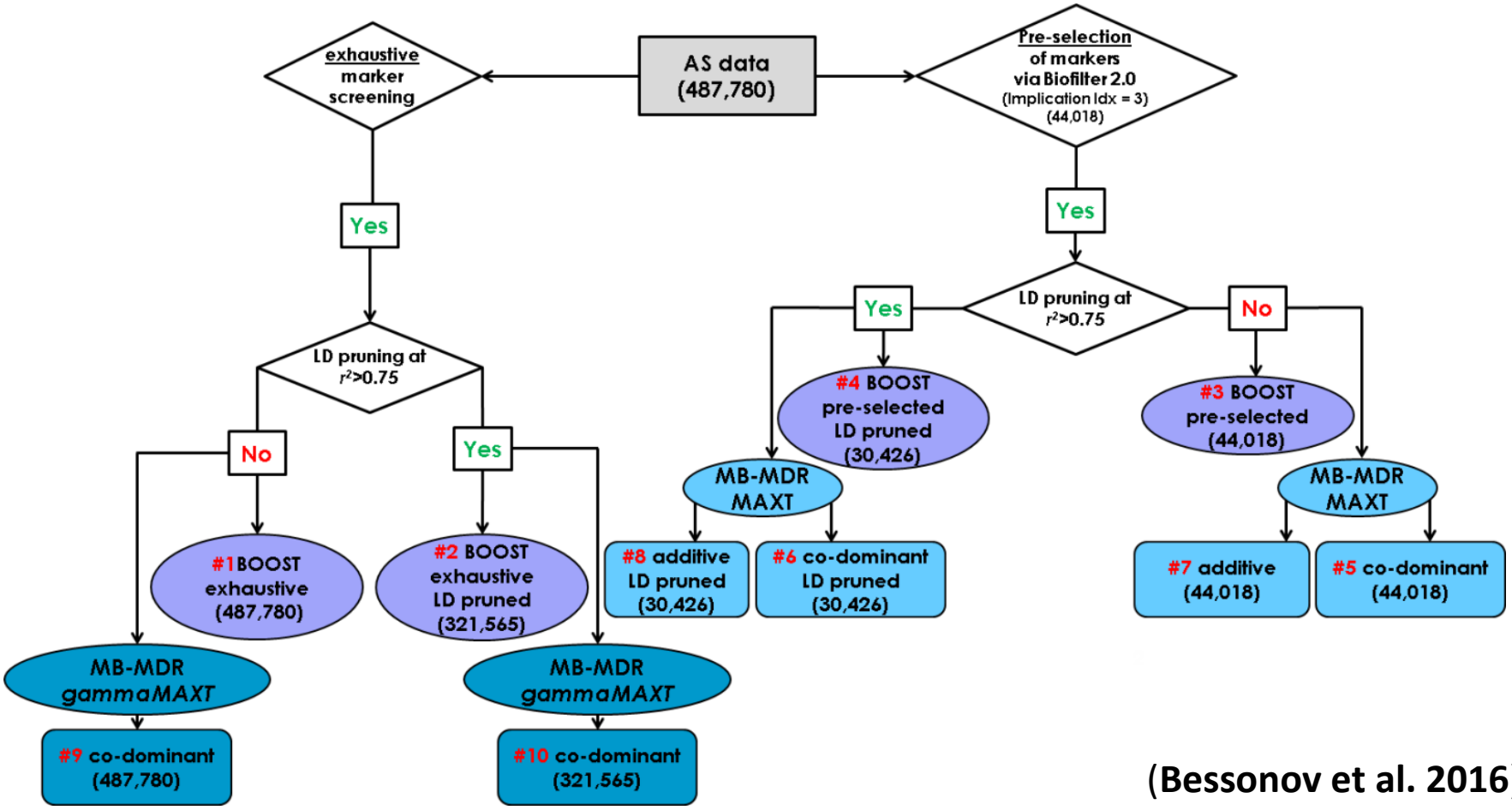


MOGPLOT

Manhattan plot (main effects) over GRAIL plot (interactions)

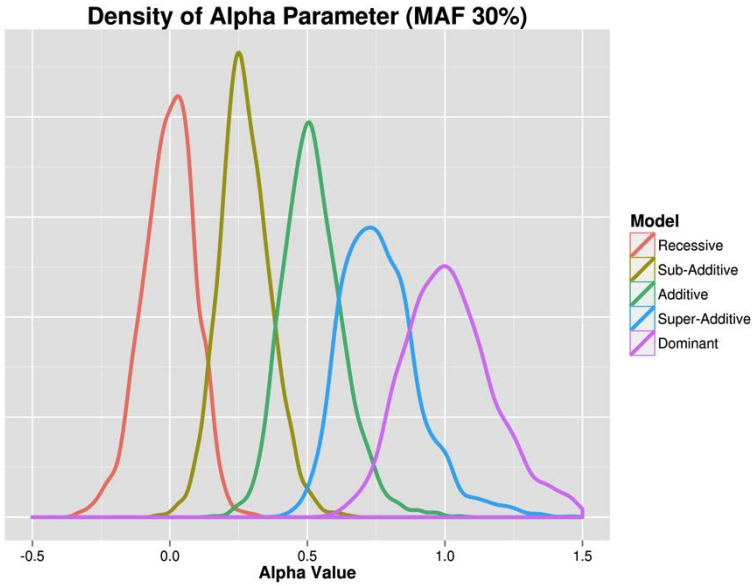
(**Van Lishout et al. 2016**)

