

Genome-wide Association Studies

Kristel Van Steen, PhD²

Systems Genomics – GIGA-R, ULg
Systems Medicine – Human Genetics, KUL

kristel.vansteen@uliege.be

Genome-wide Association Studies

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 GWAs

3 Study Design Elements

3.a Marker level

3.b Subject level

3.c Gender level (not considered in this course)

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 / Regression

5.b Causation

6 Post Association Analysis Steps

6.a Replication and validation

6.b GWA Interpretation and follow-up

1 Setting the pace

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

The use of saliva

- People spit for a variety of reasons. We've all employed the technique to remove a hair or some other distasteful object from our mouths. People who chew tobacco do it for obvious reasons. Ball players do it because they're nervous, bored or looking to showcase their masculinity. And people in many different cultures spit on their enemies to show disdain.
- Thanks to a phenomenon known as **direct-to-consumer genetic testing** or **at-home genetic testing**, people are spitting today for a much more productive (and perhaps more sophisticated) reason -- to get a glimpse of their own DNA.

(science.howstuffworks.com)

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



Commercial kits



Do not eat, drink, smoke, chew gum, brush your teeth, or use mouthwash for at least 30 minutes prior to providing your sample.




Collect the recommended volume of saliva. The recommended volume of saliva to provide is 2 mL, or about $\frac{1}{2}$ teaspoon. Your saliva sample should be just above the fill line.



Provide your sample and add the stabilization buffer within 30 minutes. The full saliva sample should be collected within 30 minutes and the funnel contents should be released into the tube immediately. Waiting longer than 30 minutes may decrease the yield and quality of your DNA.




Cap securely before shipping. Remember to remove and discard the funnel lid and place the tube cap on securely before mailing your sample to our laboratory.



All from home. No blood. No needles. Just a small saliva sample.

[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.



Your estimated lifetime risk

[Print this page](#)

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%	
Multiple sclerosis You: 0.28% Avg: 0.27%	NEW Deep vein thrombosis You: 2.9% Avg: 3.6%			

Key to your results

Condition name

Your results
You: 45%

Population Average
Avg: 30%

[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 (850) 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>)

ETHNIC AND RACIAL STUDIES, 2016
VOL. 39, NO. 2, 142–161
<http://dx.doi.org/10.1080/01419870.2016.1105990>



In the blood: the myth and reality of genetic markers of identity

Mark A. Jobling^a, Rita Rasteiro^{a,b} and Jon H. Wetton^{a,b}

^aDepartment of Genetics, University of Leicester, Leicester, UK; ^bSchool of History, University of Leicester, Leicester, UK

ABSTRACT

The differences between copies of the human genome are very small, but tend to cluster in different populations. So, despite the fact that low inter-population differentiation does not support a biological definition of races statistical methods are nonetheless claimed to be able to predict successfully the population of origin of a DNA sample. Such methods are employed in commercial genetic ancestry tests, and particular genetic signatures, often in the male-specific Y-chromosome or maternally-inherited mitochondrial DNA, have become widely identified with particular ancestral or existing groups, such as Vikings, Jews, or Zulus. Here, we provide a primer on genetics, and describe how genetic markers have become associated with particular groups. We describe the conflict between population genetics and individual-based genetics and the pitfalls of over-simplistic genetic interpretations, arguing that although the tests themselves are reliable, the interpretations are unreliable and strongly influenced by cultural and other social forces.



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

NewStatesman



SCIENCE & TECH 15 JANUARY 2015

23andMe: Why bother with predictions about yourself when you are almost certainly average?

Want to understand your genes? Call your parents.

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>)

From “risk prediction” to “my origin” to “SNP-based genetic tests”

- As we will see, we can measure (genetic) **variation between individuals** at several positions on the genome, using so-called **molecular markers** such as **Single Nucleotide Polymorphisms (SNPs)**
- To run a SNP test, scientists can embed a subject's DNA into for instance a small silicon chip containing *reference DNA* from both healthy individuals and individuals with certain diseases.
- By analyzing how the SNPs from the subject's DNA match up with SNPs from the reference DNA, the scientists can determine if the subject might be predisposed to certain diseases or disorders.

Talking about “references”: reference genomes

- A reference genome (also known as a reference assembly) is a digital nucleic acid sequence database, assembled by scientists as a representative example of a species' set of genes.
- As they are often assembled from the sequencing of DNA from a number of donors, reference genomes do not accurately represent the set of genes of any single person. Instead a reference provides a haploid mosaic of different DNA sequences from each donor.
- For example GRCh37, the Genome Reference Consortium human genome (build 37) is derived from thirteen anonymous volunteers from Buffalo, New York



"Wellcome genome bookcase" by Russ London at en.wikipedia.
Licensed under CC BY-SA 3.0 via Commons -
https://commons.wikimedia.org/wiki/File:Wellcome_genome_bookcase.png#/media/File:Wellcome_genome_bookcase.png

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.

Can you handle the truth?

Identifying Genetic Markers ©2009 HowStuffWorks

Service Provider:	23andMe	deCODEme	Navigenics
Arthritis	✱	✱	✱
Asthma	✱	✱	
Bipolar/Depression	✱		
Cardiovascular Disease	✱	✱	✱
Multiple Sclerosis	✱	✱	✱
Osteoporosis	✱		
Parkinson's Disease			
Schizophrenia	✱		
Thrombosis	✱	✱	
Type 1/2 Diabetes	✱	✱	✱



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

Reference genomes based on >> 13 individuals



RESEARCH ARTICLE

Comparison of HapMap and 1000 Genomes Reference Panels in a Large-Scale Genome-Wide Association Study

Paul S. de Vries^{1,2}, Maria Sabater-Lleal³, Daniel I. Chasman^{4,5}, Stella Trompet^{6,7}, Tarunveer S. Ahluwalia^{8,9}, Alexander Teumer¹⁰, Marcus E. Kleber¹¹, Ming-Huei Chen^{12,13}, Jie Jin Wang¹⁴, John R. Attia^{15,16}, Riccardo E. Marioni^{17,18,19}, Maristella Steri²⁰, Lu-Chen Weng²¹, Rene Pool^{22,23}, Vera Grossmann²⁴, Jennifer A. Brody²⁵, Cristina Venturini^{26,27}, Toshiko Tanaka²⁸, Lynda M. Rose⁴, Christopher Oldmeadow^{15,16}, Johanna Mazur²⁹, Saonli Basu³⁰, Mattias Frånberg^{3,31}, Qiong Yang^{13,32}, Symen Ligthart¹, Jouke J. Hottenga²², Ann Rumley³³, Antonella Mulas²⁰, Anton J. M. de Craen⁷, Anne Grotevendt³⁴, Kent D. Taylor^{35,36}, Graciela E. Delgado¹¹, Annette Kifley¹⁴, Lorna M. Lopez^{17,37,38}, Tina L. Berentzen³⁹, Massimo Mangino^{27,40}, Stefania Bandinelli⁴¹, Alanna C. Morrison¹, Anders Hamsten³, Geoffrey Tofler⁴², Moniek P. M. de Maat⁴³, Harmen H. M. Draisma^{22,44}, Gordon D. Lowe⁴⁵, Magdalena Zoledziewska²⁰, Naveed Sattar⁴⁶, Karl J. Lackner⁴⁷, Uwe Völker⁴⁸, Barbara McKnight⁴⁹, Jie Huang⁵⁰, Elizabeth G. Holliday⁵¹, Mark A. McEvoy¹⁶, John M. Starr^{17,52}, Pirro G. Hysi²⁷, Dena G. Hernandez⁵³, Weihua Guan³⁰, Fernando Rivadeneira^{1,54}, Wendy L. McArdle⁵⁵, P. Eline Slagboom⁵⁶, Tanja Zeller^{57,58}, Bruce M. Psaty^{59,60}, André G. Uitterlinden^{1,54}, Eco J. C. de Geus^{22,23}, David J. Stott⁶¹, Harald Binder⁶², Albert Hofman^{1,63}, Oscar H. Franco¹, Jerome I. Rotter^{64,65}, Luigi Ferrucci²⁸, Tim D. Spector²⁷, Ian J. Deary^{17,66}, Winfried März^{11,67,68}, Andreas Greinacher⁶⁹, Philipp S. Wild^{70,71,72}, Francesco Cucca²⁰, Dorret I. Boomsma²², Hugh Watkins⁷³, Weihong Tang²¹, Paul M. Ridker^{4,5}, Jan W. Jukema^{6,74,75}, Rodney J. Scott^{76,77}, Paul Mitchell¹⁴, Torben Hansen⁷⁸, Christopher J. O'Donnell^{13,79}, Nicholas L. Smith^{60,80,81}, David P. Strachan⁸², Abbas Dehghan^{1,83*}



OPEN ACCESS

Citation: de Vries PS, Sabater-Lleal M, Chasman DI, Trompet S, Ahluwalia TS, Teumer A, et al. (2017) Comparison of HapMap and 1000 Genomes Reference Panels in a Large-Scale Genome-Wide Association Study. PLoS ONE 12 (1): e0167742. doi:10.1371/journal.pone.0167742

1.b Speaking the 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.

What is genetic epidemiology?

Epidemiology is usually defined as “the study of the distribution, determinants [and control] of health-related states and events in populations”.¹ By contrast, genetic epidemiology means different things to different people.²⁻⁷ We regard it as a discipline closely allied to traditional epidemiology that focuses on the familial, and in particular genetic, determinants of disease and the joint effects of genes and non-genetic determinants. Crucially, appropriate account is taken of the biology that underlies the action of genes and the

close. The marker and the causative variant need not be within the same gene. This principle is the basis of genetic linkage analysis (see a later paper in this series¹²), which has achieved many of the breakthroughs in the genetics of disease causation. Many such breakthroughs involve conditions caused by variants in a single gene and have been achieved by geneticists and clinical geneticists who would not view themselves as genetic epidemiologists. Nevertheless, linkage analysis is one of the most important tools available to the genetic epidemiologist.

Lancet 2005; 366: 941–51

See Comment page 880

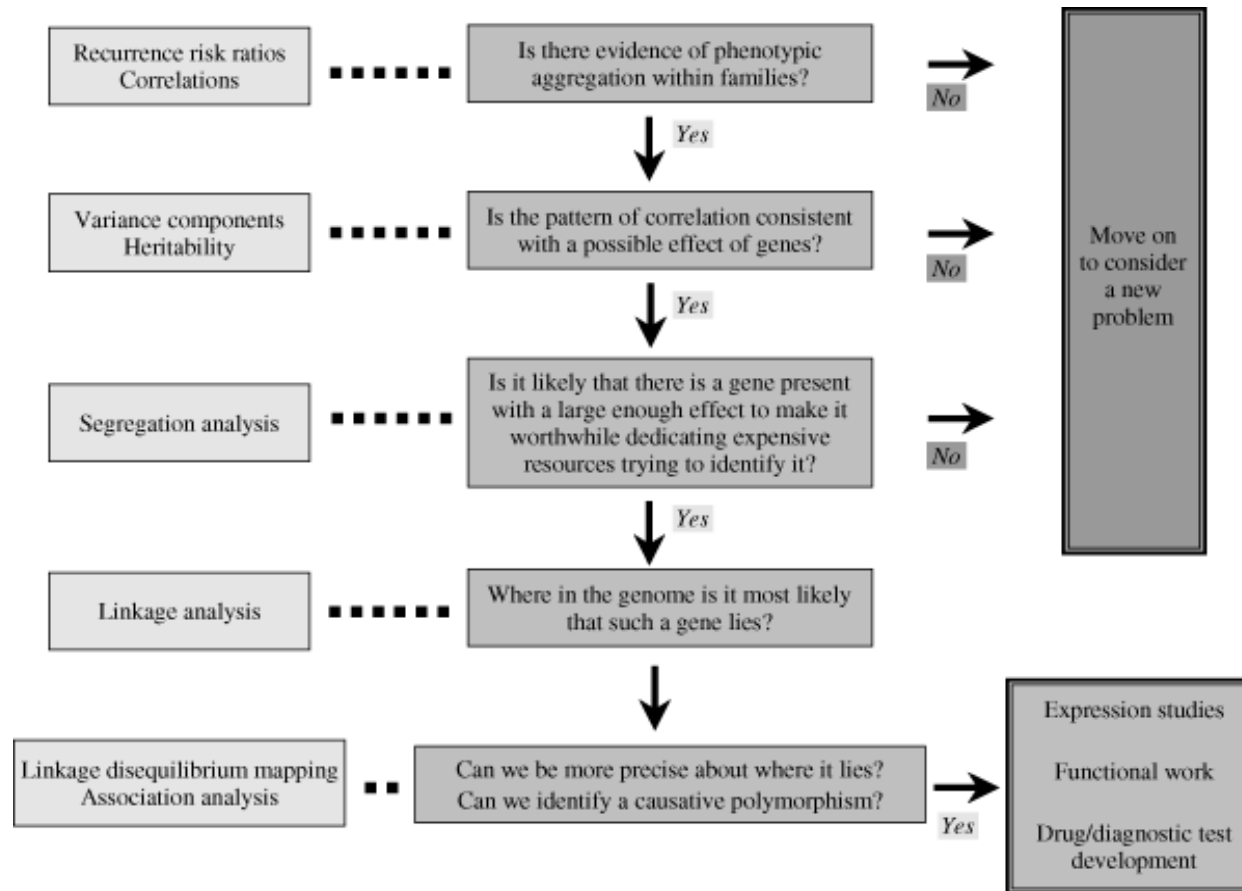
This is the first in a Series of seven papers on genetic epidemiology.

Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, UK

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

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

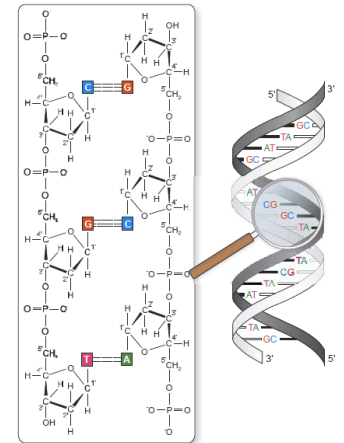
What are relevant questions in genetic epidemiology?



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

Where is the genetic information located?

- Cell has nucleus
- Nucleus carries genetic information in chromosomes
- Chromosomes 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)

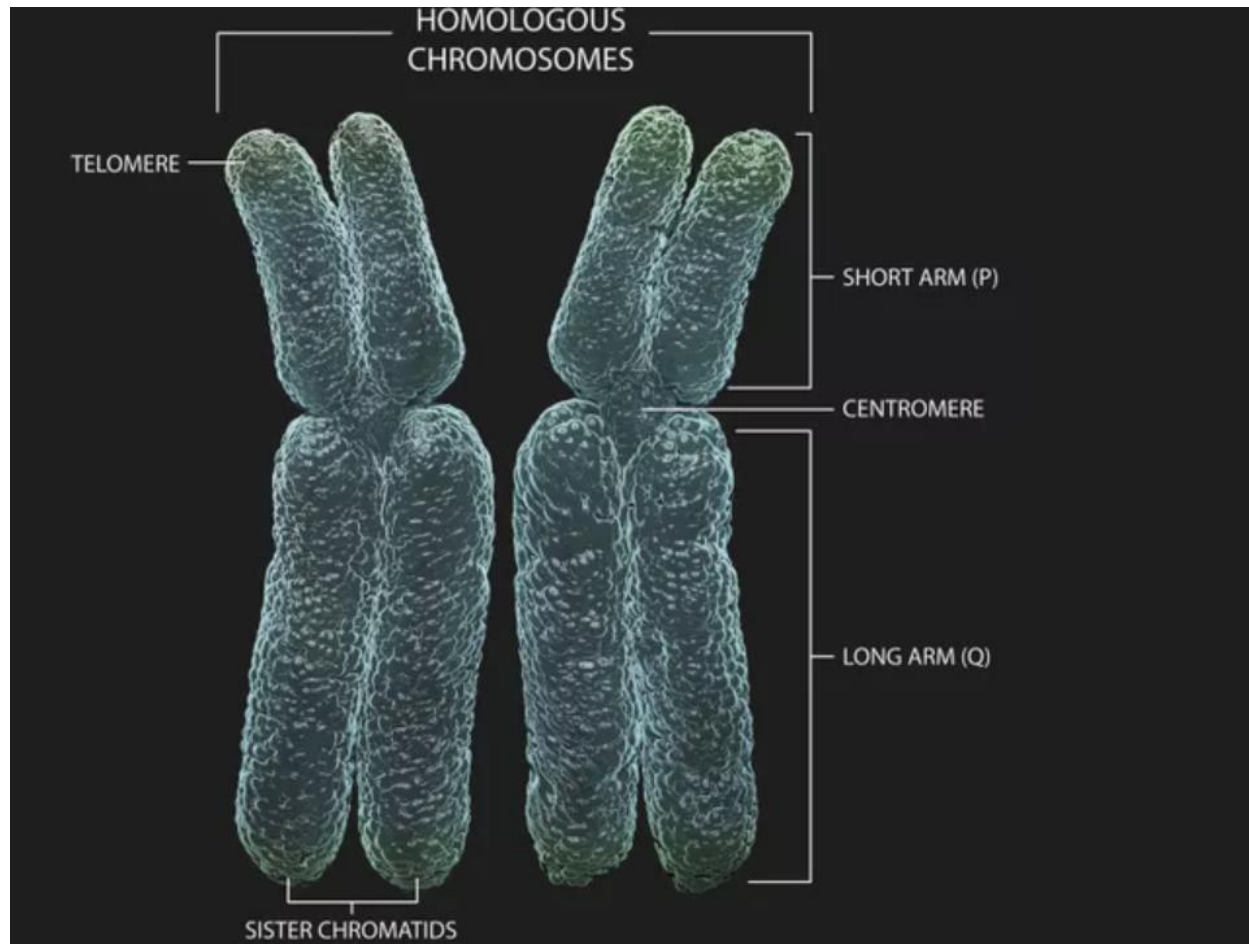


(Ziegler and Van Steen, Brazil 2010)

Where is the genetic information located?

- Chromosomes are
 - Linear arrangements of DNA
 - 22 autosomal pairs in humans
 - 2 sex chromosomes (X and Y)
- Pair of chromosomes called homologs
- Meiosis: special type of cell division
- Crossover: chromosomal segment exchange between homologs during meiosis
- Average # crossovers: $55 \times$ in males, $1.5 \times$ higher in females

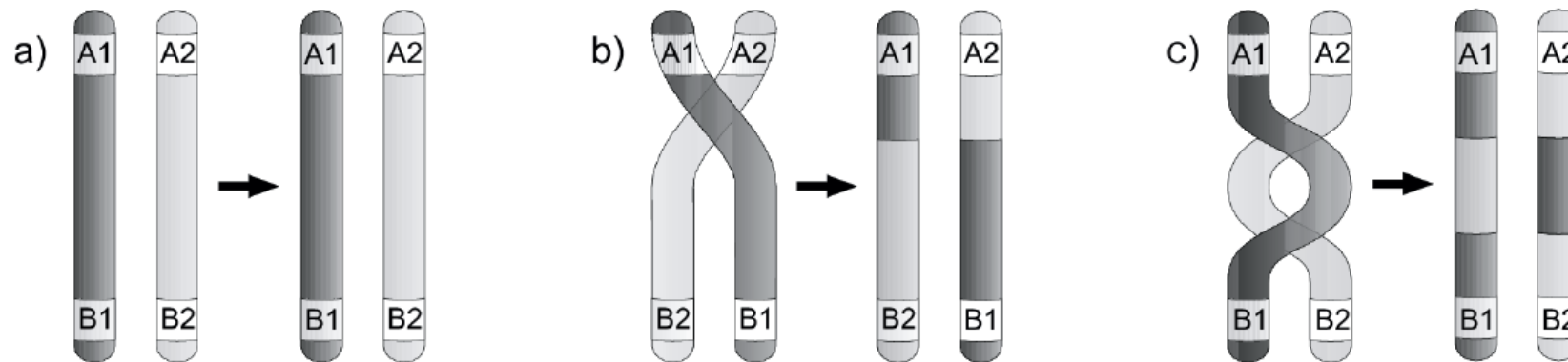
(Ziegler and Van Steen, Brazil 2010)



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

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)

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 $\approx 10^{-4}$ to $10^{-6} \Rightarrow$ 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

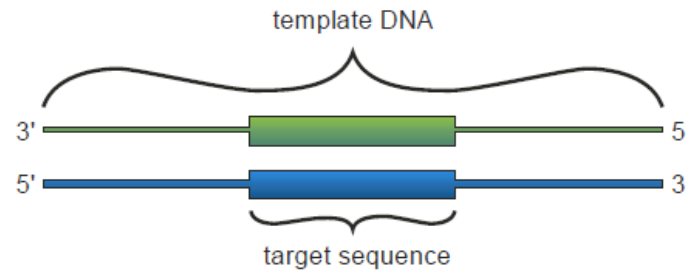


<http://www.dorak.info/epi/glosge.html>

How can individual differences be detected?

Mimic PCR

1st cycle:



1st step: Denaturation



2nd step: Hybridization



3rd step: Elongation



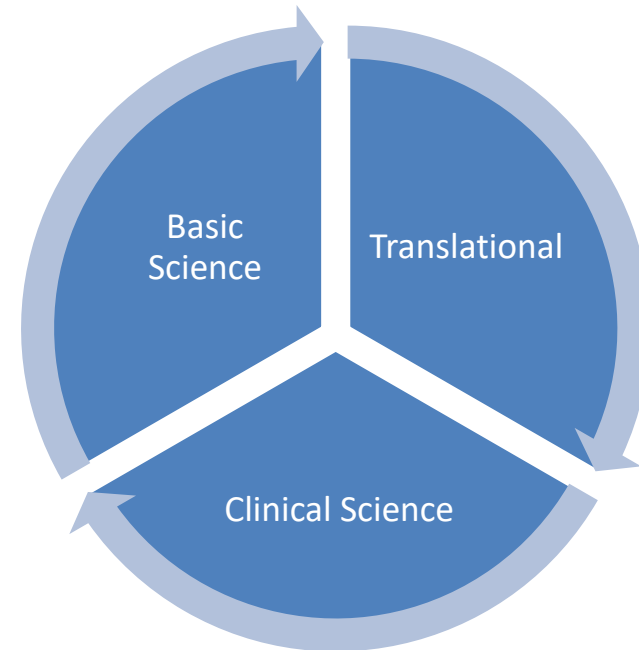
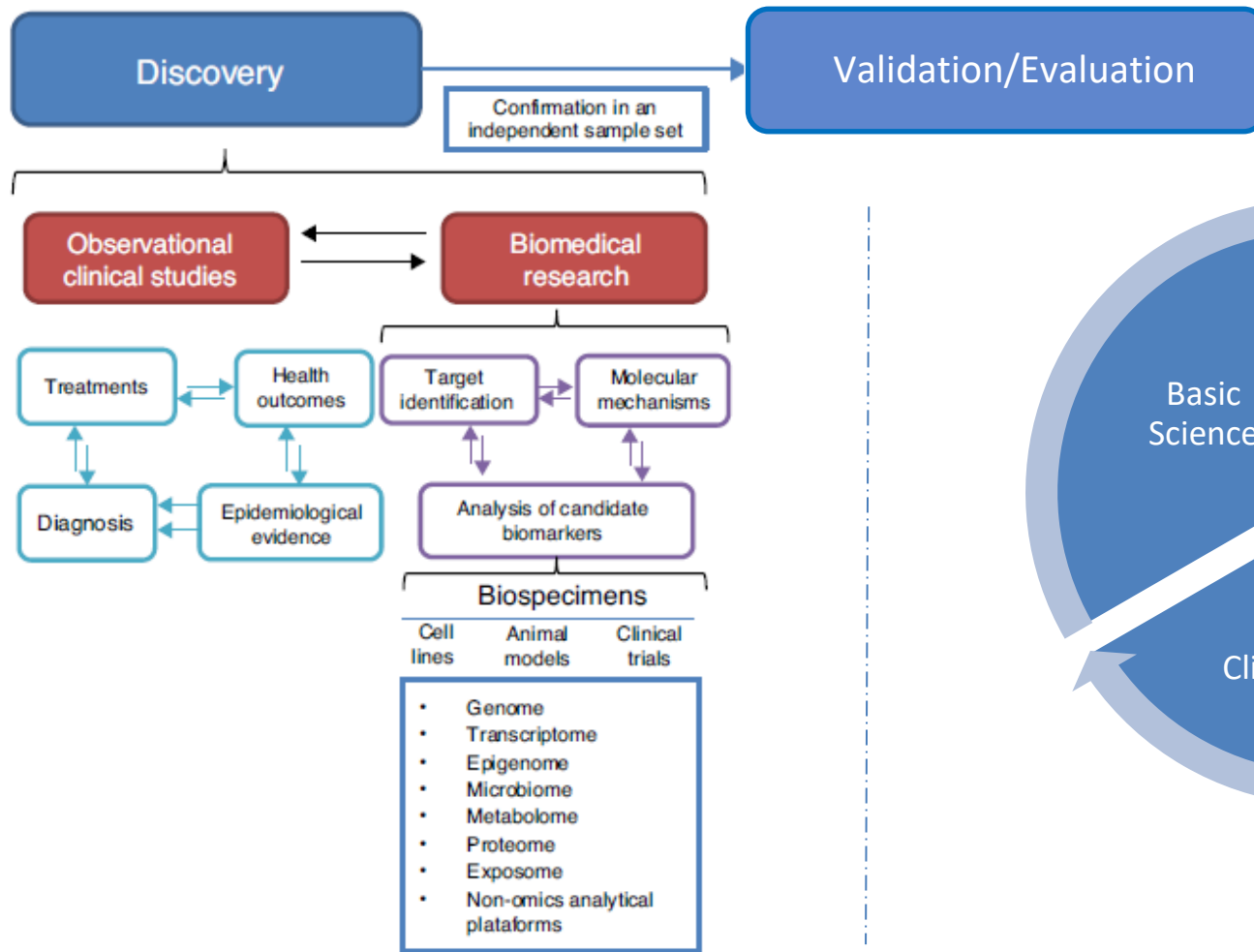
(Ziegler and Van Steen, Brazil 2010)

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?

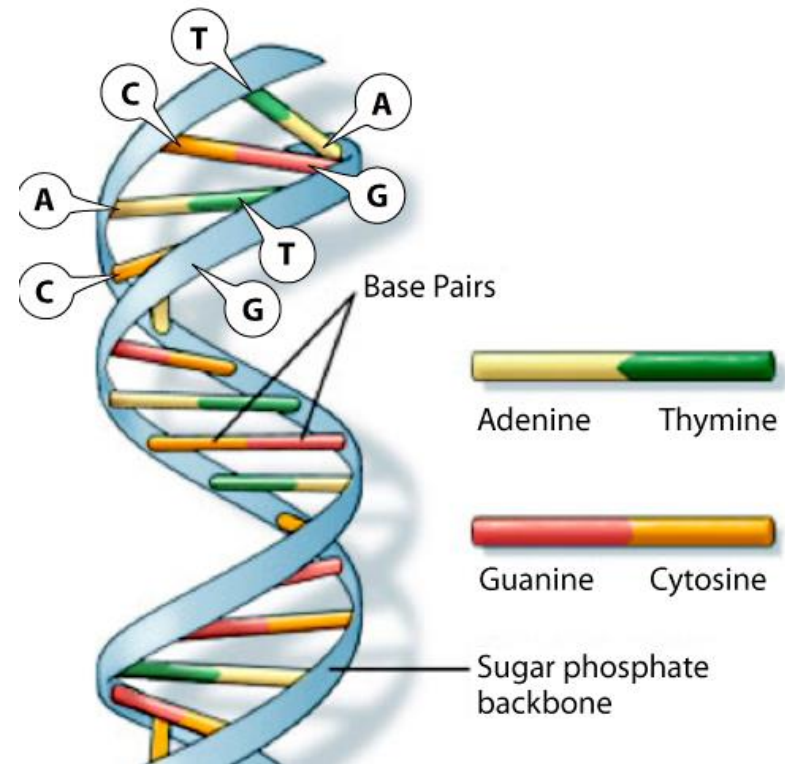
Single Nucleotide Polymorphisms (SNPs)

- 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 $\geq 1\%$
 - 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?

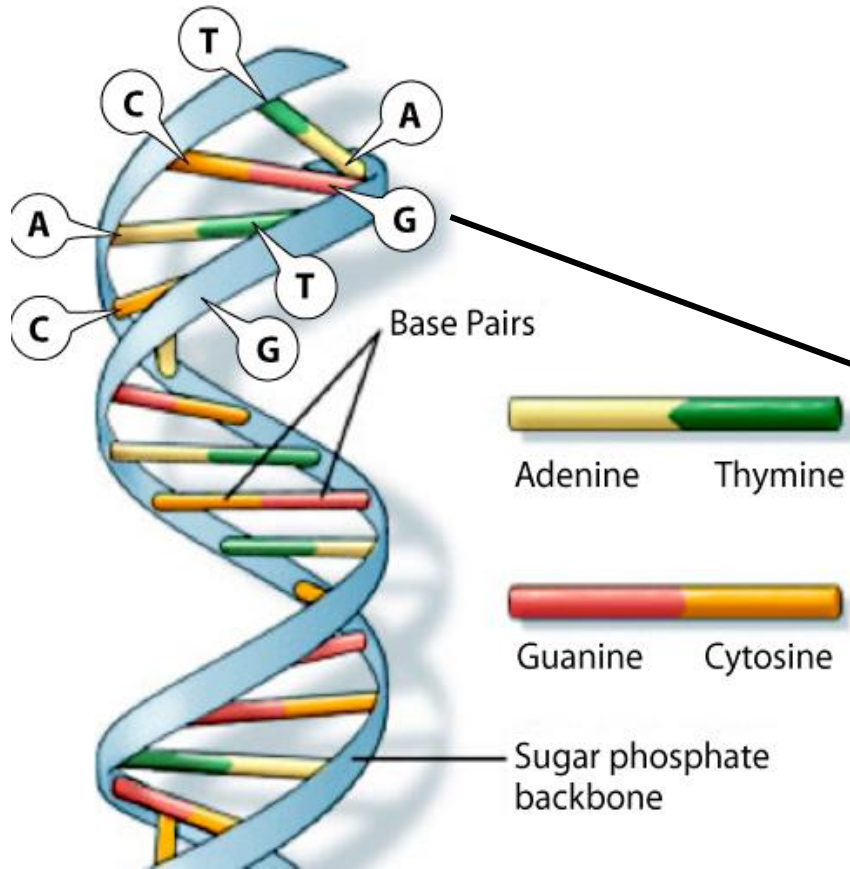
- Any two people plucked at random off the street are on average 99.9 percent the same, DNA-wise (> 3 million positional differences)
- Most genome variations are relatively small and simple, involving only a few bases—an A substituted for a T here, a G left out there, a short sequence such as CG added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



(U.S. National Library of Medicine)

Common genetic variations

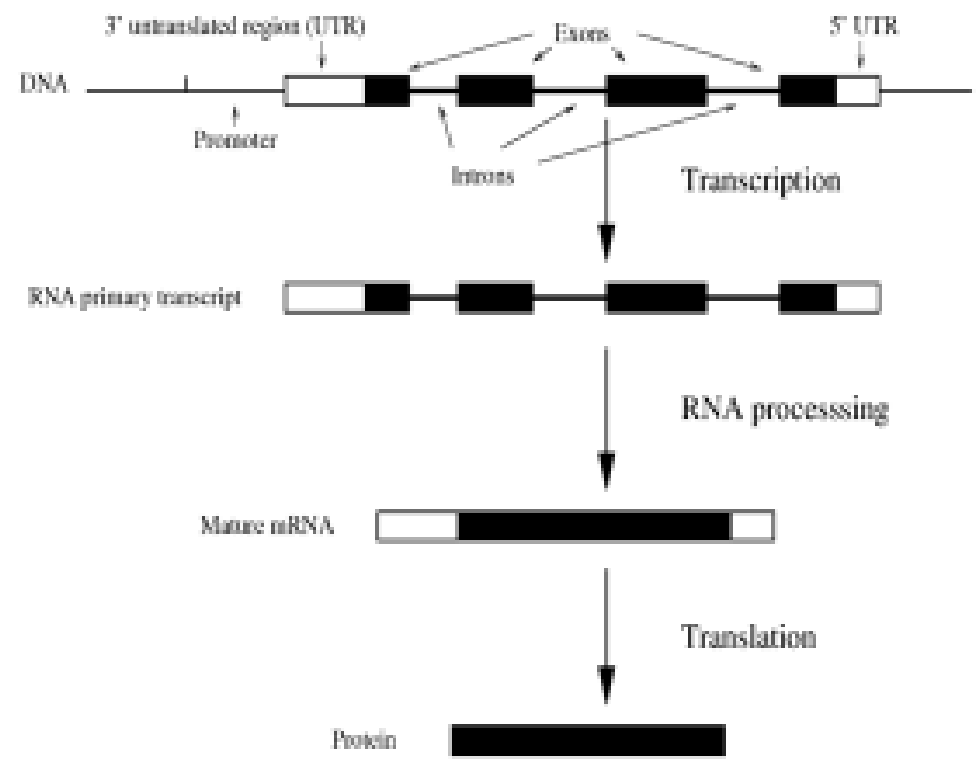
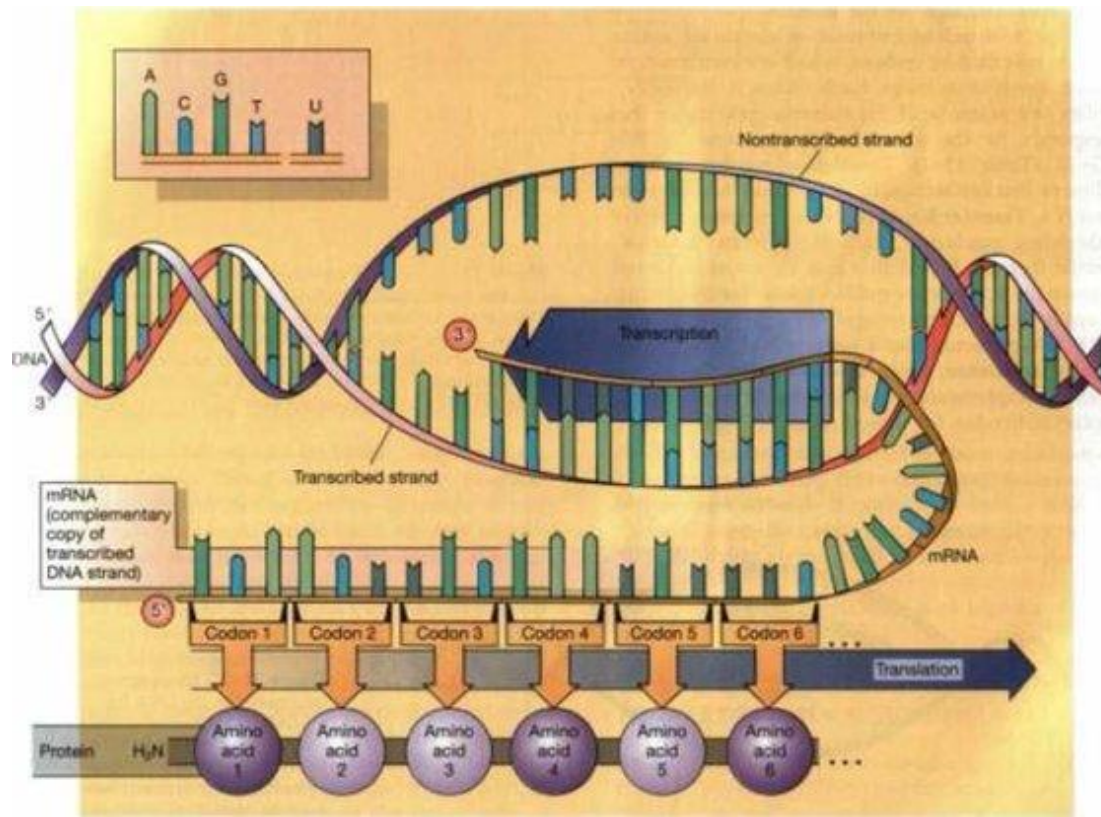
MAF (minor allele frequency)



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

What is the central dogma of molecular biology?

Simplified version



Information is everywhere: the programming of life

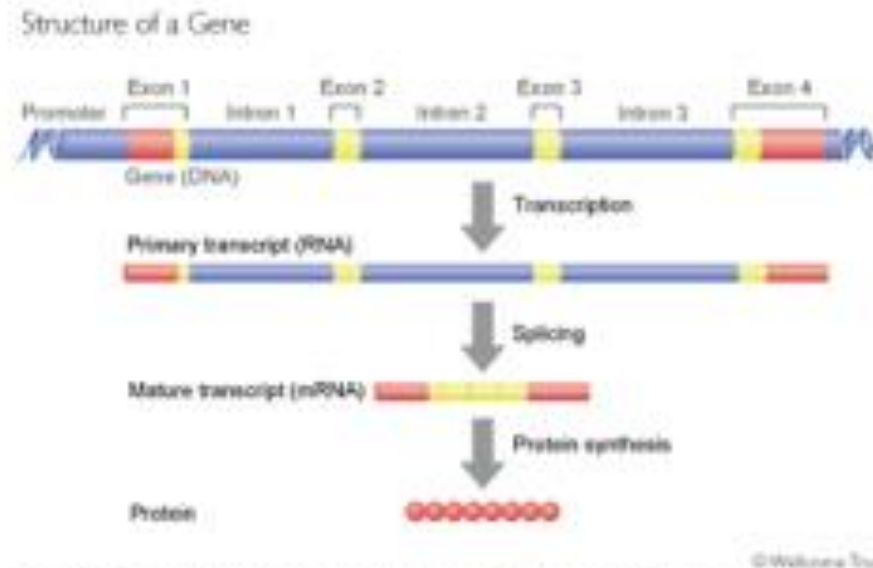
<http://www.youtube.com/watch?v=00vBqYDBW5s>

“Information:

that which can be communicated through symbolic language”

What are genes?

Defining genes by their structure



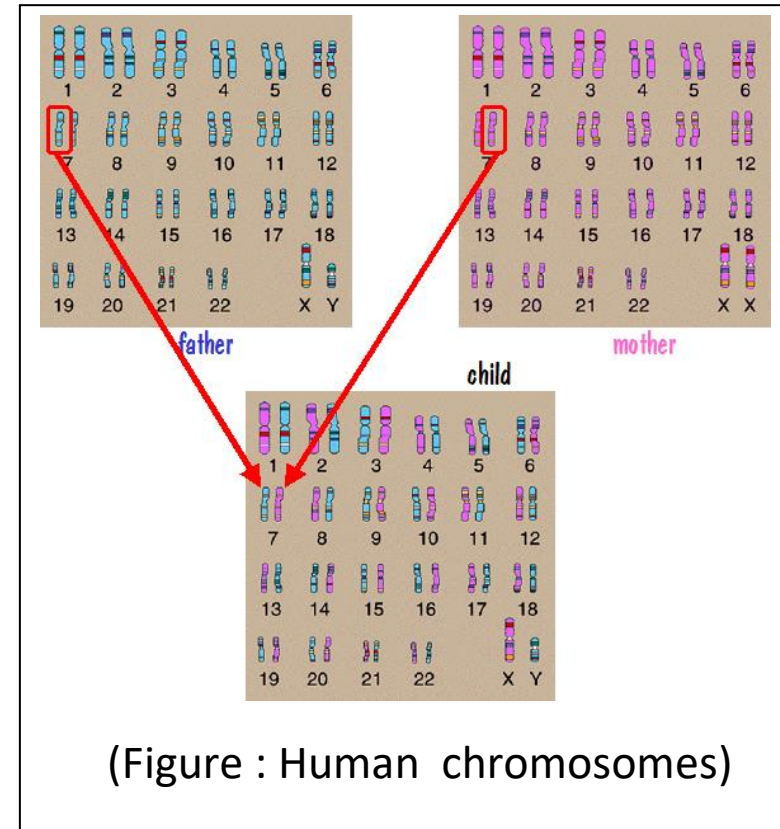
Simplified overview of gene structure and expression (for eukaryotes). A protein-coding gene is defined by the extent of the primary transcript. The gene is first transcribed to yield a primary transcript, which is processed to remove the introns. The mature transcript (messenger RNA, mRNA) is then translated 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 are genes?

- 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.

Defining genes as units of inheritance

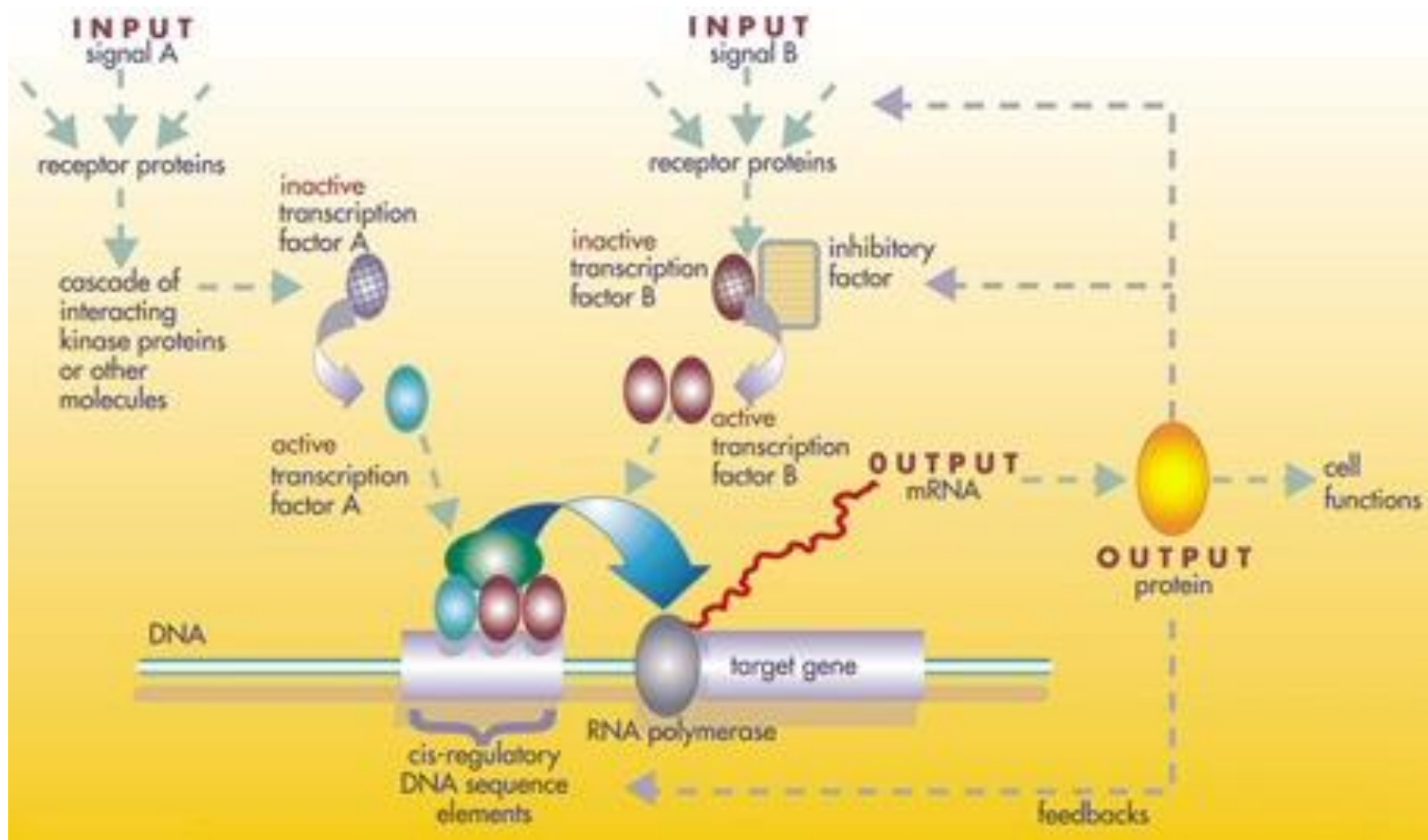


What is gene annotation?