Stability of results: analytic REPLICATION



(Bessonov et al. 2016)

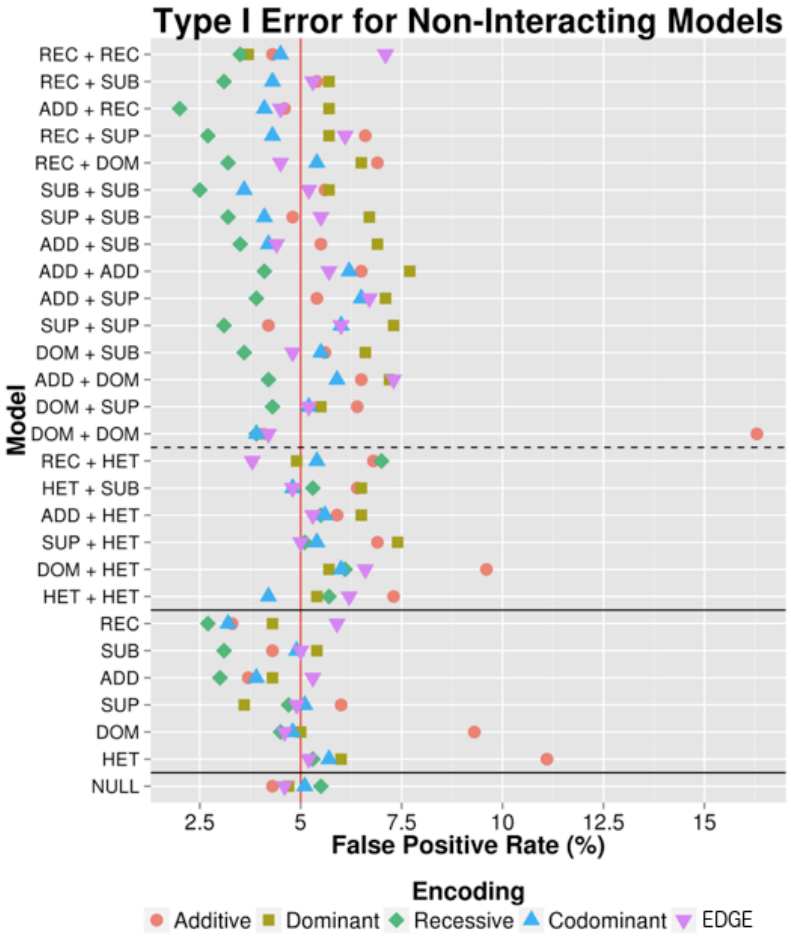
Importance of SNP encoding scheme (Hall et al. 2019 – submitted)



"Biological Action"	Homozygous Referent	Heterozygous	Homozygous Alternate
EDGE	0	**α**	1

$$Y \sim \beta_{Het}SNP_{Het} + \beta_{HA}SNP_{HA}$$

$$\alpha = \beta_{Het} / \beta_{HA}$$



Encoding (Hall et al. 2020 – under review)

Table 1. Examples of possible proportional genotype risk underlying genetic loci

Biological Action	Homozygous Referent (AA)	Heterozygous (Aa)	Homozygous Alternate (aa)
Recessive (REC)	0%	0%	100%
Sub-Additive (SUB	0%	25%	100%
Additive (ADD)	0%	50%	100%
Super-Additive (SUP)	0%	75%	100%
Dominant (DOM)	0%	100%	100%

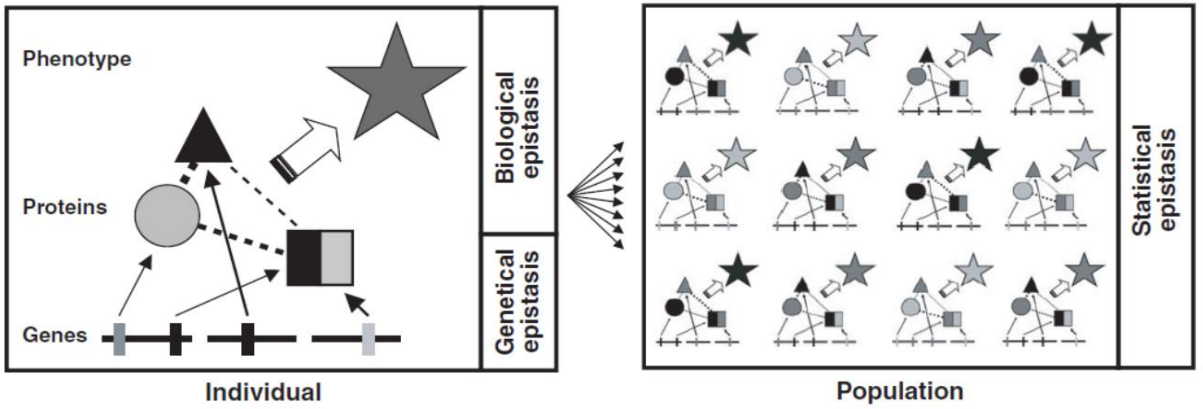
Lack of obvious correspondence between biology and statistics

- From the literature (~ interaction-specific vs two-locus hypotheses):
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387
  - …
  - Moore and Williams (2005) BioEssays 27:637–646
  - Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540
  - Wang, Elston and Zhu (2010) Hum Hered 70:269-277
  - …

  - **Van Steen et al. (2012) Brief Bioinform. 13(1):1-19**
  - **Aschard et al. (2012) Hum Genet 131(10):1591-1613**
  - **Gusareva and Van Steen (2014) Hum Genet 133(11):1343-58**

- In either case: statistical interactions DO imply JOINT involvement

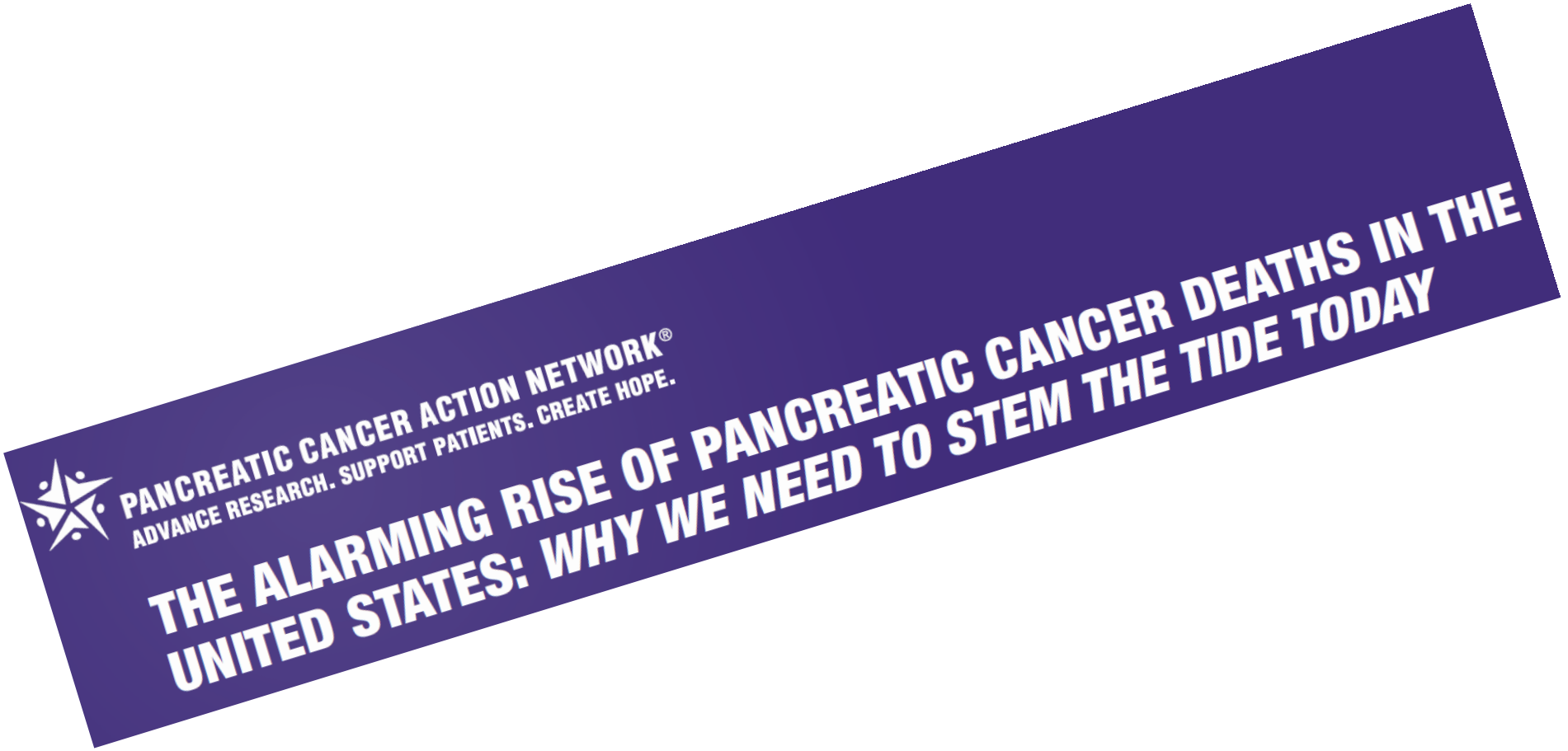
Lack of obvious correspondence between biology and statistics



(Moore 2005)

- **Epistasis: Grown into a more general theory and applications framework** for the analysis of interactions across and between -omics strata.