- 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



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.

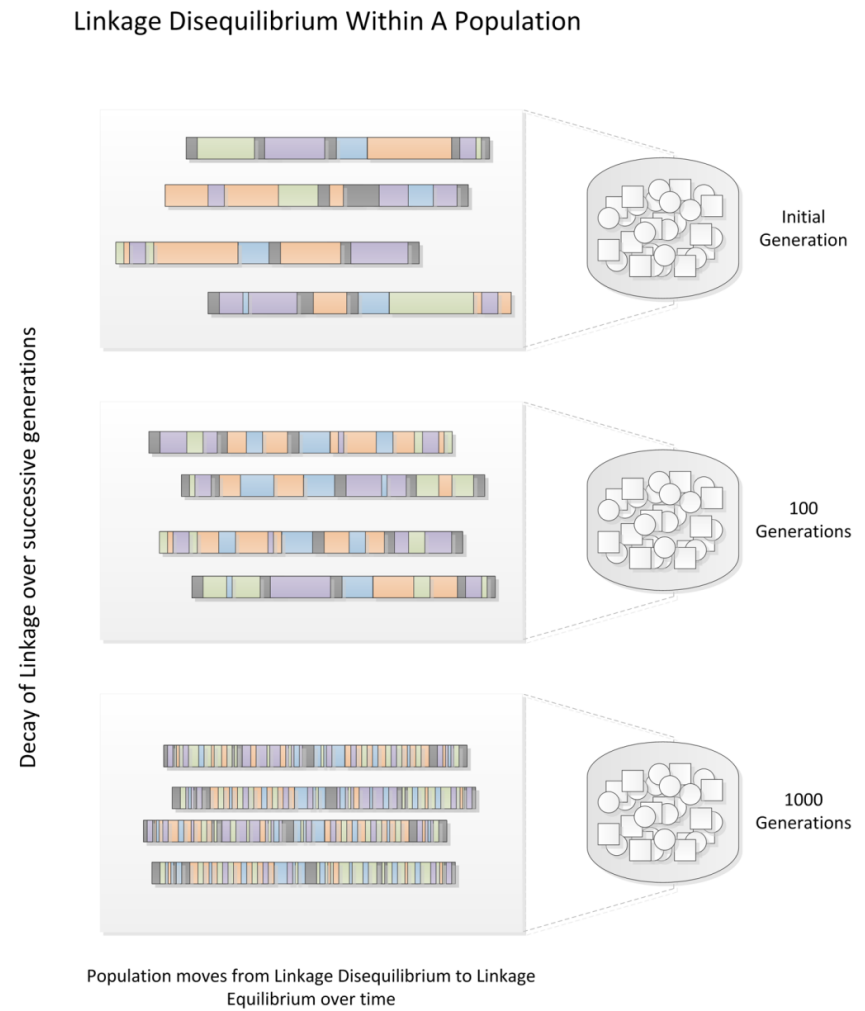
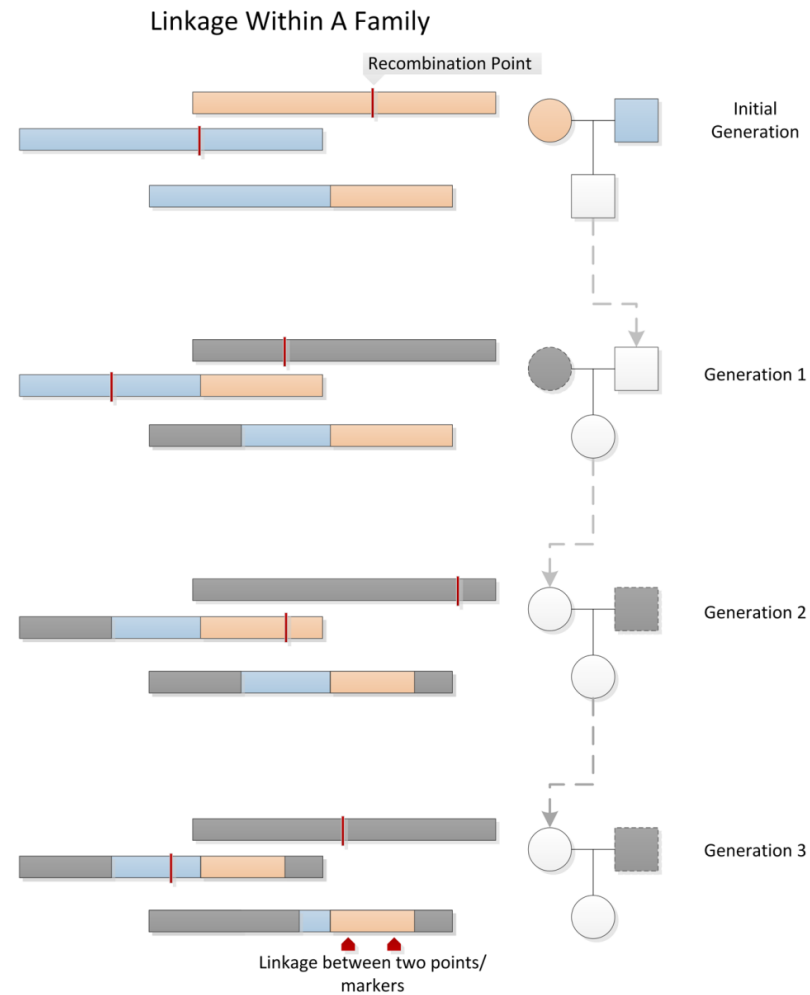
How is this possible?

How to generate a genetic map? (continued)

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

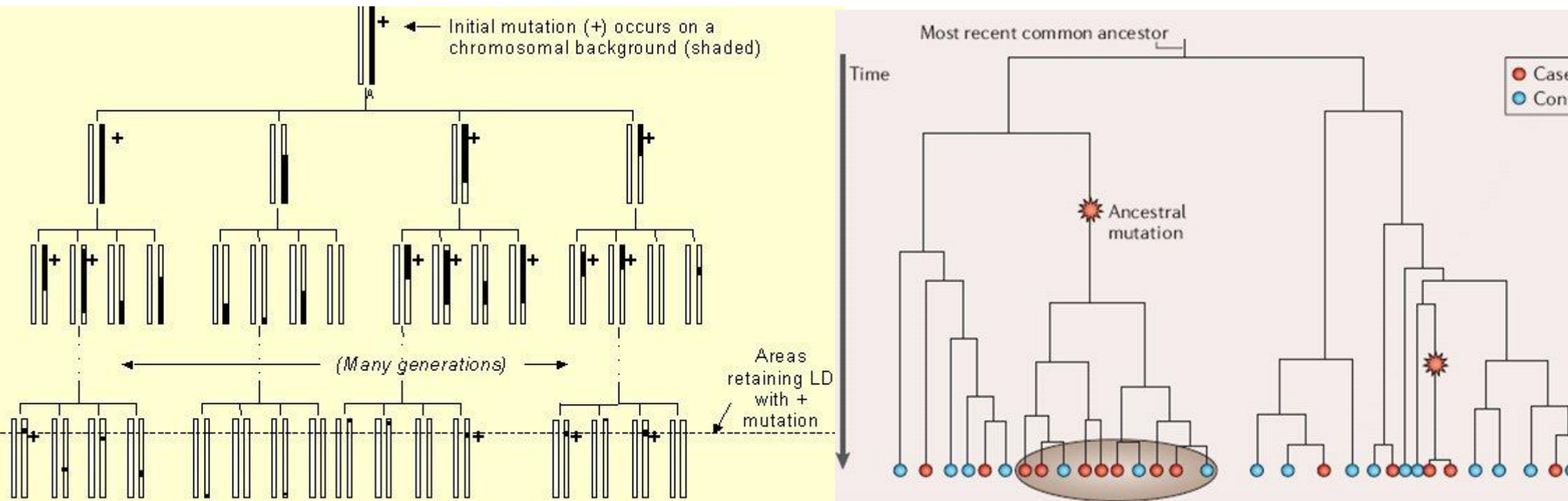
As eggs or sperm develop within a person's body, the 23 pairs of chromosomes within those cells exchange - or recombine - genetic material. If a particular gene is close to a DNA marker, the gene and marker will likely stay together during the recombination process, and be passed on together from parent 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#a1-3>)

1.c “The Human Genome Project”

genome.gov
National Human Genome Research Institute
National Institutes of Health

Research Funding | Research at NHGRI | Health | **Education** | Issues in Genetics | Newsroom | Careers & Training | About | For You

Home > Education > All About The Human Genome Project (HGP)

Education

- All About The Human Genome Project (HGP)
- Education Archive
- Fact Sheets
- Genetic Education Resources for Teachers
- NHGRI Webinar Series
- National DNA Day
- Online Genetics Education Resources
- Smithsonian NHGRI Genome Exhibition
- Talking Glossary
- Understanding the Human Genome Project

All About The Human Genome Project (HGP)

The Human Genome Project (HGP) was one of the great feats of exploration in history - an inward voyage of discovery rather than an outward exploration of the planet or the cosmos; an international research effort to sequence and map all of the genes - together known as the genome - of members of our species, *Homo sapiens*. Completed in April 2003, the HGP gave us the ability, for the first time, to read nature's complete genetic blueprint for building a human being.

In this section, you will find access to a wealth of information on the history of the HGP, its progress, cast of characters and future.

- 🔗 [Educational Resources](#)
- 🔗 [General Information](#)
- 🔗 [Research](#)
- 🔗 [Model Organisms](#)

Educational Resources

- [An Interactive Timeline of the Human Genome](#) [unlockinglifescode.org]
An interactive, hyper-linked timeline of genetics that takes the reader from Mendel (1865) to the completion of the mapping of the human genome (2003).
- [The Human Genome: A Decade of Discovery, Creating a Healthy Future](#)
A workshop for science reporters about the 10th anniversary of the completion of the draft sequence of the human genome and to look at the future of genomic research.
- [Understanding the Human Genome Project](#)
NHGRI's Online Education Kit
- [An Overview of the Human Genome Project](#)
A brief overview of the HGP.
- [50 Years of DNA: From Double Helix to Health](#)
Information about the celebration of the completion of the HGP and the 50th anniversary of the discovery of the

See Also:

- 📺 [White House Announcement](#)
June 26, 2000
- [Extramural Research Program](#)
- [Other Federal Agencies Involved in Genomics](#)

On Other Sites:

- [Human Genome Resources](#)
Access to the full human sequence

Historical overview (interludium)

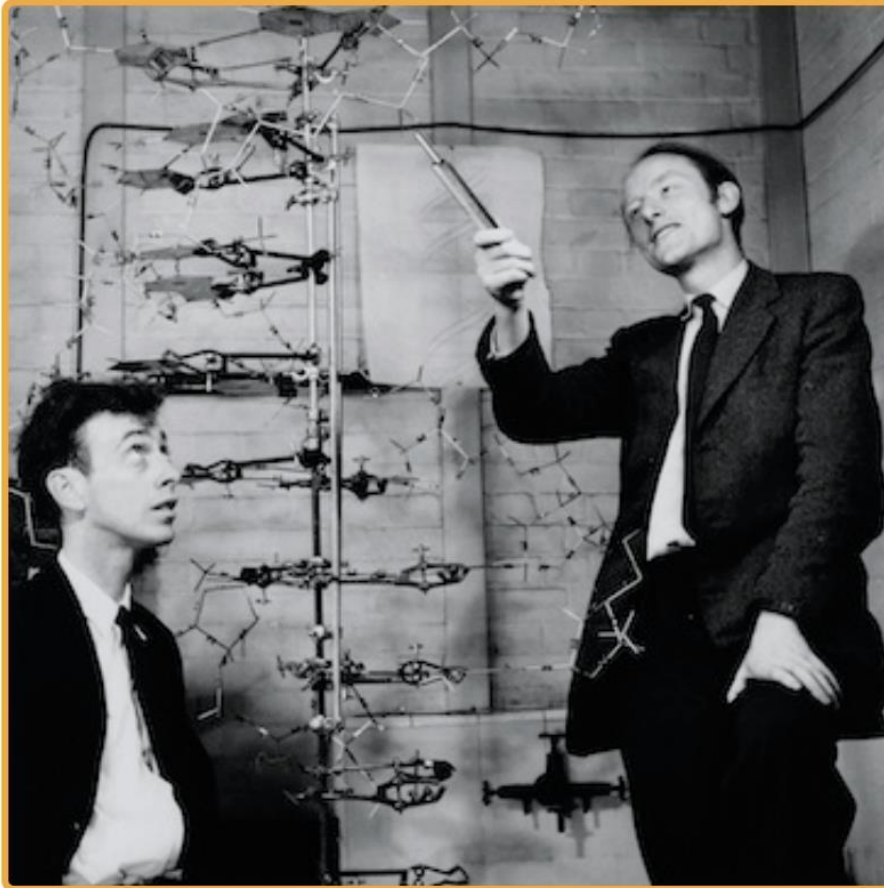


Gregor Mendel, the father of modern genetics, presents his research on experiments in plant hybridization ✕

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 $\frac{3}{4}$ yellow and $\frac{1}{4}$ green. Mendel's work, which was presented in 1865, showed that what we now call "genes" determine traits in predictable ways.