Disease context: complex "complex diseases"



Addressing complexity in “complex diseases” - pancreatic cancer

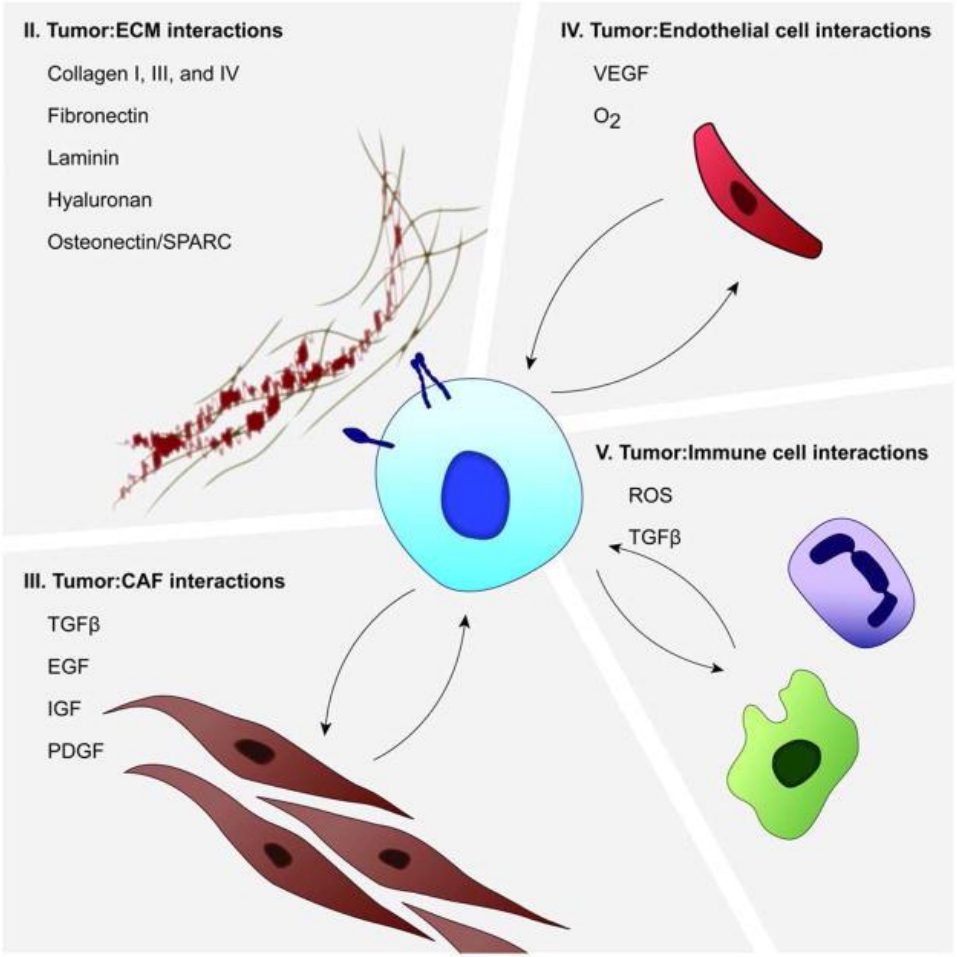
“Because effective systemic therapy capable of controlling the aggressive pancreatic cancer biology is currently lacking, the need for a better understanding of detailed mechanisms underlying pancreatic cancer development and progression is **URGENT**”

(Xie and Xie 2015)

Examples of interactions in pancreatic cancer

Tumor-stromal interactions

- Treatments focusing on pancreatic cancer cells alone have failed to significantly improve patient outcome over many decades
- Research efforts have now moved to understanding the pathophysiology of the stromal reaction and its role in cancer Progression (Whatcott et al. 2014)



Gene-environment interactions

(Jansen et al. 2015)

Formal definition of gene-environment interactions

- Also gene-environment interactions can be defined in a statistical or a biological way.
- A **biological gene-environment** interaction occurs when one or more genetic and one or more environmental factors participate in the same causal mechanism in the same individual (Yang and Khoury 1997; Rothman et al. 2008)
- As with gene-gene interactions, a **statistical gene-environment** interaction does not imply any inference about a specific biological mode of action. It is based on modeling a sample of individuals.

Formal definition of epistasis

- In practice, when modeling or testing, it may only be possible to detect **effect modification** from real-life data and not **interaction**, or interaction but not effect modification.
- Whereas an interaction effect for "exposures" X_1 and X_2 relies on a symmetric role for both X_1 and X_2 ,
  an effect modification relies on a conditioning argument (for instance on X_2) (VanderWeele 2009a)
- The distinction between both effect types is often concealed in regression analysis ... (Robins et al. 2000; North et al. 2005)

Comparison between gene–gene and gene–environment issues

- Conceptually many similar issues in terms of definition and mathematical modelling.
- In practice, some clear differences emerge.
- For G x E:
  - We generally have to decide which environments to measure / test; these are typically only a few (often < 100)
  - Measurement error (lifestyle) and unknown confounding
  - Risk estimation, important for screening strategies and public health interventions

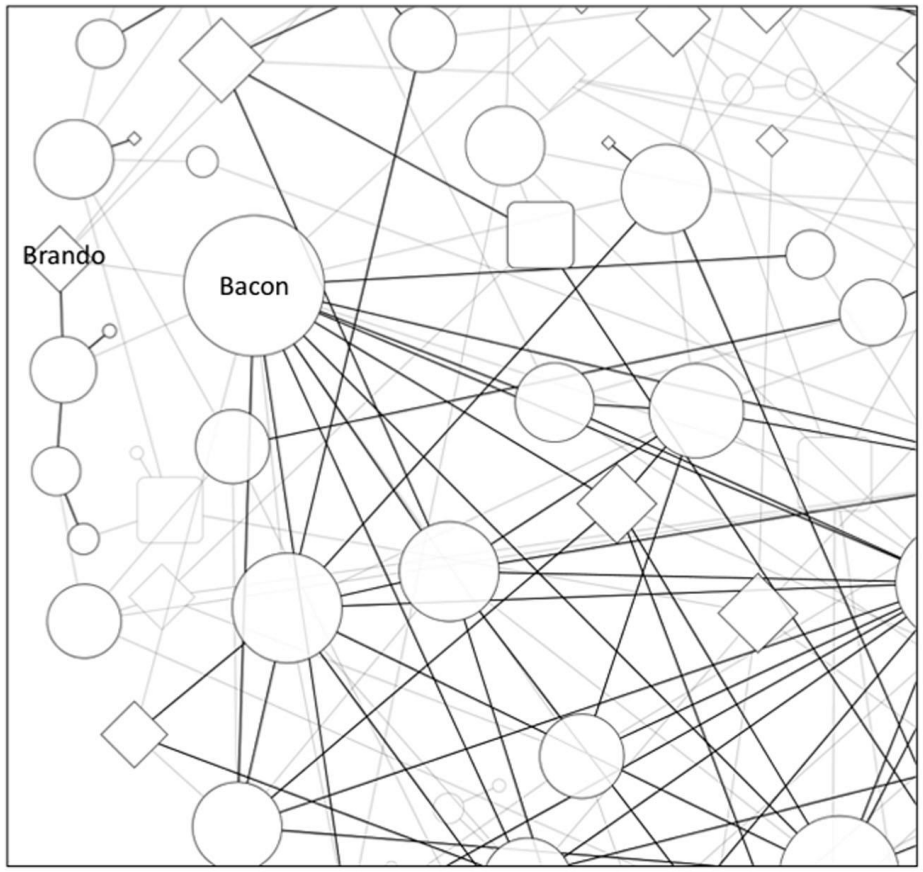
(courtesy slide EUPancreas WG2 Training School, Antwerp, 2016)

Comparison between gene–gene and gene–environment issues

- For G x G
  - Assuming we have GWAS data, we have already measured the genetic factors of interest
  - Adequate error rates (except for newer sequencing technologies)
  - (Hundred) thousands of variants
  - Higher-order interactions may reflect the complex biological wiring of complex diseases (whereas G x E often restricts attention to pairwise interactions)

(courtesy slide EUPancreas WG2 Training School, Antwerp, 2016)

Looking for higher-order interactions



Edges represent small gene–gene interactions between SNPs.

Gray nodes and edges have weaker interactions.
Circle nodes represent SNPs that do not have a significant main effect.
The diamond nodes represent significant main effect association.

The size of the node is proportional to the number of connections.

(McKinney et al 2012)

Some references

Published in final edited form as:
Hum Genet. 2012 October ; 131(10): 1591–1613. doi:10.1007/s00439-012-1192-0.

Challenges and Opportunities in Genome-Wide Environmental Interaction (GWEI) studies

Hugues Aschard[1], **Sharon Lutz**[2,*], **Bärbel Maus**[3,4,*], **Eric J. Duell**[5], **Tasha Fingerlin**[2], **Nilanjan Chatterjee**[6], **Peter Kraft**[1,7], and **Kristel Van Steen**[3,4]