1865

Historical overview

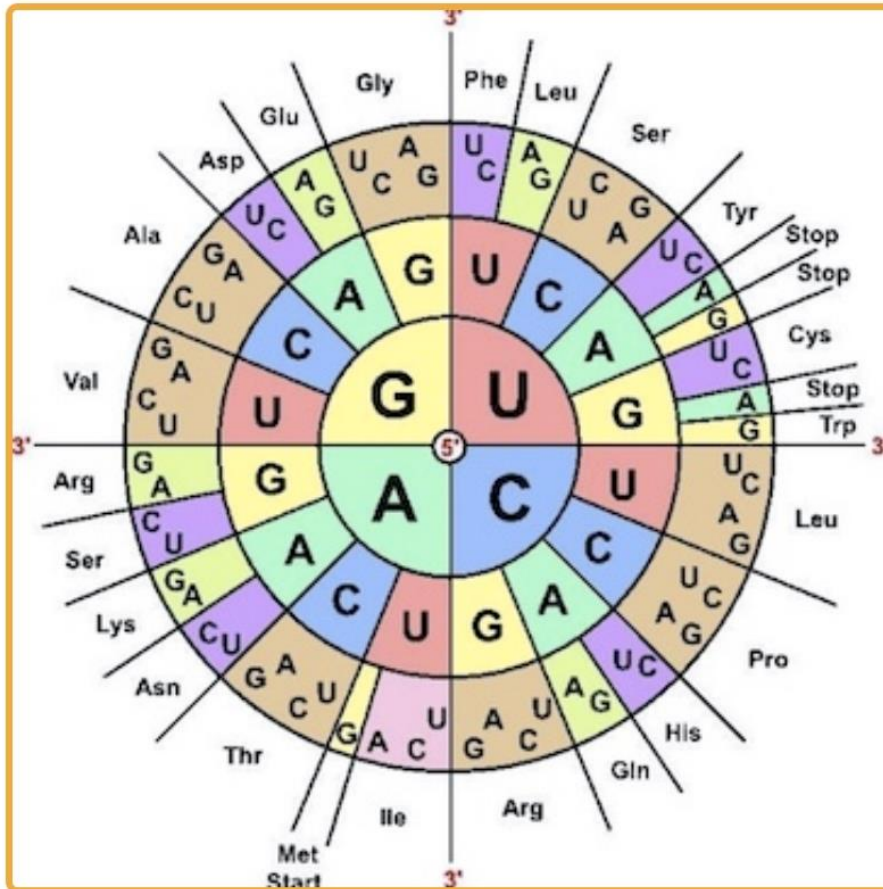


James Watson and Francis Crick discover the double helix structure of DNA ✕

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

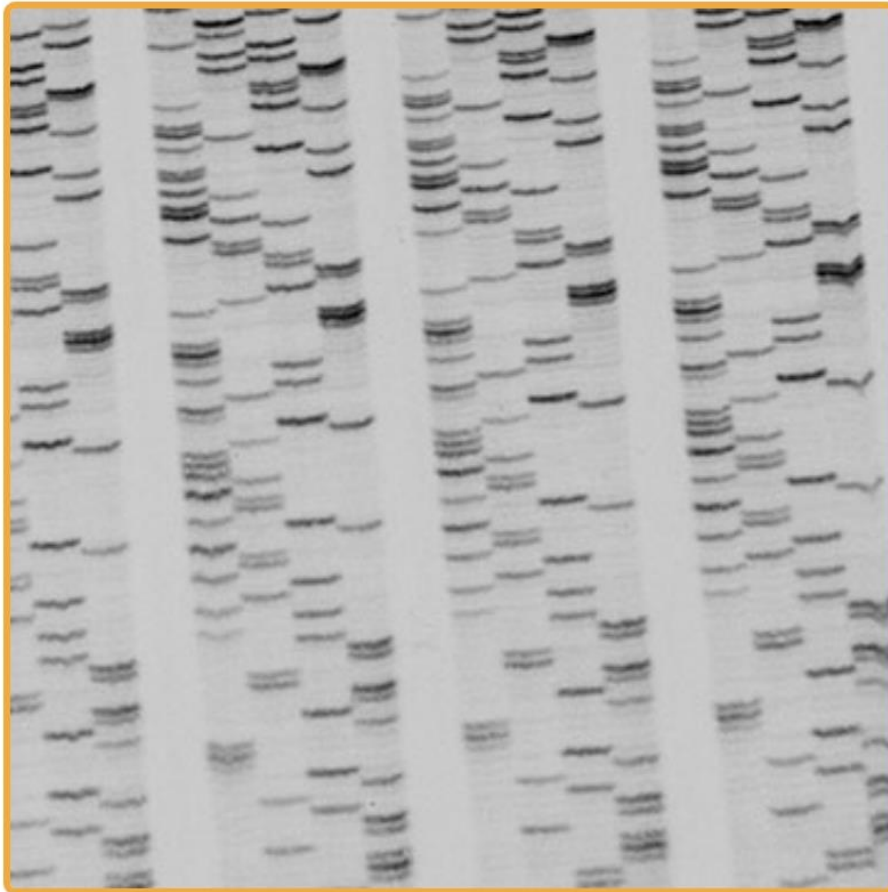


Marshall Nirenberg cracks the genetic code for protein synthesis

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

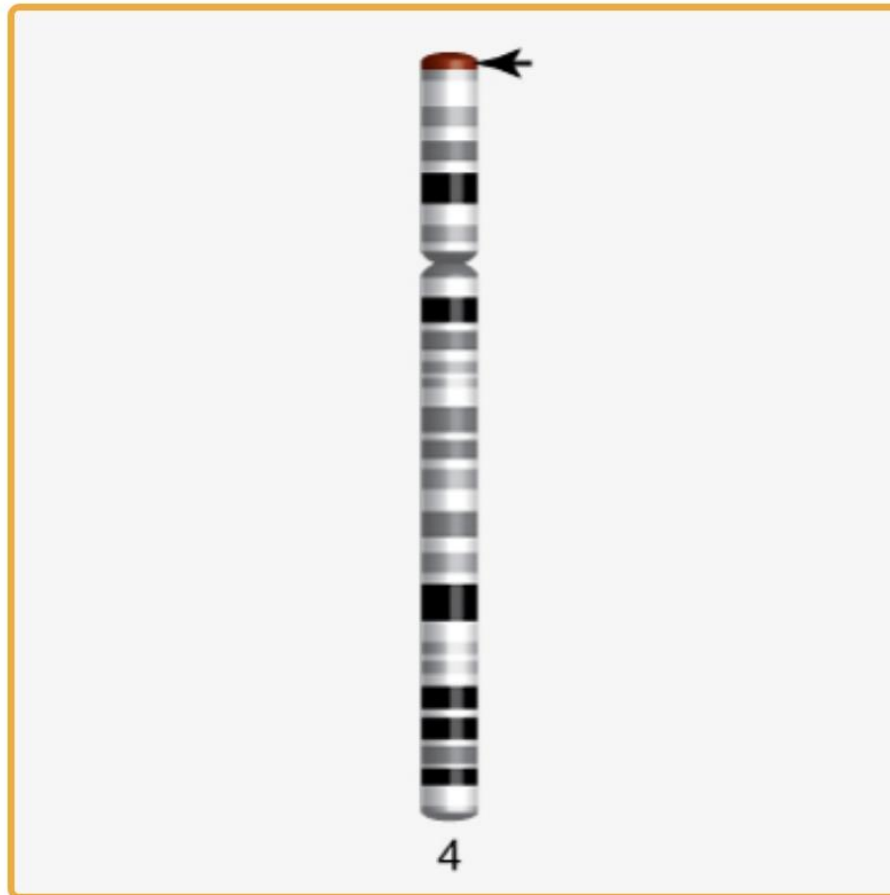


Frederick Sanger develops rapid DNA sequencing technique ✕

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

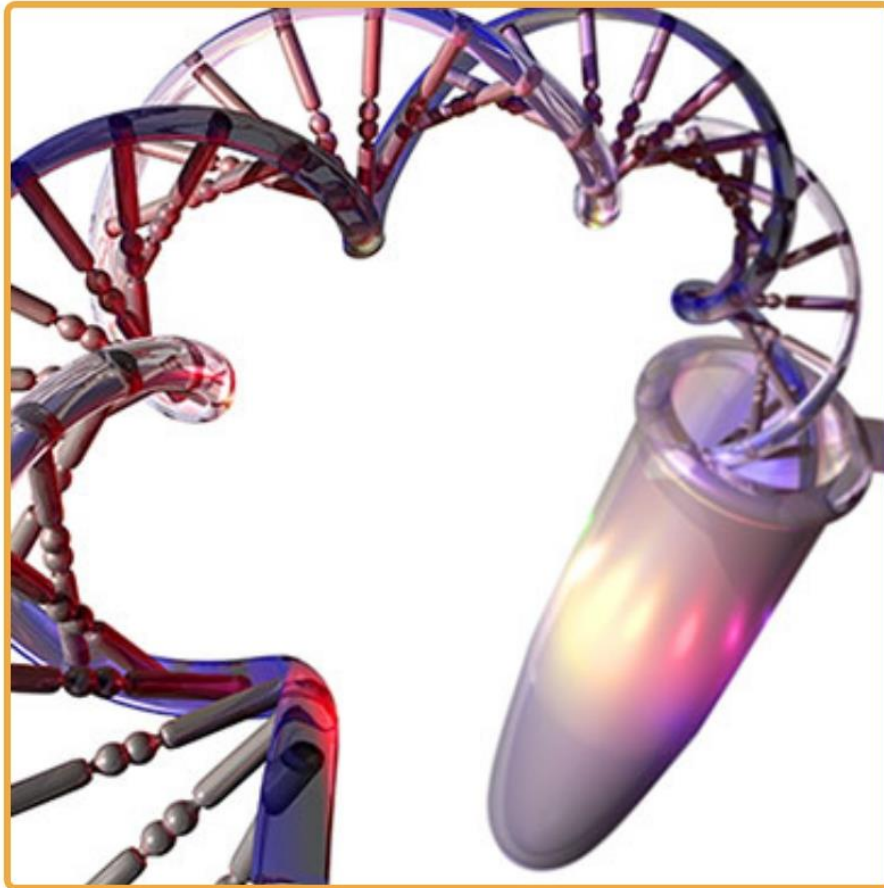


First genetic disease mapped, [✗] Huntington's Disease

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



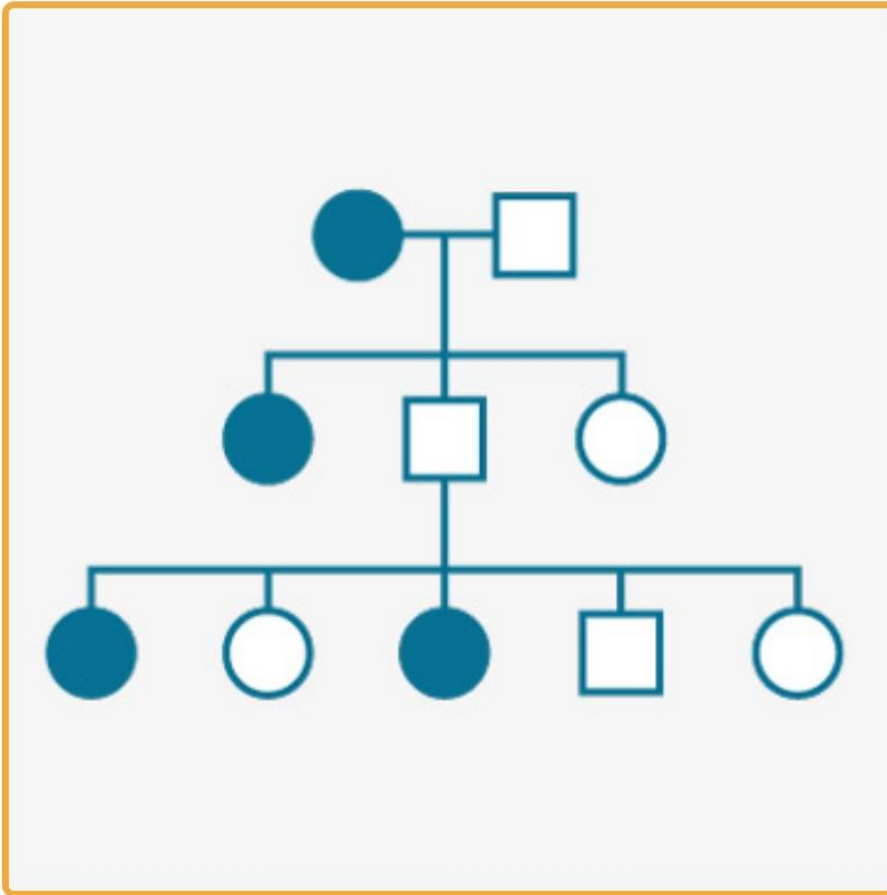
Invention of polymerase chain reaction (PCR) technology for amplifying DNA



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

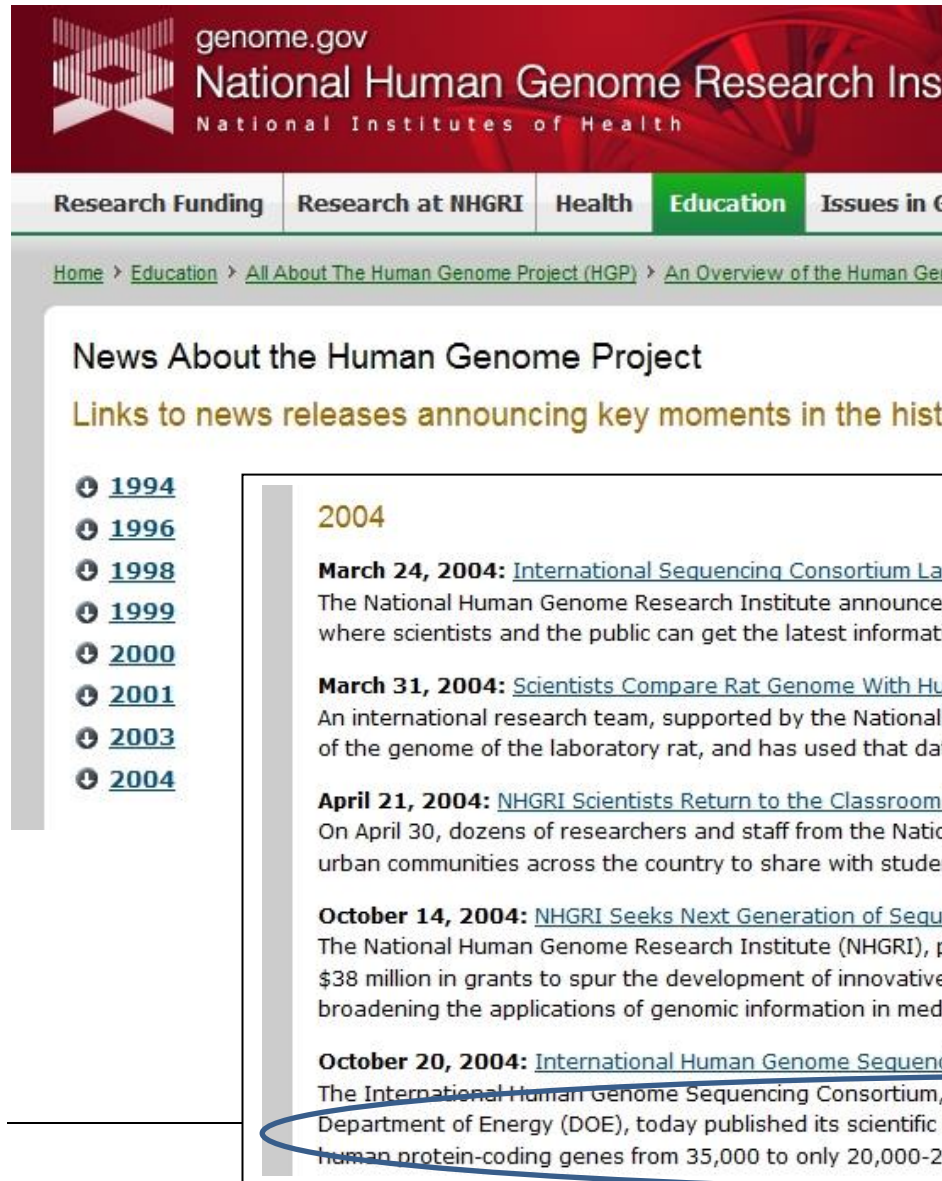
The Sequence of the Human Genome

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. Brans,¹ Robert A. Holt,¹ Jeannine D. Goczyns,¹ Peter Amanatides,¹ Richard M. Ballaw,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 Mal Simon,⁹ Carolyn Slayman,¹⁰ Michael Hunkapiller,¹¹ Randall Bolanos,¹ Arthur Delcher,¹ Ian Dew,¹ Daniel Fasulo,¹ Michael Flanagan,¹ Liliana Flores,¹ Aaron Halpern,¹ Sridhar Hannenhalli,¹ Saul Kravitz,¹ Samuel Levy,¹ Clark Moberly,¹ Knut Rainart,¹ Karin Remington,¹ Jane Abu-Threideh,¹ Ellen Beasley,¹ Kandra Biddick,¹ Vivian Bonazzi,¹ Rhonda Brandon,¹ Michele Cargill,¹ Ishwar Chandramouleswaran,¹ Rosana Charlab,¹ Kabir Chaturvedi,¹ Zuoming Deng,¹ Valentina Di Francesco,¹ Patrick Dunn,¹ Karan Elbeck,¹ Carlos Brangollista,¹ Andrei E. Gabrielian,¹ Weiniu Gan,¹ Wangmao Ge,¹ Fangchang 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 Li,¹ Yiding Lei,¹ Zherya Li,¹ Jiyin Li,¹ Yong Liang,¹ Xiaoying Lin,¹ Fu Lu,¹ Gennady V. Markulov,¹ Natalia Milshina,¹ Helen M. Moore,¹ Ashwinkumar K Nalk,¹ Vaibhav A. Narayan,¹ Beena Neelam,¹ Daborah Nusskarn,¹ Douglas B. Rusch,¹ Steven Saltzberg,¹² Wai Shao,¹ Bkiong Shua,¹ Jingtao Sun,¹ Zhen Yuan Wang,¹ Aihui Wang,¹ Xin Wang,¹ Jian Wang,¹ Ming-Hui Wai,¹ Ron Wides,¹² Chunlin Xie,¹ Chunxun Yan,¹ Allison Yao,¹ Jane Ye,¹ Ming Zhan,¹ Weiqing Zhang,¹ Hongyu Zhang,¹ Qi Zhao,¹ Liansheng Zhang,¹ Fei Zhong,¹ Wanyan Zhong,¹ Shaoqing C. Zhu,¹ Shaying Zhao,¹² Dennis Gilbert,¹ Suzanne Baumhueter,¹ Gene Slier,¹ Christine Carter,¹ Anibal Cravchik,¹ Trevor Woodage,¹ Farooz Ali,¹ Huljin An,¹ Aderonke Awo,¹ Danita Baldwin,¹ Holly Baden,¹ Mary Barnstead,¹ Ian Barrow,¹ Karen Beeson,¹ Dana Busam,¹ Amy Carver,¹ Angela Carter,¹ Ming Lai Cheng,¹ Liz Curry,¹ Steve Danaher,¹ Lionel Davanport,¹ Raymond Desillets,¹ Susanne Dietz,¹ Kristina Dodson,¹ Lisa Doup,¹ Steven Ferriera,¹ Naha Garg,¹ Andreas Gluecksmann,¹ Brit Hart,¹ Jason Haynes,¹ Charles Haynes,¹ Cheryl Halner,¹ Suzanne Hladin,¹ Damon Hostin,¹ Jarratt Houck,¹ Timothy Howland,¹ Chinyere Ibegwam,¹ Jeffery Johnson,¹ Francis Kalush,¹ Lesley Kline,¹ Shashi Koduru,¹ Amy Love,¹ Falecia Mann,¹ David May,¹ Steven McCawley,¹ Tina McIntosh,¹ Ivy McHullen,¹ Moe 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 Ruftal,¹ Richard Scott,¹ Cynthia Sittar,¹ Michelle Smallwood,¹ Erin Stewart,¹ Renee Strong,¹ Ellen Suh,¹ Reginald Thomas,¹ Ni Ni Tint,¹ Sukyae Tsa,¹ Claire Vech,¹ Gary Wang,¹ Jeremy Watter,¹ Sherita Williams,¹ Monica Williams,¹ Sandra Windsor,¹ Emily Winn-Daen,¹ Kerillan Wolfe,¹ Jayshree Zaveri,¹ Karana Zaveri,¹ Josep F. Abril,¹⁴ Roderic Guigo,¹⁴ Michael J. Campbell,¹ Kimman V. Sjolander,¹ Brian Karlak,¹ Anish Kajarwal,¹ Hualyu Mi,¹ Betty Lazarova,¹ Thomas Hatton,¹ Apurva Narechania,¹ Karan 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 Basti,¹ James Baxendale,¹ Louis Blick,¹ Marcelo Caminha,¹ John Carnes-Stino,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 Foster,¹ Harold Gire,¹ Stephen Glanowski,¹ Kenneth Glasser,¹ Anna Glodak,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 Pock,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 Melyuan Wen,¹ David Wu,¹ Mitchell Wu,¹ Ashley Xia,¹ Ali Zandieh,¹ Xiaohong Zhu¹

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



genome.gov
National Human Genome Research Institute
National Institutes of Health

Research Funding | Research at NHGRI | Health | **Education** | Issues in C

Home > Education > All About The Human Genome Project (HGP) > An Overview of the Human Ge

News About the Human Genome Project

Links to news releases announcing key moments in the hist

- 1994
- 1996
- 1998
- 1999
- 2000
- 2001
- 2003
- 2004

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


National Human Genome Research Institute
National Institutes of Health

Research Funding
Research at NHGRI
Health
Education
Issues in Genetics
Newsroom
Careers & Training
About
For You

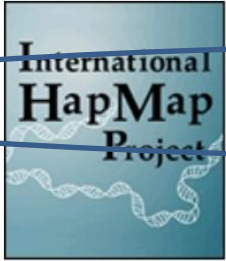



Home > [Education](#) > [Understanding the Human Genome Project](#) > [Dynamic Timeline](#) > [2004-The Future](#) > **2005b: HapMap Project Completed**



Online Education Kit: 2004–The Future

- 2004a: Rat and Chicken Genomes Sequenced
- 2004b: FDA Approves First Microarray
- 2004c: Refined Analysis of Complete Human Genome Sequence
- 2004d: Surgeon General Stresses Importance of Family History
- 2005a: Chimpanzee Genomes Sequenced
- 2005b: HapMap Project Completed**
- 2005c: Trypanosomatid Genomes Sequenced
- 2005d: Dog Genomes Sequenced
- 2006a: The Cancer Genome Atlas (TCGA) Project Started
- 2006b: Second Non-human Primate Genome is Sequenced
- 2006c: Initiatives to Establish the Genetic and Environmental Causes of Common Diseases Launched
- The Future

2005: HapMap Project Completed



The International HapMap Consortium published a catalog of human genetic variation that is expected to help speed the identification of genes associated with common diseases such as asthma, cancer, diabetes, and heart disease. While the Human Genome Project focused on the DNA sequence from a single individual, the HapMap project focused on variation in the genome and on human populations. The \$138 million project was a three-year collaboration between more than 200 researchers from Canada, China, Japan, Nigeria and the United States. The new paper described the completion of a Phase I HapMap that contains more than 1 million markers of genetic variation. At the time of the publication, the consortium was nearing completion of a Phase II HapMap that would contain more than 3 million genetic markers.

 Share
  Print

See Also:

[2005 Release: International Consortium Completes Map](#)


[International HapMap Project](#)


On Other Sites:

[International HapMap Project](#)
Web page for the International HapMap Consortium

More Information


References:

The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nature Genetics*, 5: 467-475. 2004. [[Full Text](#)] 

International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437: 1229-1320. 2005. [[Full Text](#)] 

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., et al. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308: 385-389. 2005. [[PubMed](#)]

To view the PDFs on this page, you will need Adobe Reader.



Historical overview

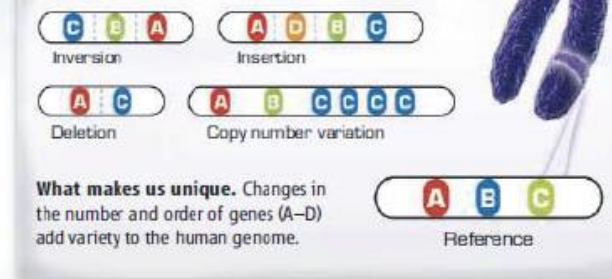
BREAKTHROUGH OF THE YEAR

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.

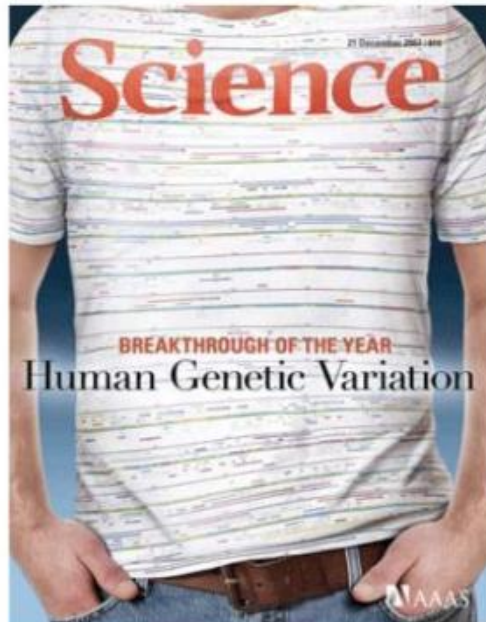
Less than a year ago, the big news was triangulating variation between us and our primate cousins to get a better handle on genetic changes along the evolutionary tree that led to humans. Now, we have moved from asking what in our DNA makes us human to striving to know what in my DNA makes me me.



Pennisi 2007 Science 318:1842-3

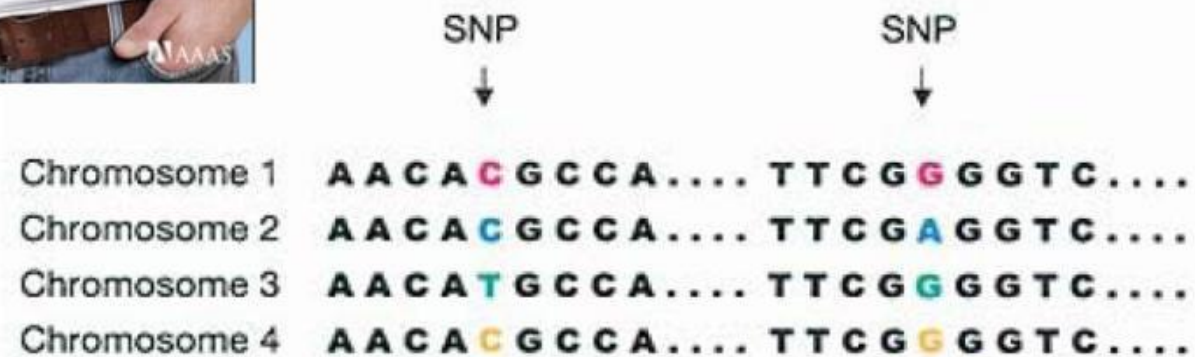
2007 SCIENTIFIC BREAKTHROUGH OF THE YEAR

Science Magazine, December 21, 2007



“It’s all about me!”

Single Nucleotide Polymorphisms (SNPs)



Historical overview: associating genetic variation to disease outcomes



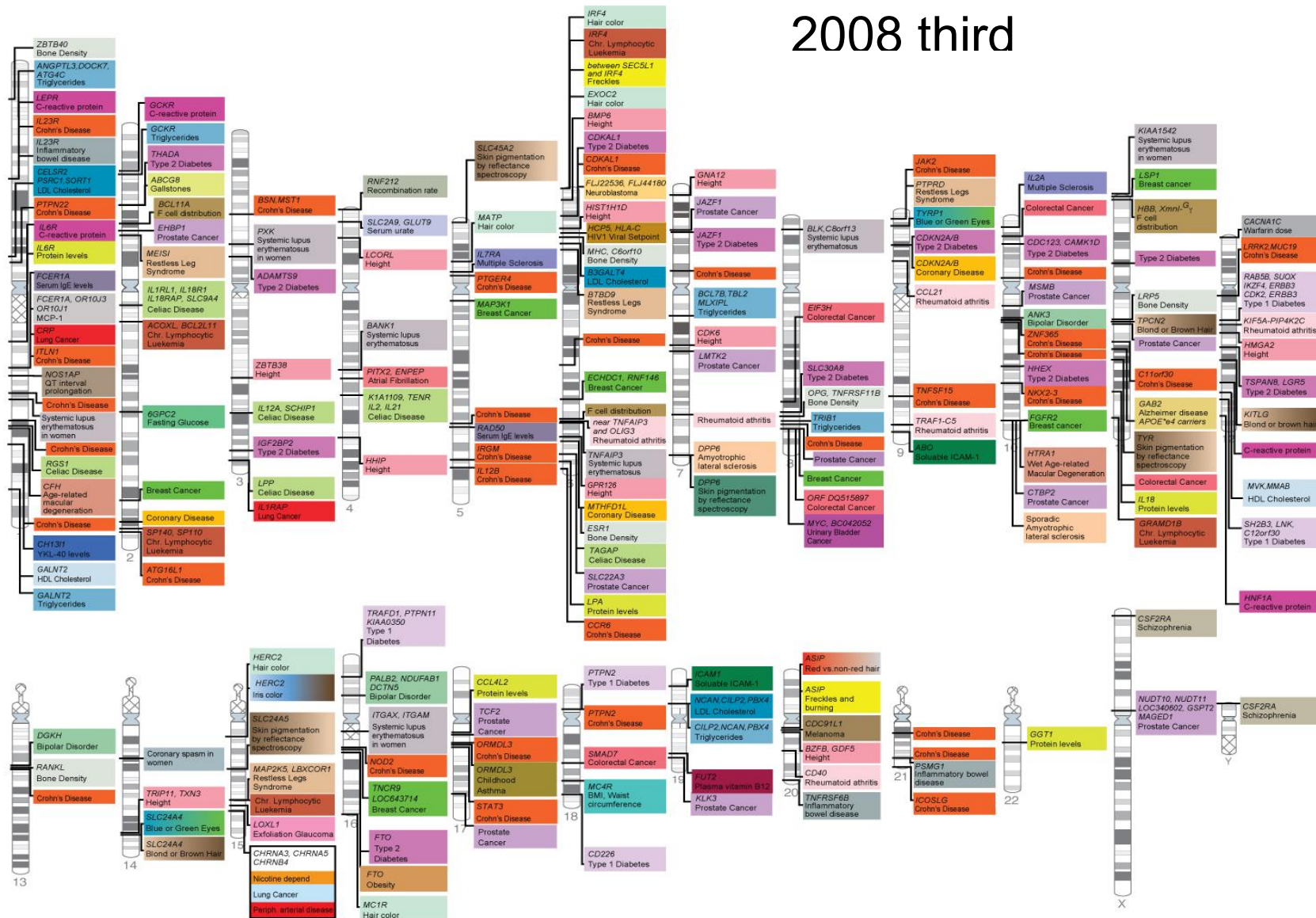
BREAKTHROUGH OF THE YEAR: The Runners-Up

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

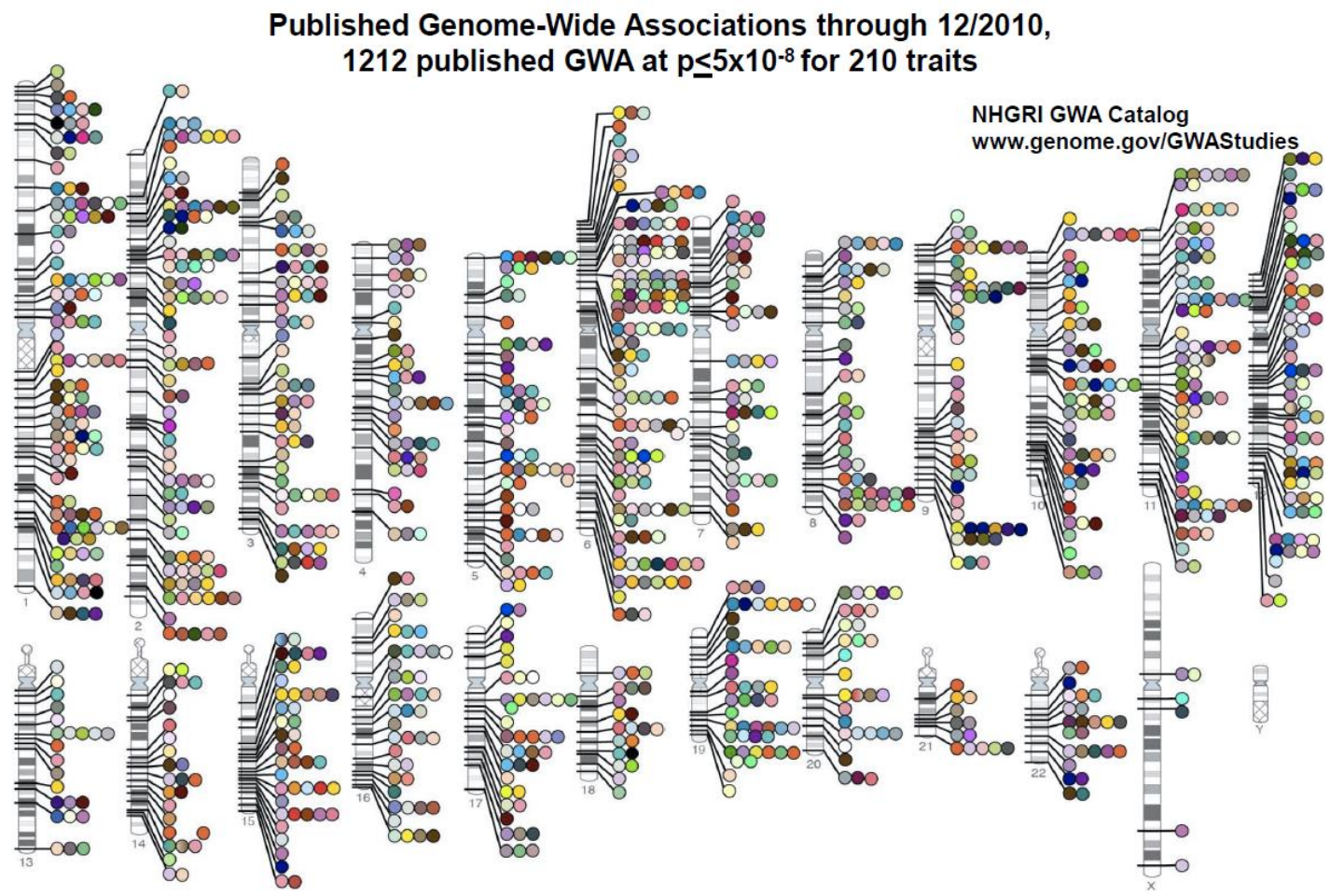
Areas to Watch in 2007

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?

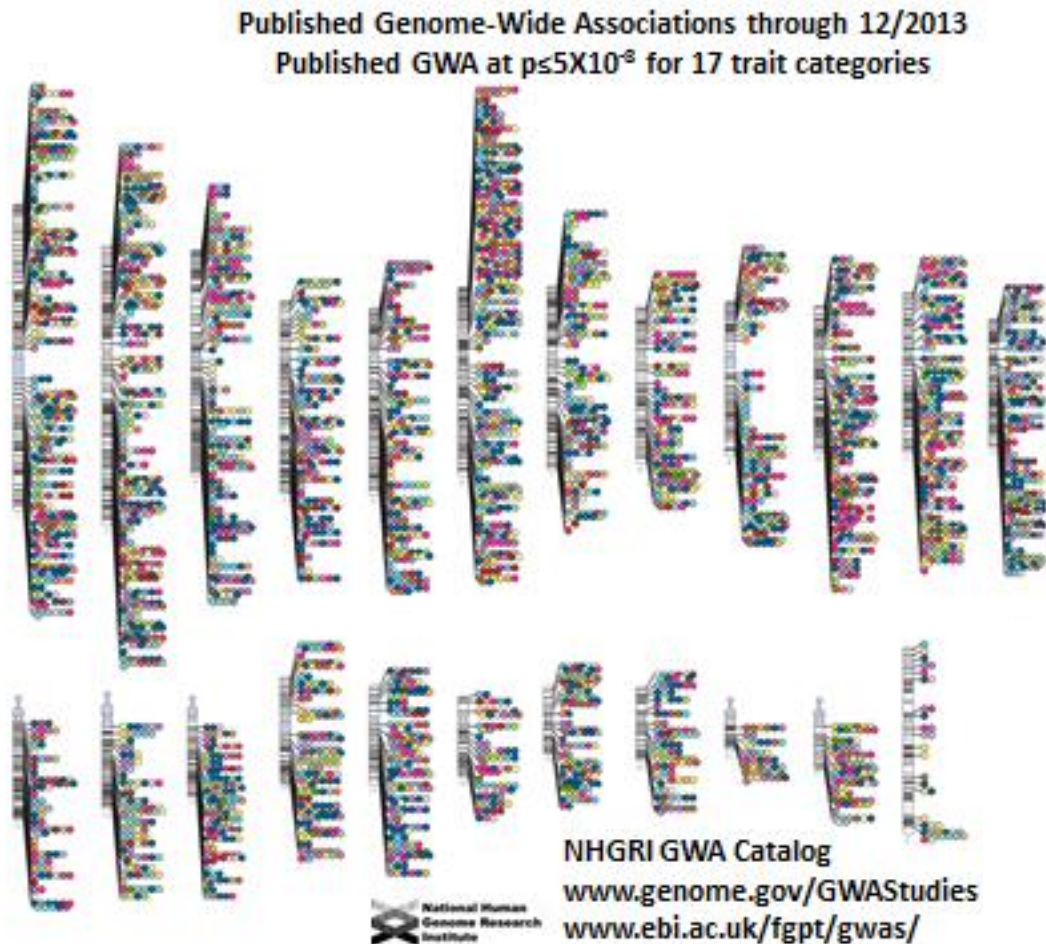
Historical overview: GWAs as a tool to “map” diseases



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 measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

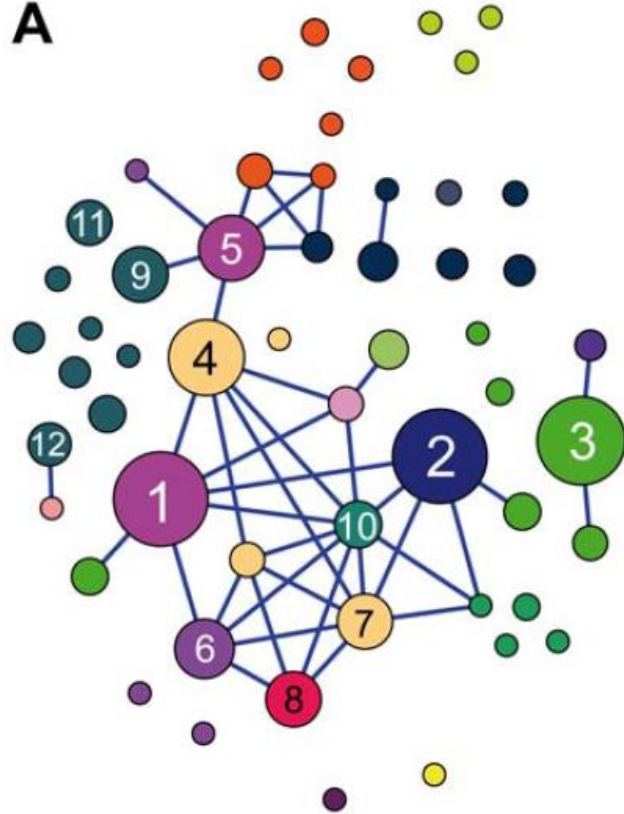
Historical overview: trait categories and nr of SNPs

Date: 3/3/2020



Historical overview: inter-relationships (networks)

A

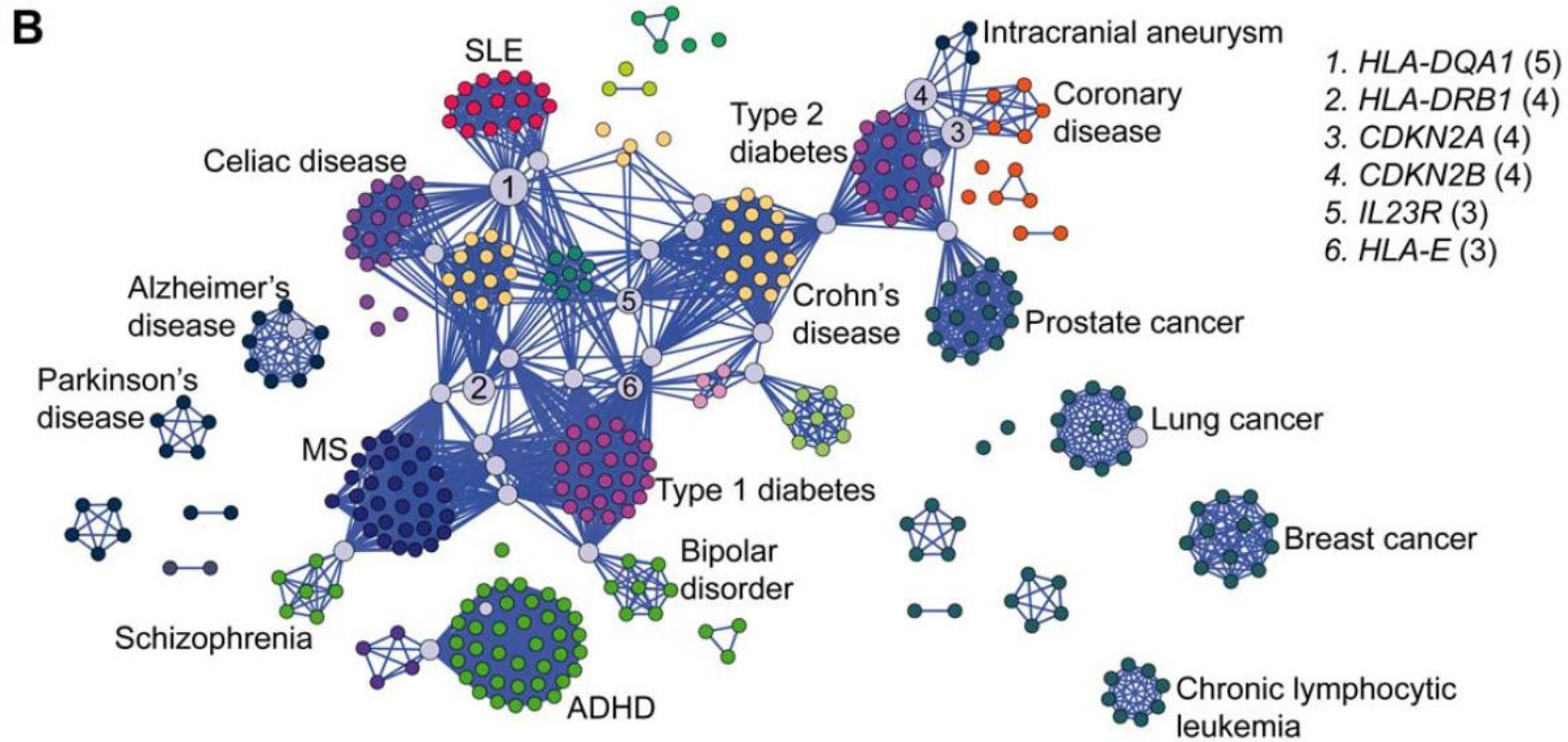


1. Type 1 diabetes (36)
2. Multiple sclerosis (36)
3. ADHD and conduct disorder (33)
4. Crohn's disease (27)
5. Type 2 diabetes (22)
6. Celiac disease (19)
7. Ulcerative colitis (17)
8. Systemic lupus erythematosus (17)
9. Prostate cancer (17)
10. Rheumatoid arthritis (13)
11. Breast cancer (12)
12. 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)
- Nm, Es, Is
- Skin and connective tissue diseases (Sc)
- Sc, Is
- Urogenital diseases

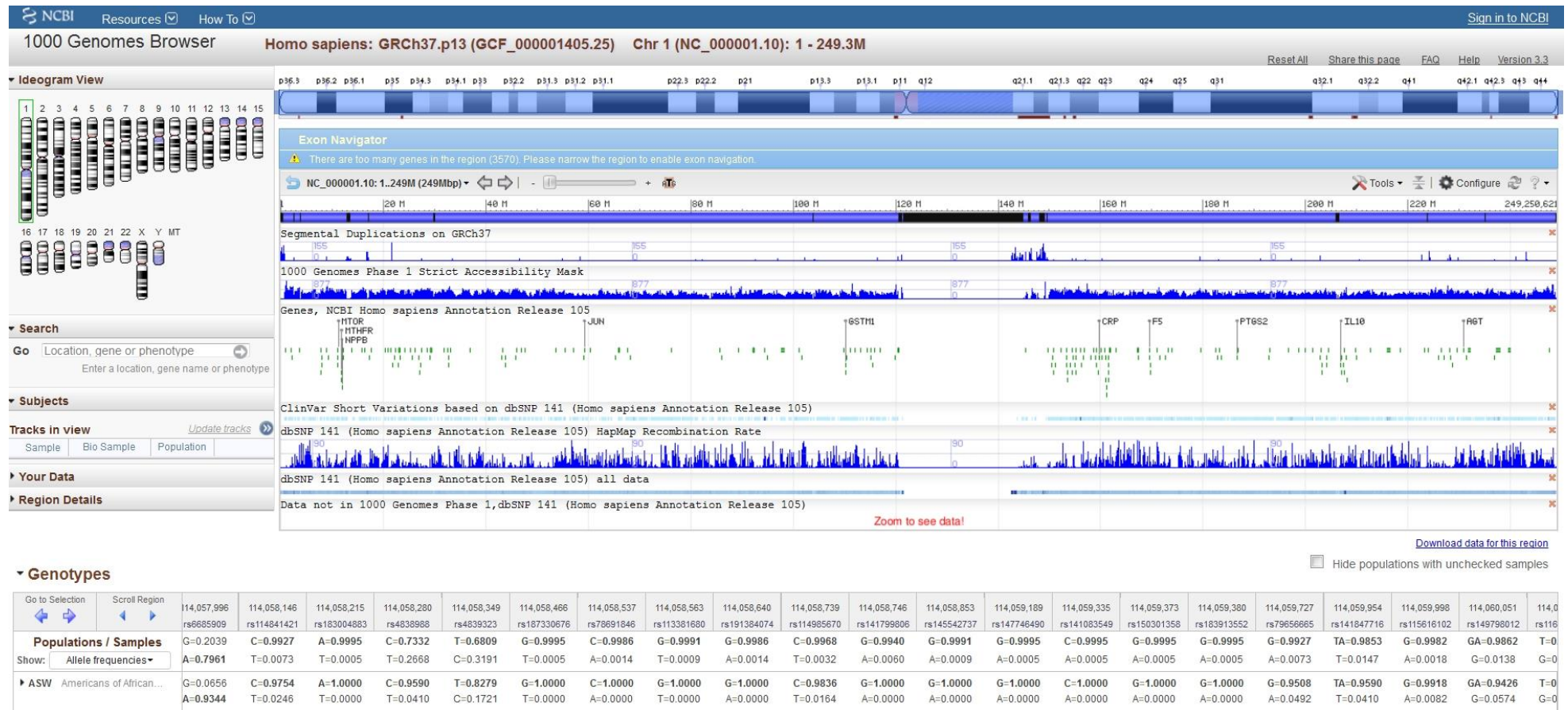
(Barrenas et al 2009: complex disease network – nodes are diseases)

Historical overview: inter-relationships (networks)



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

Historical overview: exome sequencing, full genome sequencing



Historical overview: monitoring the progress

The screenshot shows the OMIM website interface. At the top, there is a blue navigation bar with the NCBI logo, "Resources" and "How To" dropdown menus, and a "Sign in to NCBI" link. Below this is a search bar with "OMIM" in the dropdown menu and a "Search" button. There are also links for "Limits" and "Advanced" search options, and a "Help" link.

The main content area features a large banner with the OMIM logo and a classical portrait of a man. To the right of the banner, the text reads: "OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh. Its official home is omim.org."

Below the banner, there are three columns of links:

- Using OMIM**
 - [Getting Started](#)
 - [FAQ](#)
- OMIM tools**
 - [OMIM API](#)
- Related Resources**
 - [ClinVar](#)
 - [Gene](#)
 - [GTR](#)
 - [MedGen](#)

At the bottom left, it says "Last updated on: 05 Oct 2014".

OMIM: molecular dissection of human disease

- Online Mendelian Inheritance in Man (OMIM[®]) is a continuously updated **catalog of human genes and genetic disorders and traits** (i.e. coded phenotypes, where phenotype is any characteristic of the organism), with particular focus on the molecular relationship between genetic variation and phenotypic expression.
- It can be considered to be a phenotypic companion to the Human Genome Project. OMIM is a continuation of Dr. Victor A. McKusick's Mendelian Inheritance in Man, which was published through 12 editions, the last in 1998.
- OMIM is currently biocurated at the McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine.
- Frequently asked questions: <http://www.omim.org/help/faq>

Accessing OMIM

The screenshot shows the NCBI website interface. At the top, there is a search bar and navigation icons. Below the search bar is a navigation menu with the following items: NCBI Home, Resource List (A-Z), All Resources, Chemicals & Bioassays, Data & Software, DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation. The 'Resource List (A-Z)' menu is expanded, showing a list of databases including Genome, GEO DataSets, GEO Profiles, GSS, GTR, HomoloGene, Identical Protein Groups, MedGen, MeSH, NCBI Web Site, NLM Catalog, Nucleotide, OMIM, PMC, PopSet, Probe, Protein, Protein Clusters, PubChem BioAssay, and PubChem Compound. The main content area features a header with the NCBI logo and a search bar. Below the header, there is a section titled 'Welcome to NCBI' with a brief description of the center's mission. The main content is organized into six columns, each with a title and a description: 'Submit' (Submit data or manuscripts to our databases), 'Download' (Transfer NCBI data to your computer), 'Learn' (Find help documents, attend a class or watch a tutorial), 'Develop' (Build APIs and code to build applications), 'Analyze' (Identify an NCBI tool for your data analysis task), and 'Research' (Explore NCBI research and collaborative projects). To the right of the main content, there is a 'Popular Resources' section with links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Below this is an 'NCBI News & Blog' section with a recent article titled 'PubMed Labs is now part of NCBI Labs' dated 03 Oct 2017, and another article titled 'October 11 NCBI Minute: Introducing the New RefSeq Functional Elements Project'.