Hum Genet (2014) 133:1343–1358
DOI 10.1007/s00439-014-1480-y

REVIEW PAPER

Practical aspects of genome-wide association interaction analysis

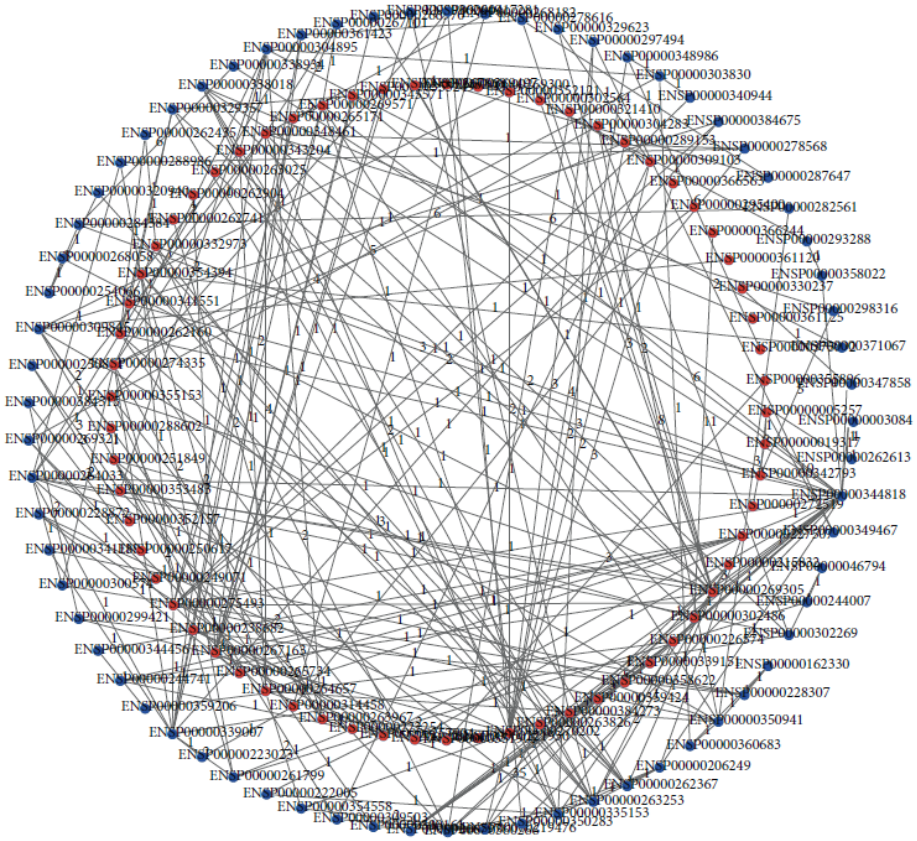
Elena S. Gusareva · Kristel Van Steen

Protein-protein interactions

(Yuan et al. 2015)

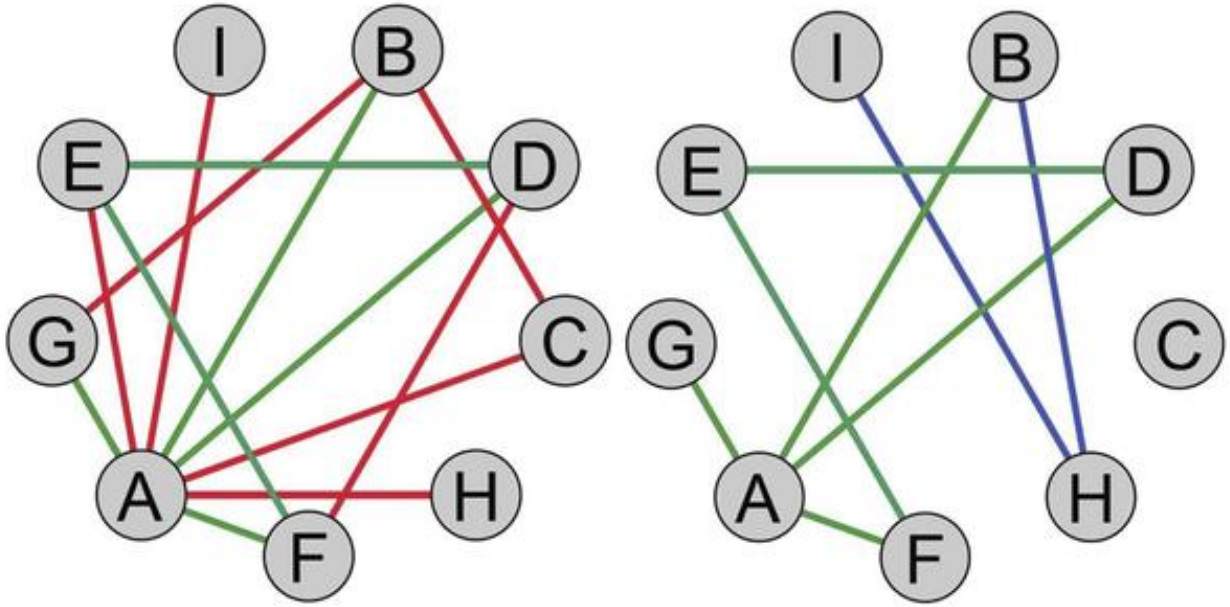
A graph consisting of 2,080 shortest paths:

- The nodes on the inner circle (red nodes) represent 65 PC-related genes.
- The nodes on the outer circle (blue nodes) represent 69 shortest path genes.
- The numbers on the edges represent the weights of the edges.