Finding the trees for the forest

NCBI Resources How To Sign in to NCBI

MedGen MedGen asthma Search

Create alert Limits Advanced Help

See MedGen results with **asthma** as a clinical feature (41)

Summary 20 per page

Send to:

Filter your results:

- All (319)
- Records in GTR (76)
- Records in OMIM (68)
- Diseases (172)
- Records in Orphanet (27)
- Records in HPO (9)
- Recommended for clinicians (77)

Manage Filters

Search results

Items: 9

- [Asthma](#)
 1. A chronic respiratory disease manifested as difficulty breathing due to the narrowing of bronchial passageways. [from [NCI](#)]

MedGen UID: 2109 • Concept ID: [C0004096](#) • Disease or Syndrome

[GTR](#) [ClinVar](#) [Genes](#) [OMIM](#) [GeneReviews](#)

- [Exercise-induced asthma](#)

Find related data

Database: Select

Finding the trees for the forest (click on “Asthma”)

NCBI Resources How To
Sign in to NCBI

MedGen

Search
Limits Advanced Help

Full Report Send to:

Asthma
 MedGen UID: 2109 • Concept ID: C0004096 • Disease or Syndrome

Synonyms: Bronchial asthma; Reactive airway disease

SNOMED CT: Airway hyperreactivity (195967001); Asthmatic (195967001); Bronchial asthma (195967001); Asthma (195967001)

HPO: HP:0002099

Go to:

[Definition](#)

A chronic respiratory disease manifested as difficulty breathing due to the narrowing of bronchial passageways. [from NCI]

Go to:

[Term Hierarchy](#)

▼ Diseases, Respiratory Tract

Table of contents

Definition

Term Hierarchy

Conditions with this feature

Recent clinical studies

Recent systematic reviews

Genetic Testing Registry

Deletion/duplication analysis (18)

Sequence analysis of the entire coding

Finding the trees for the forest (Click on ClinVar)

ClinVar [Advanced](#) [Help](#)

- [Home](#)
- [About](#) ▾
- [Access](#) ▾
- [Help](#) ▾
- [Submit](#) ▾
- [Statistics](#) ▾
- [FTP](#) ▾

Clinical significance

- Conflicting interpretations (0)
- Benign (0)
- Likely benign (0)
- Uncertain significance (6)
- Likely pathogenic (3)
- Pathogenic (2)
- Risk factor (1)

Molecular consequence

- Frameshift (1)
- Missense (0)
- Nonsense (0)
- Splice site (0)
- ncRNA (0)

Tabular ▾ 100 per page ▾ Sort by Relevance ▾

Download: ▾


SIDEBAR

Links from MedGen

Items: 12

	Variation <i>Location</i>	Gene(s)	Protein change	Condition(s)	Clinical significance <i>(Last reviewed)</i>	Review status	Acco
<input type="checkbox"/>	1. NM_001145775.2(FKBP5):c.106-2332 A>C <i>GRCh37:</i> Chr6:35607267 <i>GRCh38:</i> Chr6:35639490	FKBP5		Asthma	risk factor (Jun 20, 2019)	no assertion criteria provided	VCV00
<input type="checkbox"/>	2. 46:XY;t(1;14)(q42;q13)			Exotropia, Split foot, Chronic	Uncertain	criteria	VCV00

Finding the trees for the forest (click on the gene)


[Resources](#) 
[How To](#) 
[Sign in to NCBI](#)

Gene




[Advanced](#) [Help](#)

Full Report 


Send to: 

FKBP5 FKBP prolyl isomerase 5 [*Homo sapiens* (human)]

Gene ID: 2289, updated on 28-Feb-2020

 **Summary**  

Official Symbol FKBP5 provided by [HGNC](#)
Official Full Name FKBP prolyl isomerase 5 provided by [HGNC](#)
Primary source [HGNC:HGNC:3721](#)
See related [Ensembl:ENSG00000096060](#) [MIM:602623](#)
Gene type protein coding
RefSeq status REVIEWED
Organism [Homo sapiens](#)
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;

Table of contents 

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Expression
- Bibliography
- Phenotypes
- Variation
- Pathways from PubChem
- Interactions

Finding the trees for the forest (scroll down for genomic context etc)

Genomic context
⌆ ?

[See FKBP5 in Genome Data Viewer](#)

Location: 6p21.31

Exon count: 13

Annotation release	Status	Assembly	Chr	Location
109.20191205	current	GRCh38.p13 (GCF_000001405.39)	6	NC_000006.12 (35573585..35728583, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	6	NC_000006.11 (35541362..35696360, complement)

Chromosome 6 - NC_000006.12

The diagram shows a genomic map of Chromosome 6 region NC_000006.12. The coordinates range from 35497874 to 35774263. Genes and loci shown include TULP1, RPS15AP19, RPL36P9, MIR5690, LOC285847, LOC101929309, FKBP5 (highlighted in red), ARMC12, LOC112267956, and LOC100652794. Arrows indicate the direction of transcription for each feature.

Variation Viewer (GRCh37.p13)

Variation Viewer (GRCh38)

1000 Genomes Browser (GRCh37.p13)

Ensembl

UCSC

Related information

Order cDNA clone

3D structures

BioAssay by Target (List)

BioAssay by Target (Summary)

BioAssay, by Gene target

BioAssays, RNAi Target, Tested

BioProjects

Finding the trees for the forest (additional phenotypes and markers)

Phenotypes



[BioGRID CRISPR Screen Phenotypes \(2 hits/791 screens\)](#)

[Find tests for this gene in the NIH Genetic Testing Registry \(GTR\)](#)

[Review eQTL and phenotype association data in this region using PheGenI](#)

Associated conditions

Description	Tests
Major depressive disorder MedGen: C1269683 , OMIM: 608516 , GeneReviews: Not available	Compare labs

Variation



[See variants in ClinVar](#)

[See studies and variants in dbVar](#)

[See Variation Viewer \(GRCh37.p13\)](#)

[See Variation Viewer \(GRCh38\)](#)

Finding the trees for the forest (click on GTR instead of ClinVar)

NCBI Resources How To
Sign in to NCBI

GTR: GENETIC TESTING REGISTRY

Tests
Search

[Advanced search for tests](#)

Tests
(18)

Conditions
(1)

Genes
(1)

Laboratories
(1)

Filters

▼ **Test type**

Clinical (18)

▼ **Test purpose**

Diagnosis (18)

Mutation Confirmation (18)

▼ **Test method**

Results: 1 to 18 of 18

Tests names and labs	Conditions	Genes and analytes	Methods
Hyper-IgE Syndrome NGS Panel Fulgent Genetics United States	30	5	D Deletion/duplication analysis C Sequence analysis of the entire coding region
B-Negative Severe Combined Immunodeficiency NGS Panel	73	13	D Deletion/duplication analysis C Sequence analysis of the entire coding

2 The rise of GWAs



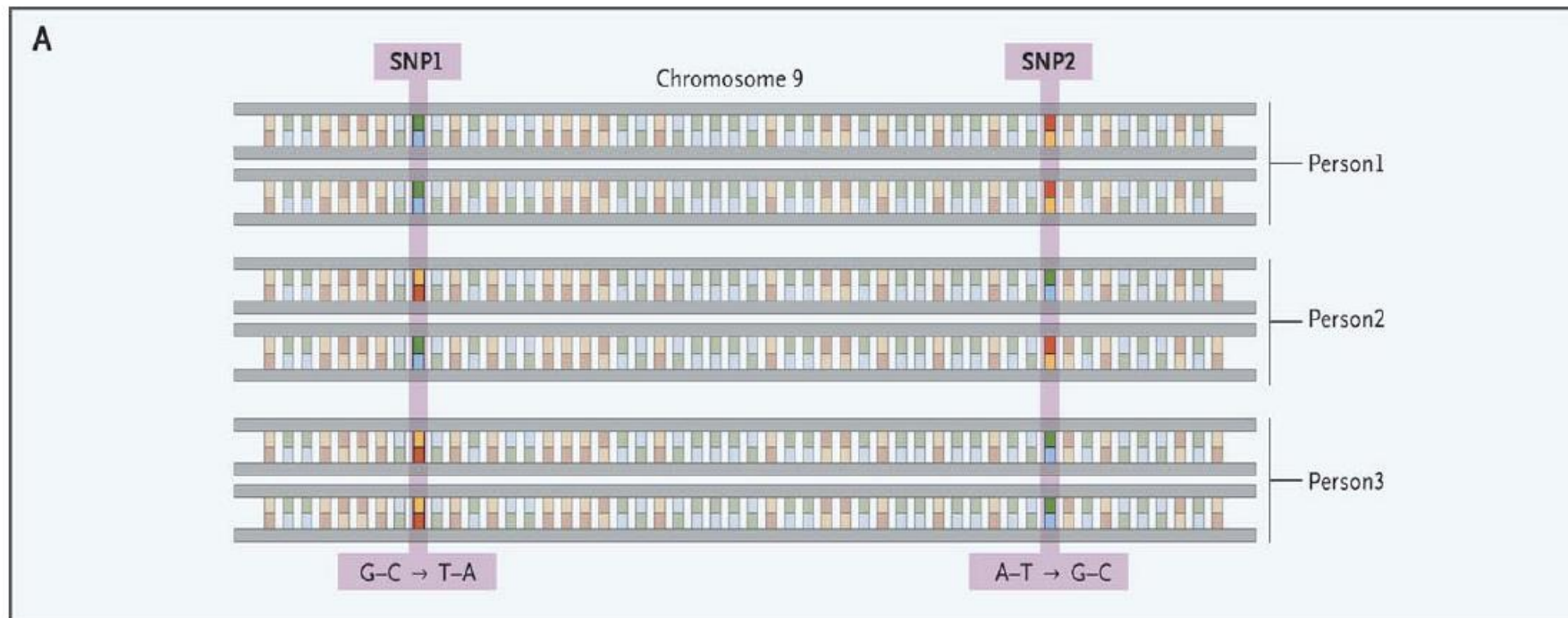
(slide Doug Brutlag 2010)

What are GWAs?

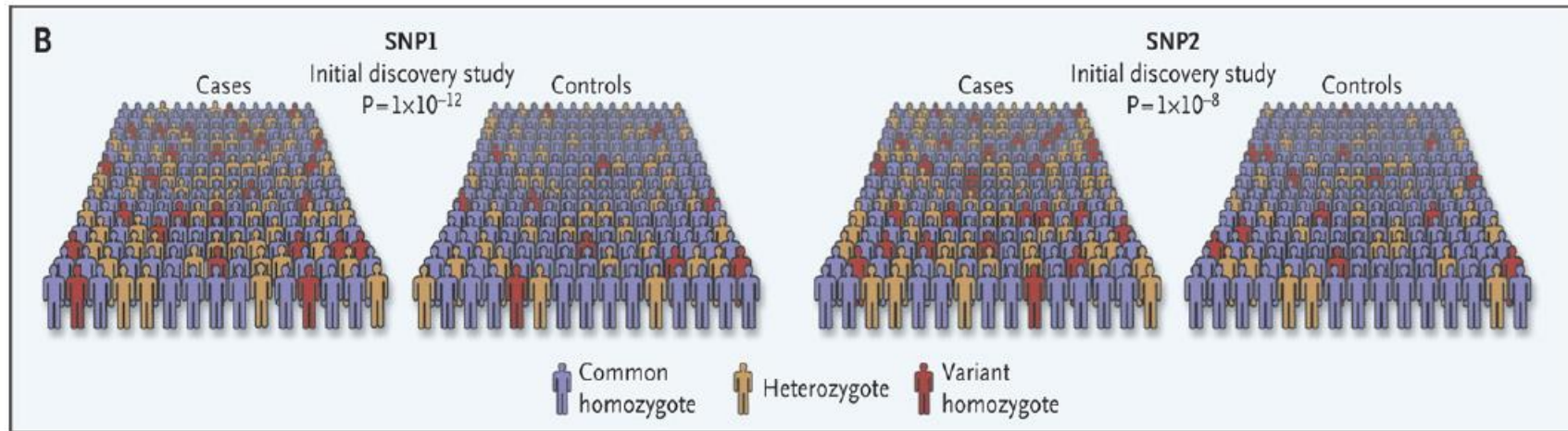
- A **genome-wide association study** is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular trait.
- **Recall:** a **trait** can be defined as a coded phenotype, a particular characteristic 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)



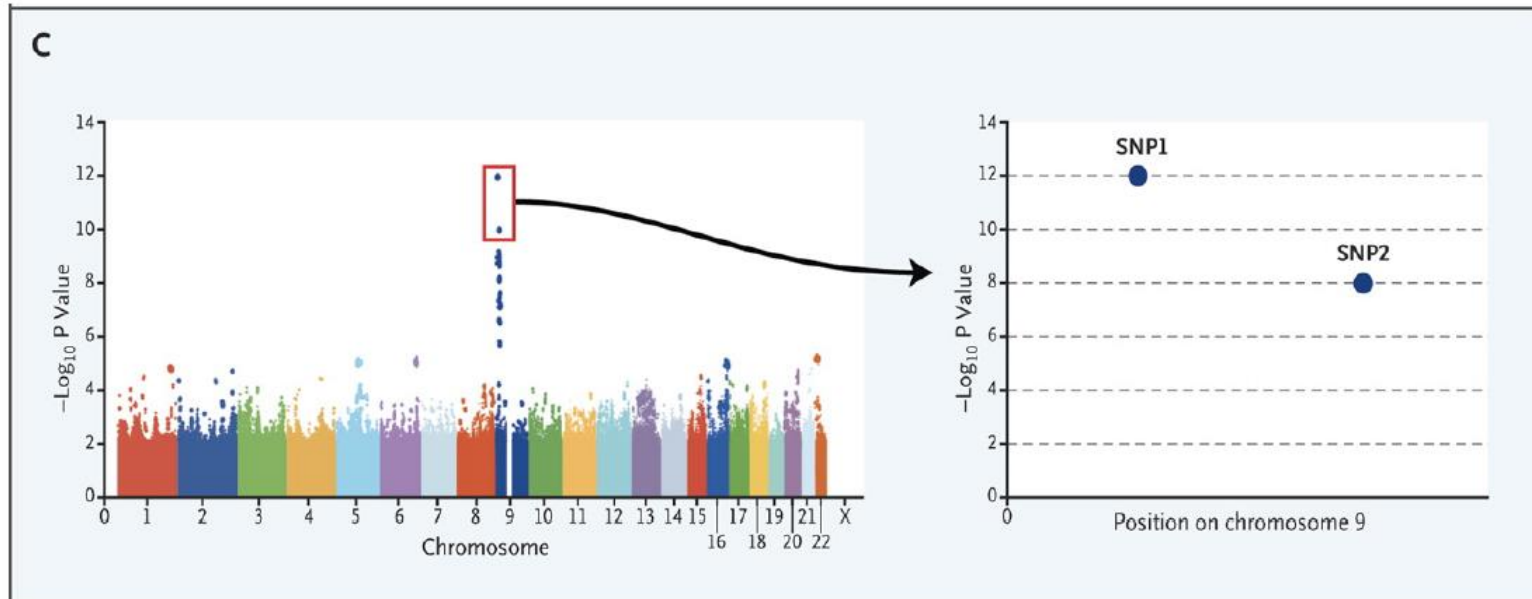
Genome-wide association studies: basic principles



- 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)

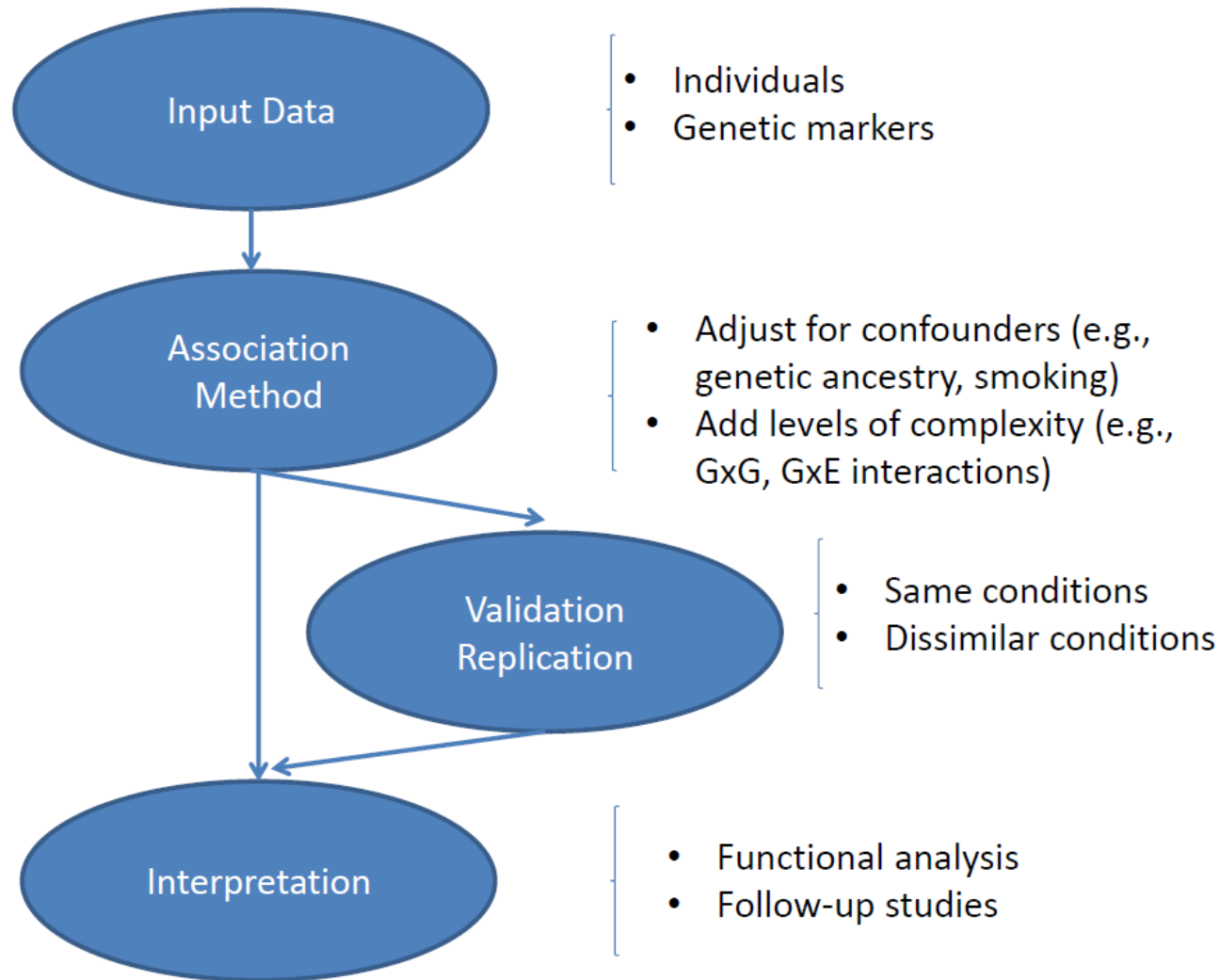
Genome-wide association studies: basic principles



- 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

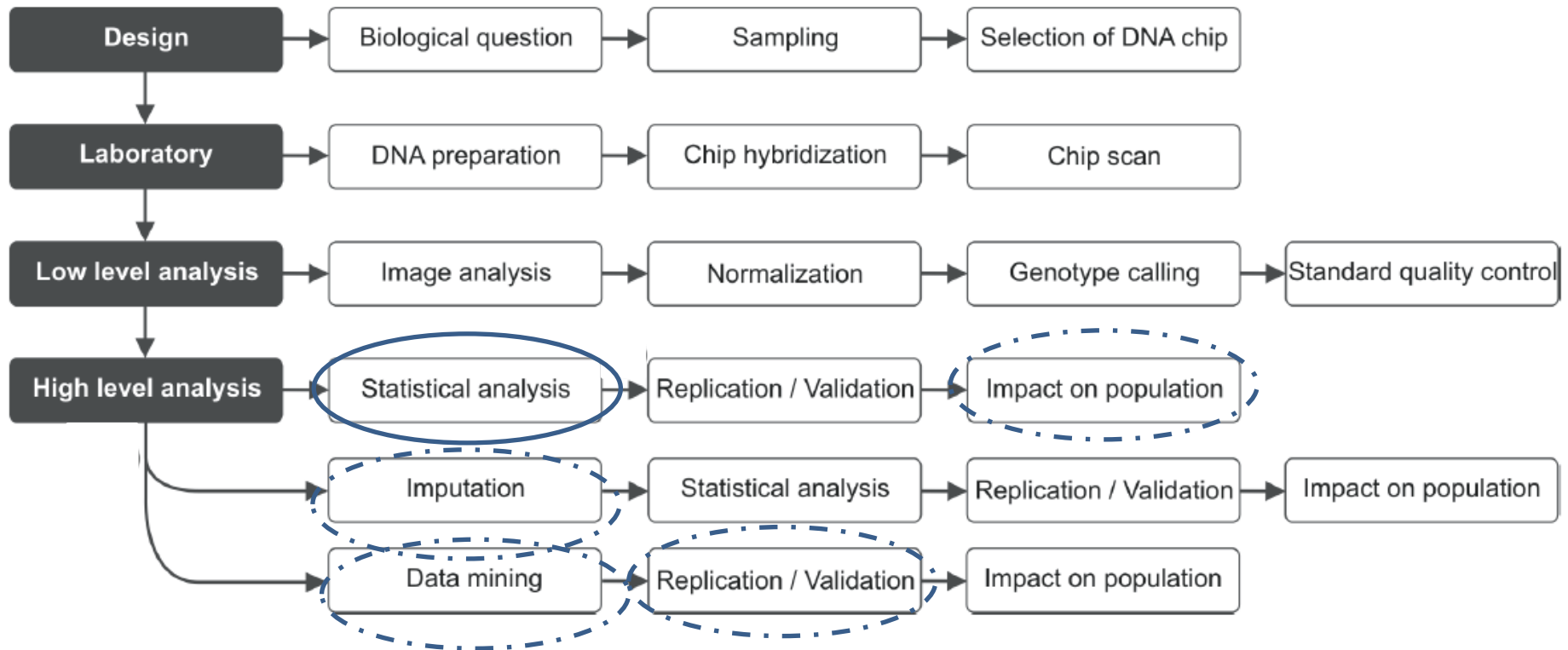


Genome-wide association studies: key components

- 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 <i>Diabetes Metab Res Rev</i> 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 individuals	559 Han Chinese ancestry individuals	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 <i>Int J Cancer</i> 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 <i>Neurobiol Aging</i> 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 Ripke², Aaron Isaacs¹ and Cornelia M. van Duijn¹

¹Department of Epidemiology and Biostatistics, Erasmus MC Rotterdam, Postbus 2040, 3000 CA Rotterdam, The Netherlands and ²Statistical 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

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

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 Li²

¹Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37232 and ²Department 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

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 Telenti³, Ioannis Xenarios¹
and Brian J. Stevenson^{1,2,*}

¹Swiss Institute of Bioinformatics, ²Ludwig Institute for Cancer Research, 1015 Lausanne and ³Institute 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

Open Access

Method

Bioconductor: open software development for computational biology and bioinformatics

Robert C Gentleman¹, Vincent J Carey², Douglas M Bates³, Ben Bolstad⁴, Marcel Dettling⁵, Sandrine Dudoit⁴, Byron Ellis⁶, Laurent Gautier⁷, Yongchao Ge⁸, Jeff Gentry¹, Kurt Hornik⁹, Torsten Hothorn¹⁰, Wolfgang Huber¹¹, Stefano Iacus¹², Rafael Irizarry¹³, Friedrich Leisch⁹, Cheng Li¹, Martin Maechler⁵, Anthony J Rossini¹⁴, Gunther Sawitzki¹⁵, Colin Smith¹⁶, Gordon Smyth¹⁷, Luke Tierney¹⁸, Jean YH Yang¹⁹ and Jianhua Zhang¹

Published: 15 September 2004

Genome Biology 2004, 5:R80

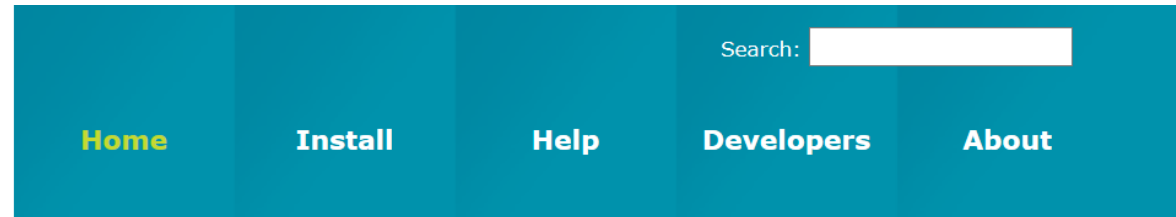
The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2004/5/10/R80>

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.



About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, [1024 software packages](#), and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker](#) images.

News

- Bioconductor [F1000 Research Channel](#) launched.
- Bioconductor [3.1](#) is available.
- Orchestrating high-throughput genomic analysis with *Bioconductor* ([abstract](#)) and other [recent literature](#)

Install »

Get started with *Bioconductor*

- [Install *Bioconductor*](#)
- [Explore packages](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn »

Master *Bioconductor* tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use »

Create bioinformatic solutions with *Bioconductor*

- [Software](#), [Annotation](#), and [Experiment](#) packages
- [Amazon Machine Image](#)
- [Latest release announcement](#)

Develop »

Contribute to *Bioconductor*

- [Use Bioc 'devel'](#)
- 'Devel' [Software](#), [Annotation](#) and [Experiment](#) packages
- [Package guidelines](#)
- [New package submission](#)

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

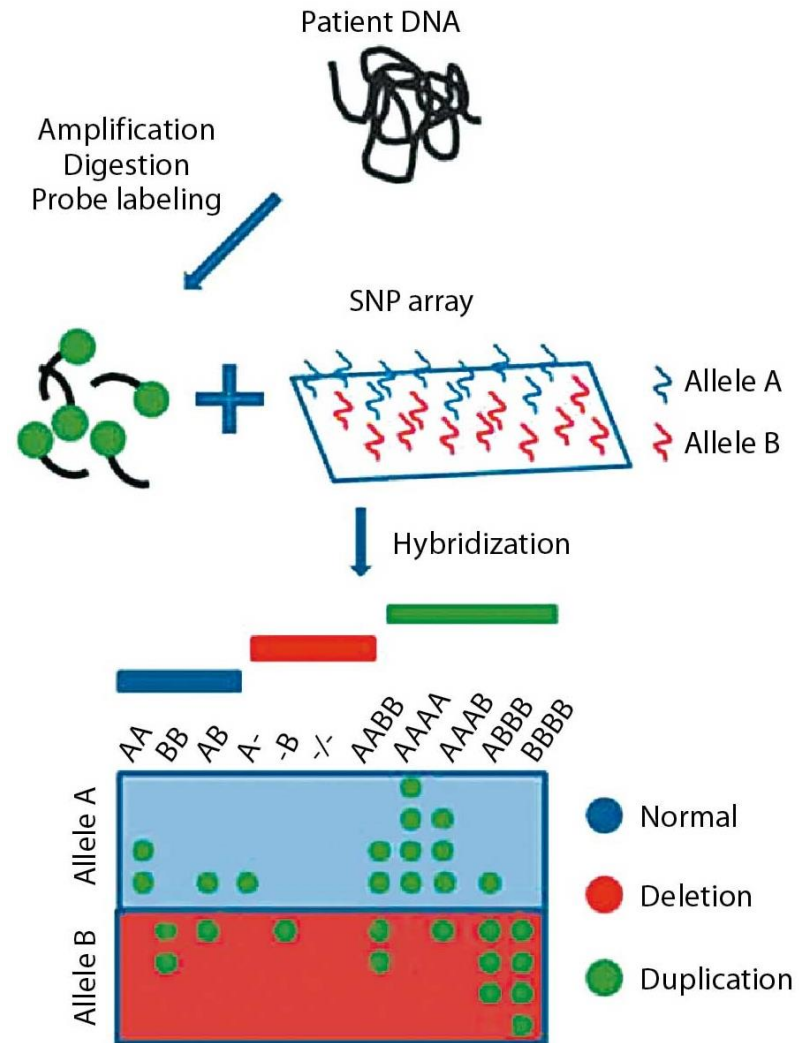
3 Study Design

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

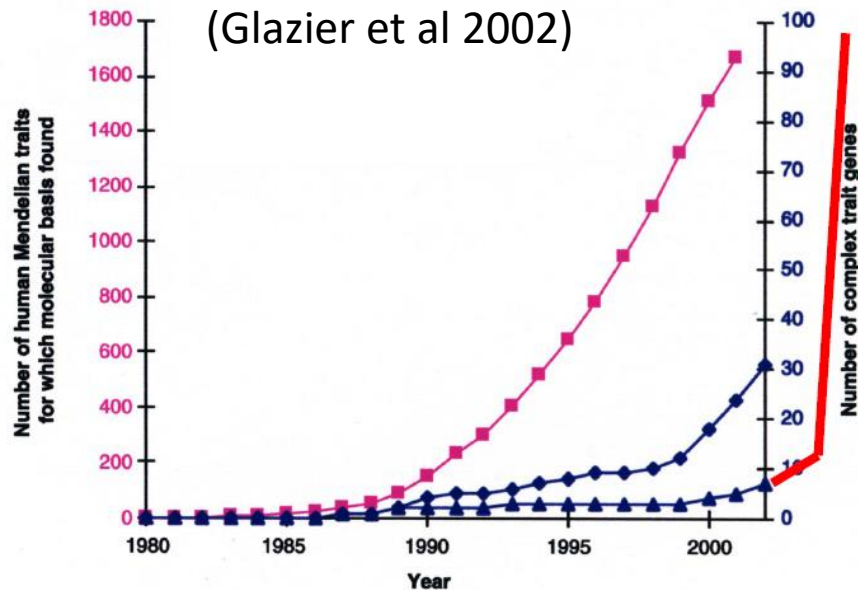
- Hypothesis 1 for GWAs: Common Disease – Common Variant (CDCV):
 - This hypothesis argues that **genetic variations with appreciable frequency** in the population at large, but **relatively low penetrance** (i.e. the probability that a carrier of the relevant variants will express the disease), are 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**



PINK : Human Mendelian traits

BLUE middle line : All complex traits

BLUE bottom line + red extension:
Human complex traits

Complex disease (definition):

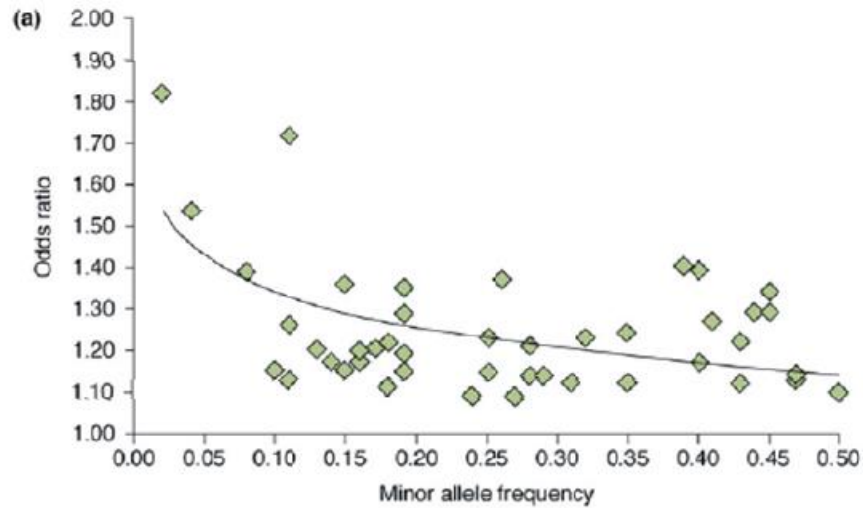
The term complex trait/disease refers to any phenotype that

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

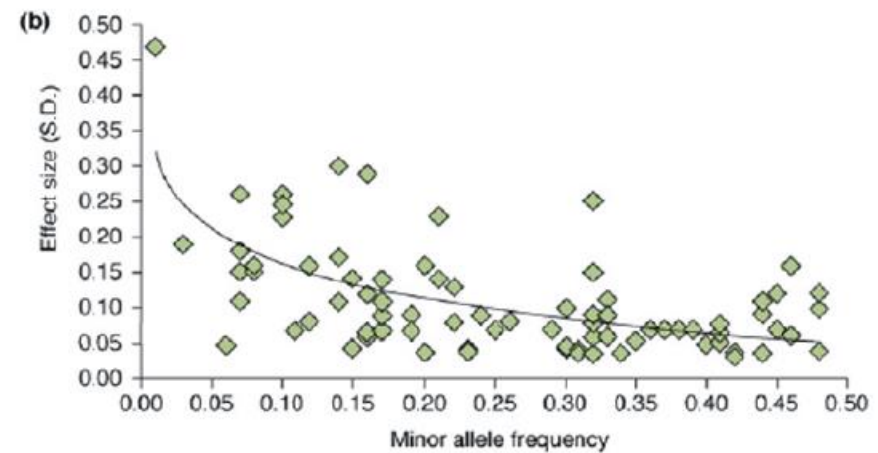
although they may exhibit familial tendencies (familial clustering, concordance among relatives).

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)

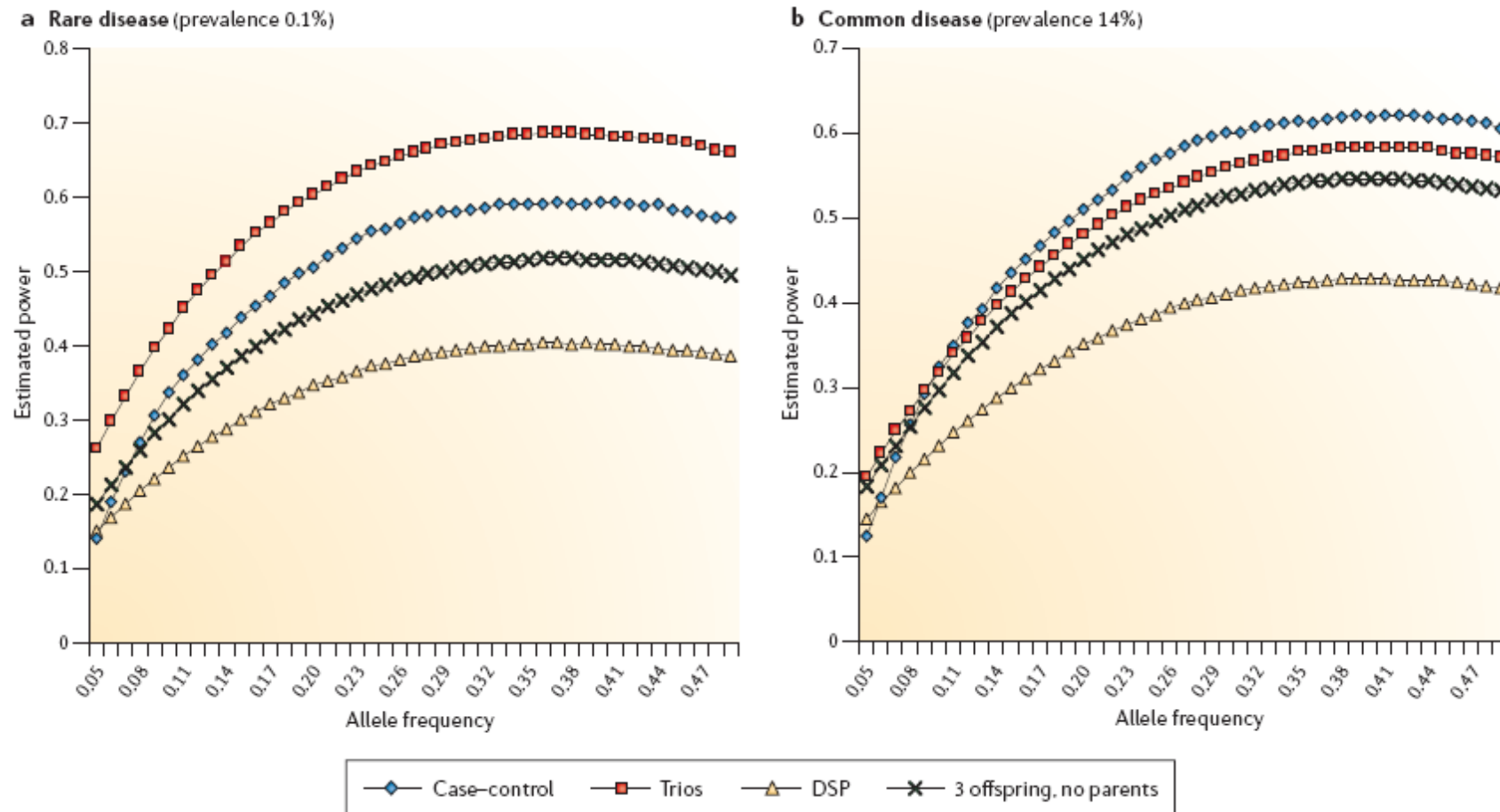
4. Simple logistics for diseases in children
5. Allows investigating imprinting (“bad allele” from father or mother?)

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

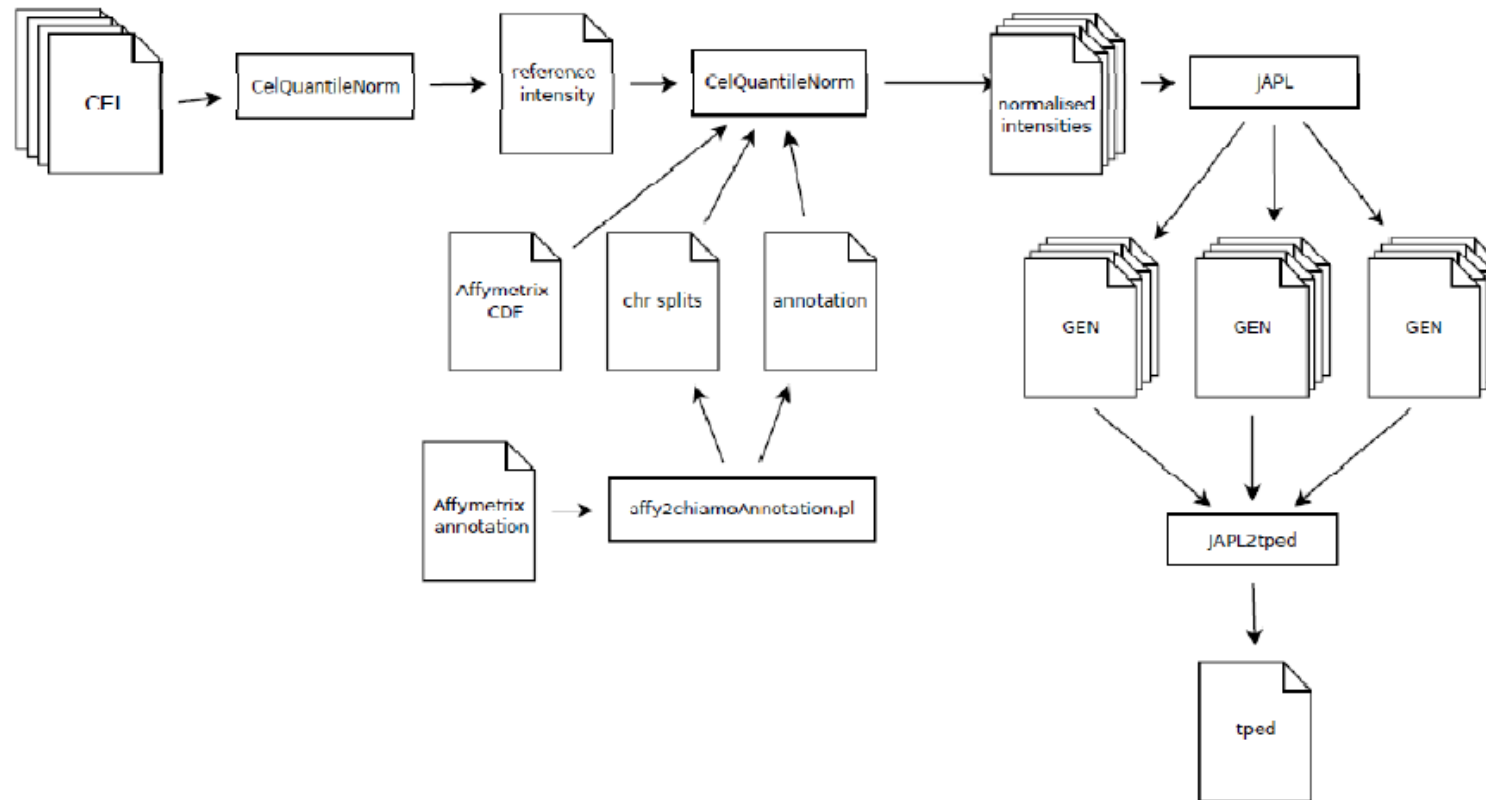
tfam file: First 6 columns of standard ped file

tped 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

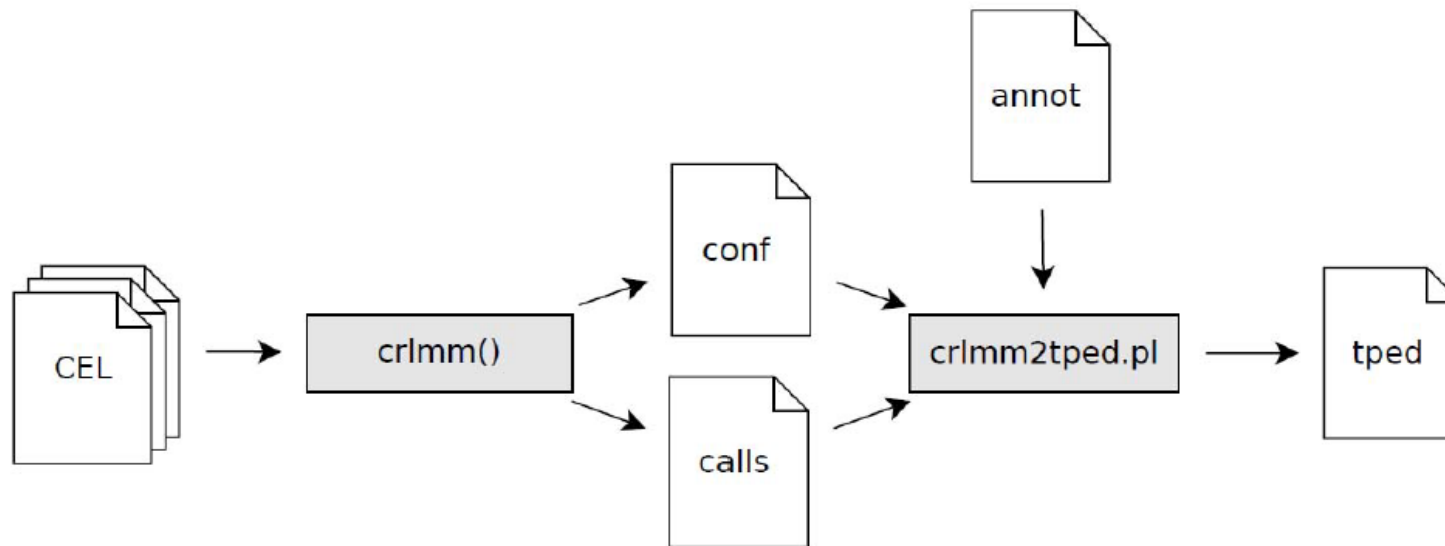
Note: data flow prior to analysis depends on calling algorithm



Genotype calling algorithm: JAPL

(Ziegler and Van Steen 2010)

Note: data flow prior to analysis depends on calling algorithm

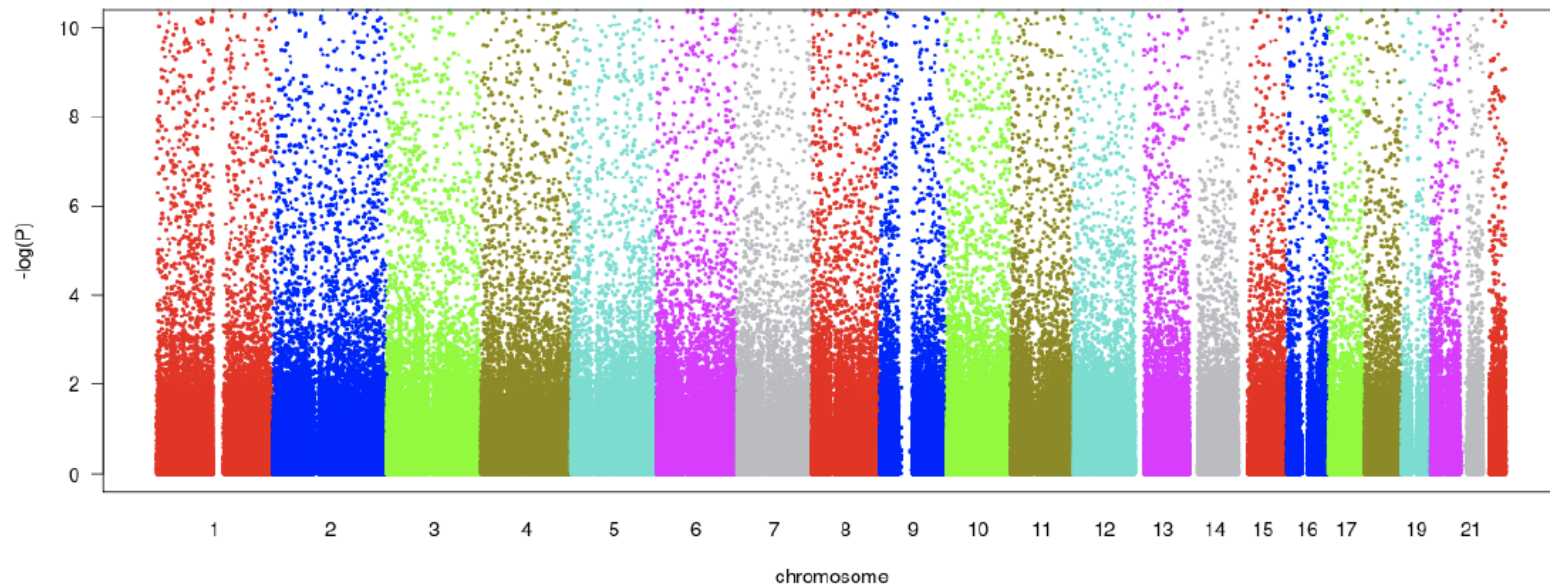


Genotype calling algorithm: CRLMM

(Ziegler and Van Steen 2010)

Why is quality control (QC) important?