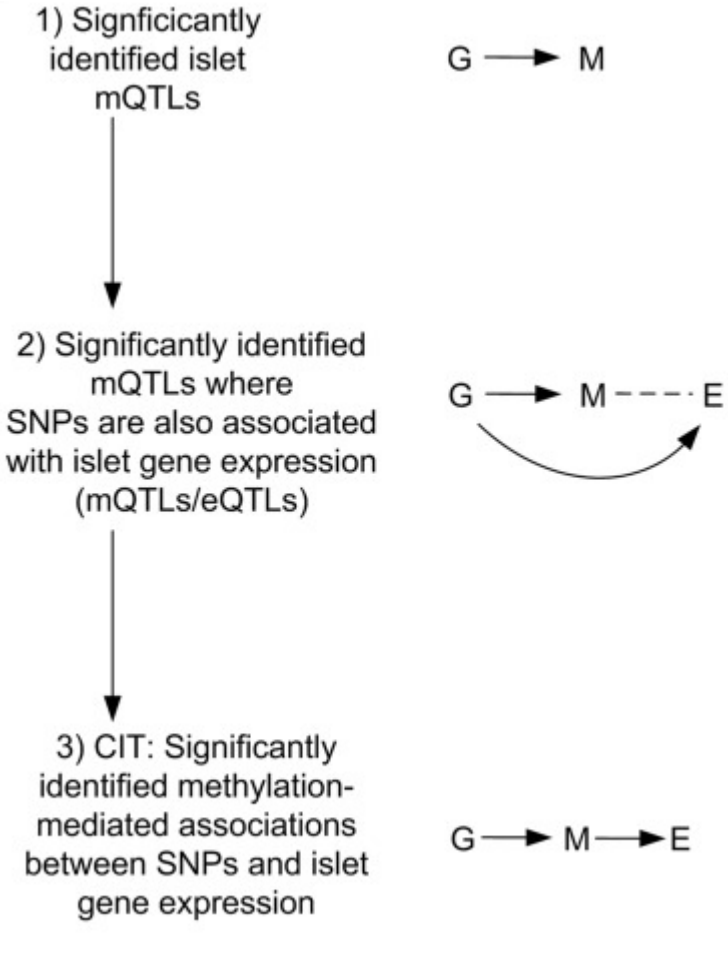
Gene-coexpression networks

(Anglani et al. 2014)



- Healthy condition on the left and disease-affected tissue on the right. Green links remain unchanged in the two phenotypes
- Red connections are loss from healthy to cancer network
- Blue edges are novel connections in the cancer tissue

Genetic-epigenetic mechanistic interactions (pancreatic islets)



cis-mQTLs:	*trans-mQTLs*:
67,438 SNP-CpG pairs (36,783 unique SNPs and 11,735 unique CpGs)	2,562 SNP-CpG pairs (1,465 unique SNPs and 383 unique CpGs)

cis-mQTLs/eQTLs:	*trans-mQTLs/eQTLs*:
2,805 SNP-CpG-mRNA combinations (243 unique SNPs, 276 unique CpGs and 46 unique mRNA transcripts)	37 SNP-CpG-mRNA combinations (22 unique SNPs, 12 unique CpGs and 8 unique mRNA transcripts)

cis-mQTLs/eQTLs:	*trans-mQTLs/eQTLs*:
28 SNP-CpG-mRNA combinations (17 unique SNPs, 14 unique CpGs and 5 unique mRNA transcripts)	4 SNP-CpG-mRNA combinations (2 unique SNPs, 4 unique CpGs and 1 unique mRNA transcript)

Gene-gene interactions using SNPs?

(Olsson et al. 2014)

GWAS Catalogue – "Pancreas Cancer"

Wolpin BM (PMID: 25086665)	2014-08-03	Nat Genet	Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer.	Pancreatic cancer	10

Initial sample description	1,582 European ancestry cases, 5,203 European ancestry controls
Initial ancestry (country of recruitment)	6785 European (U.S., Australia, France, Germany, Netherlands, Denmark, Finland, Norway, Sweden, U.K., Greece, Italy, Spain)
Replication sample description	6,101 European ancestry cases, 9,194 European ancestry controls
Replication ancestry (country of recruitment)	15295 European (Canada, U.S., France, Germany, Netherlands, Denmark, Finland, Norway, Sweden, U.K., Greece, Italy, Spain)
Platform [SNPs passing QC]	Illumina [608202]

(http://www.ebi.ac.uk/gwas/search?query=pancreas%20cancer#study)

Getting the phenotype straight …

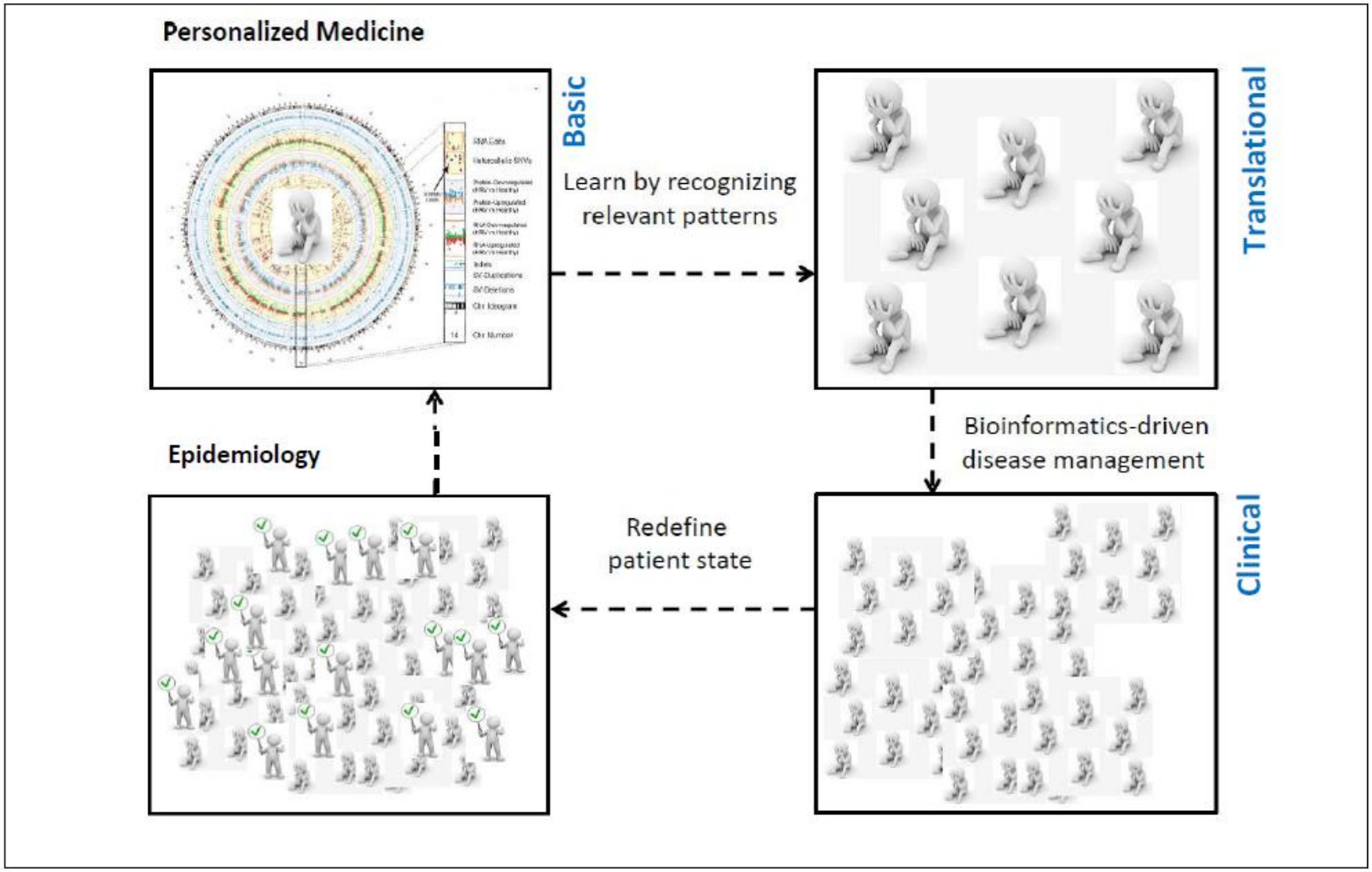
REVIEWS

Molecular subtypes of pancreatic cancer

*Eric A. Collisson[1], Peter Bailey[2], David K. Chang[2,3] and Andrew V. Biankin[2,3,4]**

Abstract | Cancers that appear morphologically similar often have dramatically different clinical features, respond variably to therapy and have a range of outcomes. Compelling evidence now demonstrates that differences in the molecular pathology of otherwise indistinguishable cancers substantially impact the clinical characteristics of the disease. Molecular subtypes now guide preclinical and clinical therapeutic development and treatment in many cancer types. The ability to predict optimal therapeutic strategies ahead of treatment improves overall patient outcomes, minimizing treatment-related morbidity and cost. Although clinical decision making based on histopathological criteria underpinned by robust data is well established in many cancer types, subtypes of pancreatic cancer do not currently inform treatment decisions. However, accumulating molecular data are defining subgroups in pancreatic cancer with distinct biology and potential subtype-specific therapeutic vulnerabilities, providing the opportunity to define a de novo clinically applicable molecular taxonomy. This Review summarizes current knowledge concerning the molecular subtyping of pancreatic cancer and explores future strategies for using a molecular taxonomy to guide therapeutic development and ultimately routine therapy with the overall goal of improving outcomes for this disease.

(https://www.nature.com/articles/s41575-019-0109-y.pdf)


Personalized Medicine
Basic
Learn by recognizing relevant patterns
Translational
Bioinformatics-driven disease management
Clinical
Redefine patient state
Epidemiology

Questions?

Main supporting doc to this class (complementing course slides)

√

Review

A brief history of human disease genetics

https://doi.org/10.1038/s41586-019-1879-7
Received: 16 July 2019
Accepted: 13 November 2019
Published online: 8 January 2020

Melina Claussnitzer[1,2,3], Judy H. Cho[4,5,6], Rory Collins[7,8], Nancy J. Cox[9], Emmanouil T. Dermitzakis[10,11], Matthew E. Hurles[12], Sekar Kathiresan[2,13,14], Eimear E. Kenny[4,6,15], Cecilia M. Lindgren[2,16,17], Daniel G. MacArthur[2,13,18], Kathryn N. North[19,20], Sharon E. Plon[21,22], Heidi L. Rehm[2,13,18,23], Neil Risch[24], Charles N. Rotimi[25], Jay Shendure[26,27,28], Nicole Soranzo[12,29] & Mark I. McCarthy[17,30,31,32]*

√

A tutorial on statistical methods for population association studies

David J. Balding

Abstract | Although genetic association studies have been with us for many years, even for the simplest analyses there is little consensus on the most appropriate statistical procedures. Here I give an overview of statistical approaches to population association studies, including preliminary analyses (Hardy–Weinberg equilibrium testing, inference of phase and missing data, and SNP tagging), and single-SNP and multipoint tests for association. My goal is to outline the key methods with a brief discussion of problems (population structure and multiple testing), avenues for solutions and some ongoing developments.

Nature reviews Genetics 2006; 5:63-70 – for those interested in technical (statistical) details

Homework assignment

+

Check out the document

“Critical evaluation of a paper/report”