BEFORE QC → true signals are lost in false positive signals

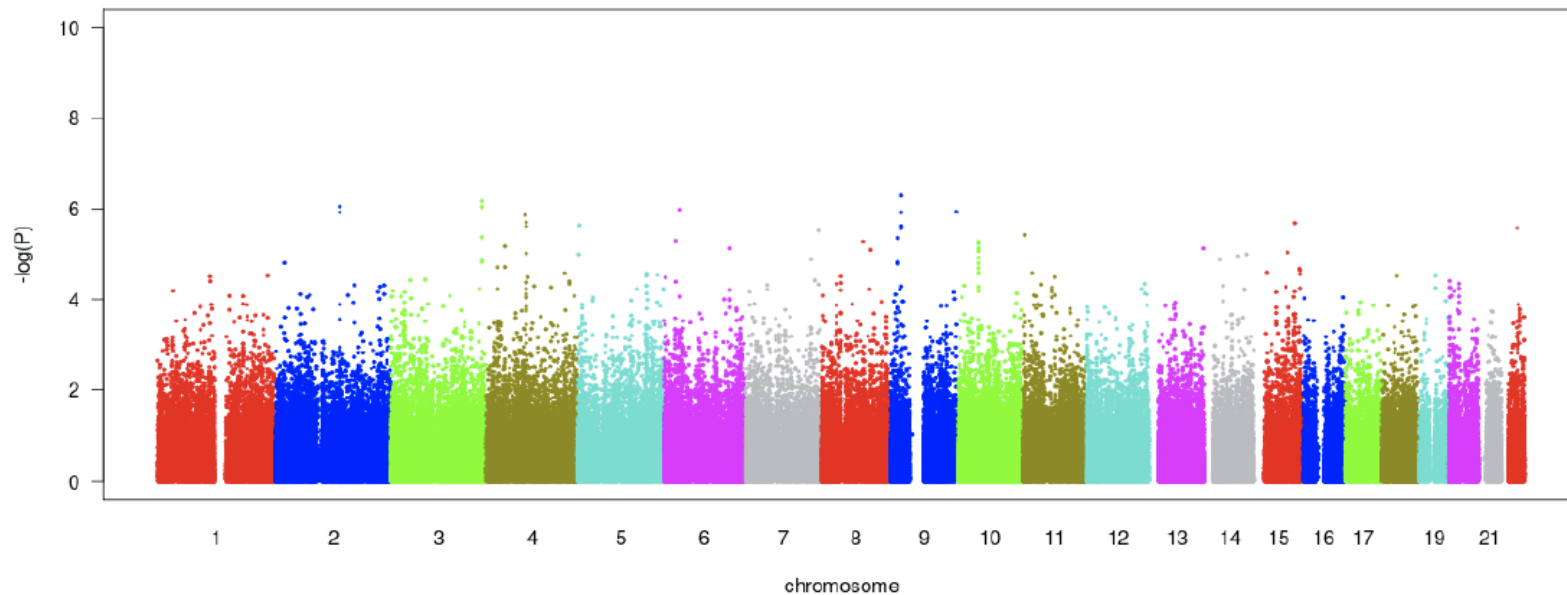


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(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

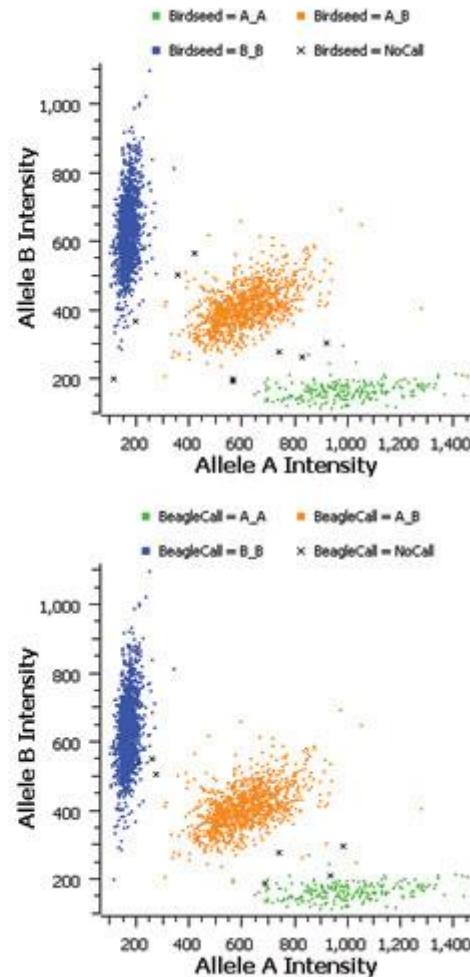
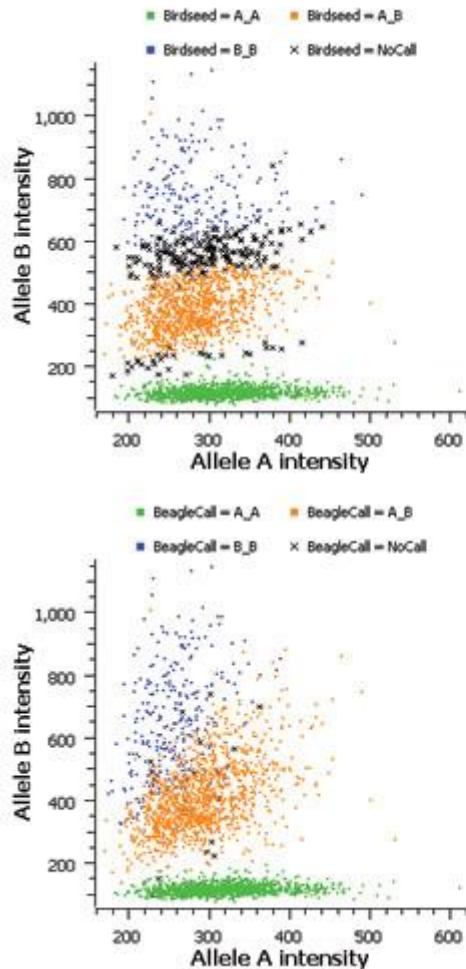
SNPs passing standard quality control: 270,701

(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)**

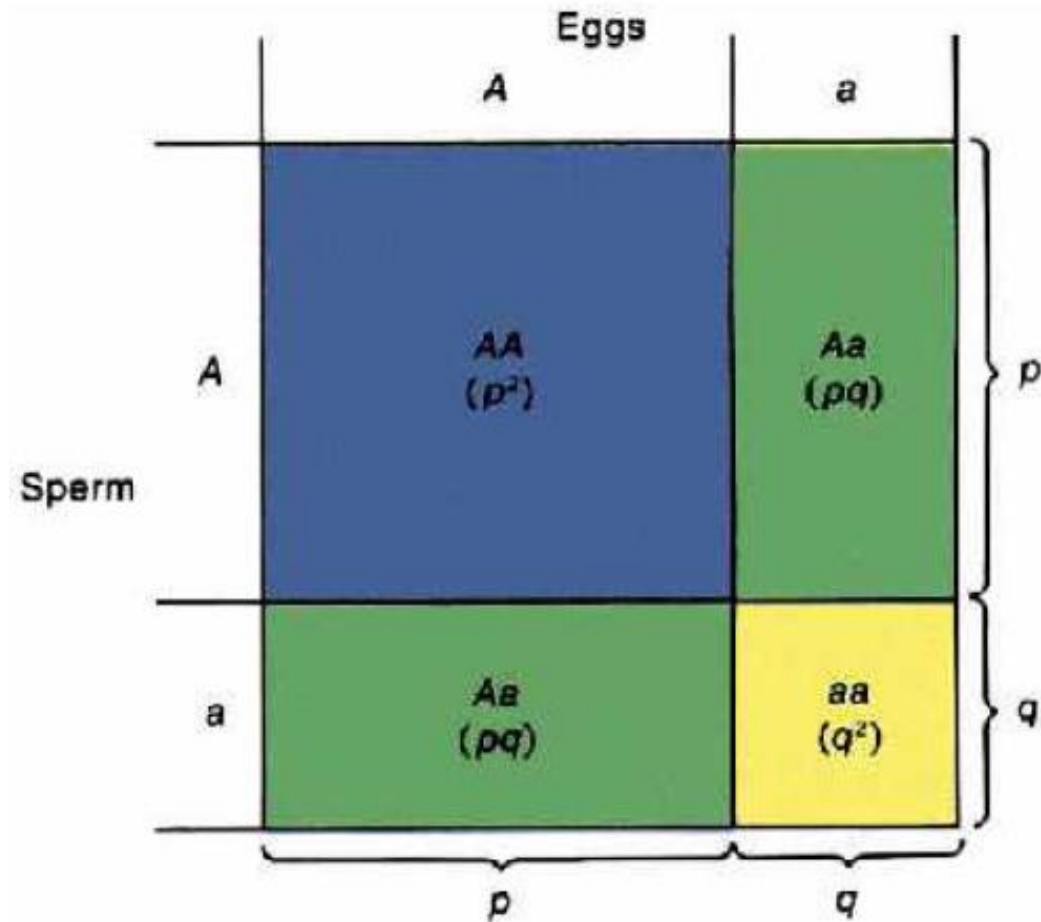
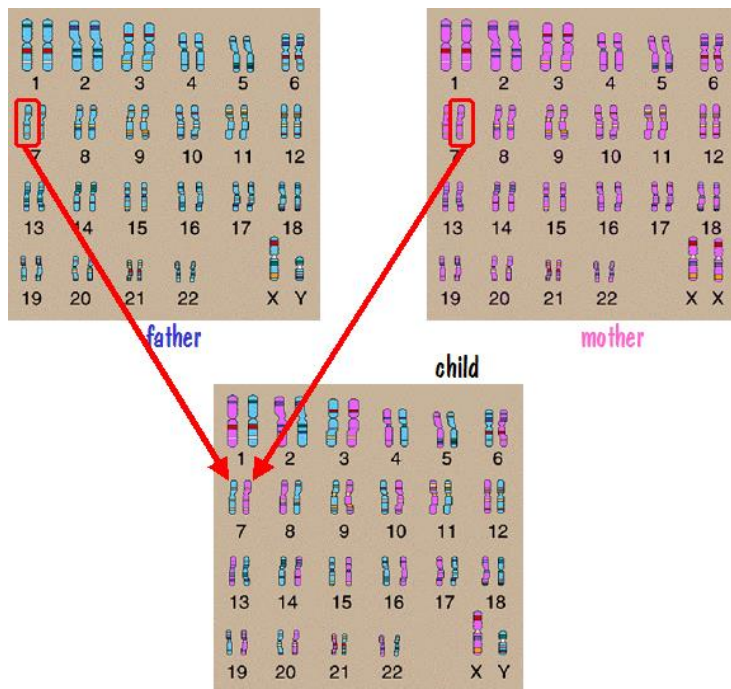
- 1 minus call rate
- MiF needs to be investigated separately in cases and controls because differential missingness may bias association results

- **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	$\geq 97\%$
	Cryptic relatedness	Study specific
	Ethnic origin	Study specific; visual inspection of principal components
	Heterozygosity	Mean \pm 3 std.dev. over all samples
	Heterozygosity by gender	Mean \pm 3 std.dev. within gender group
SNP level	MAF	$\geq 1\%$
	MiF	$\leq 2\%$ in any study group, e.g., in both cases and controls
	MiF by gender	$\leq 2\%$ 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

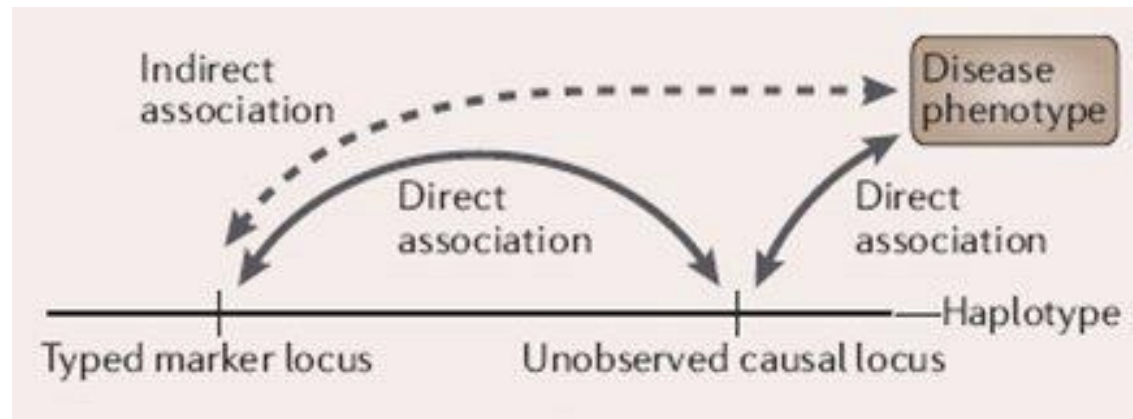
(Ziegler 2009)

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 rationale for performing genetic association studies



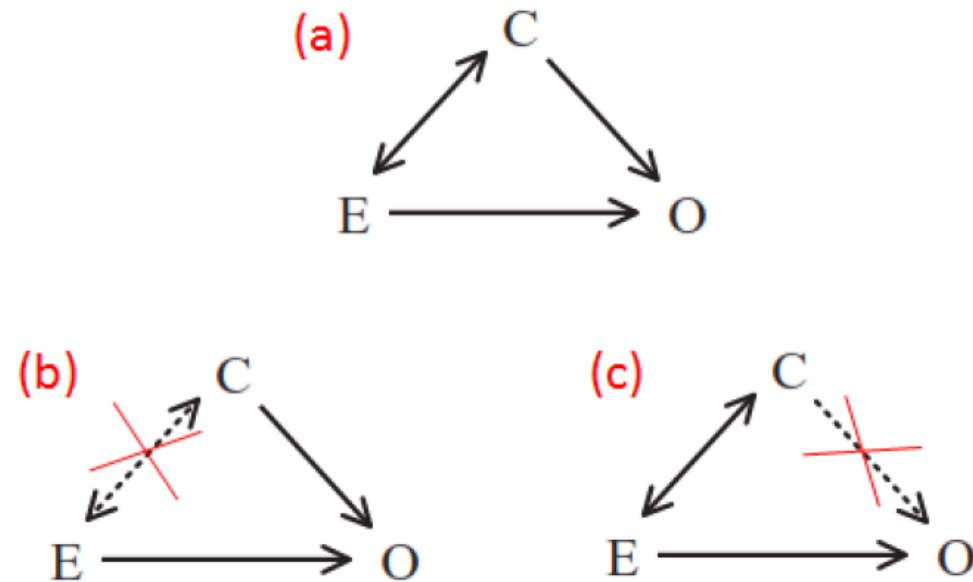
4.c Confounding by shared genetic ancestry – “population stratification”

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.

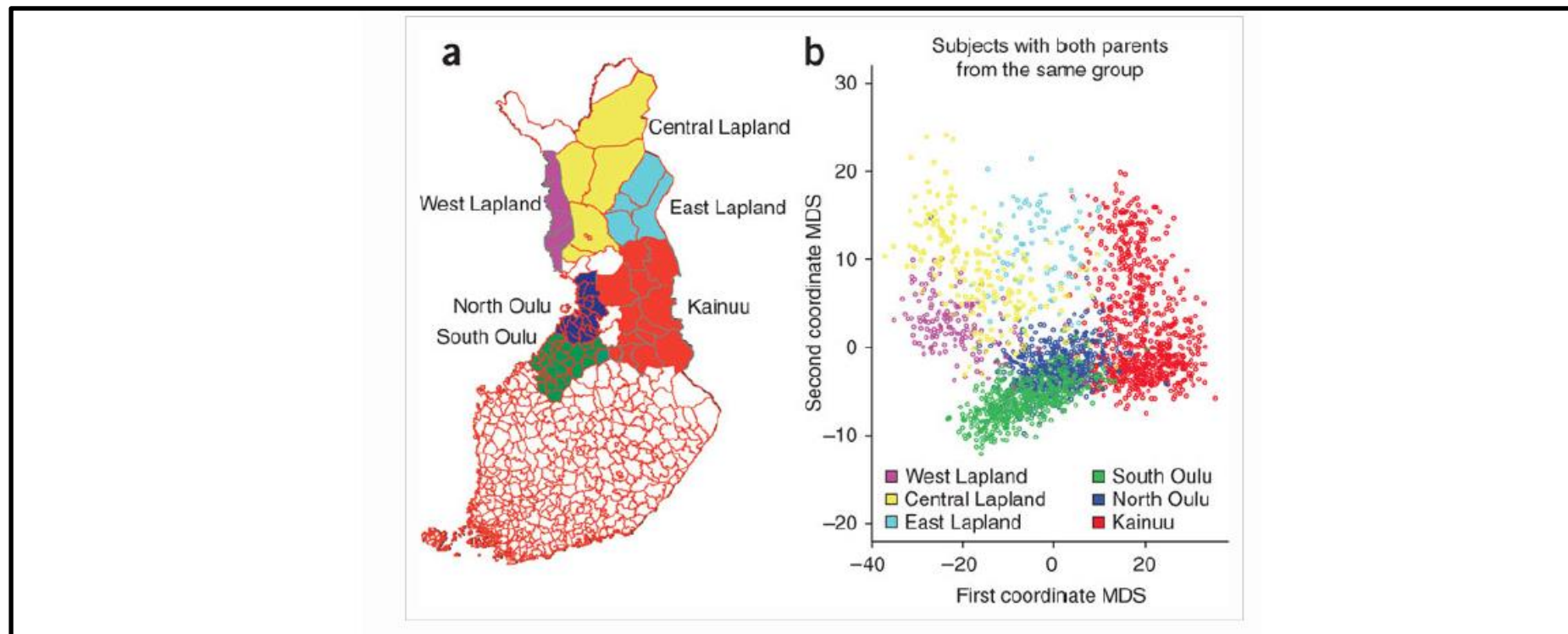


(Cois 2014)

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

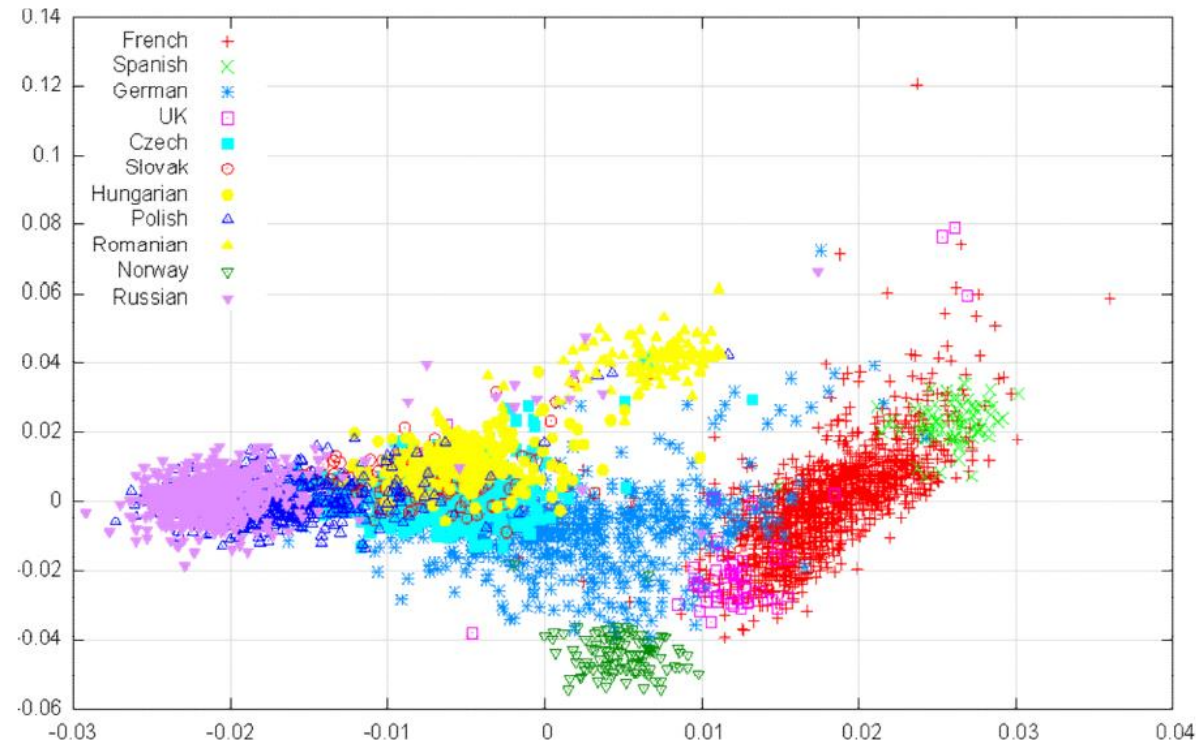
- There can be population structure in all populations, even those that appear to be relatively “homogeneous”



(Sabatti et al. 2009)

Confounding by shared genetic ancestry : creating a PC space

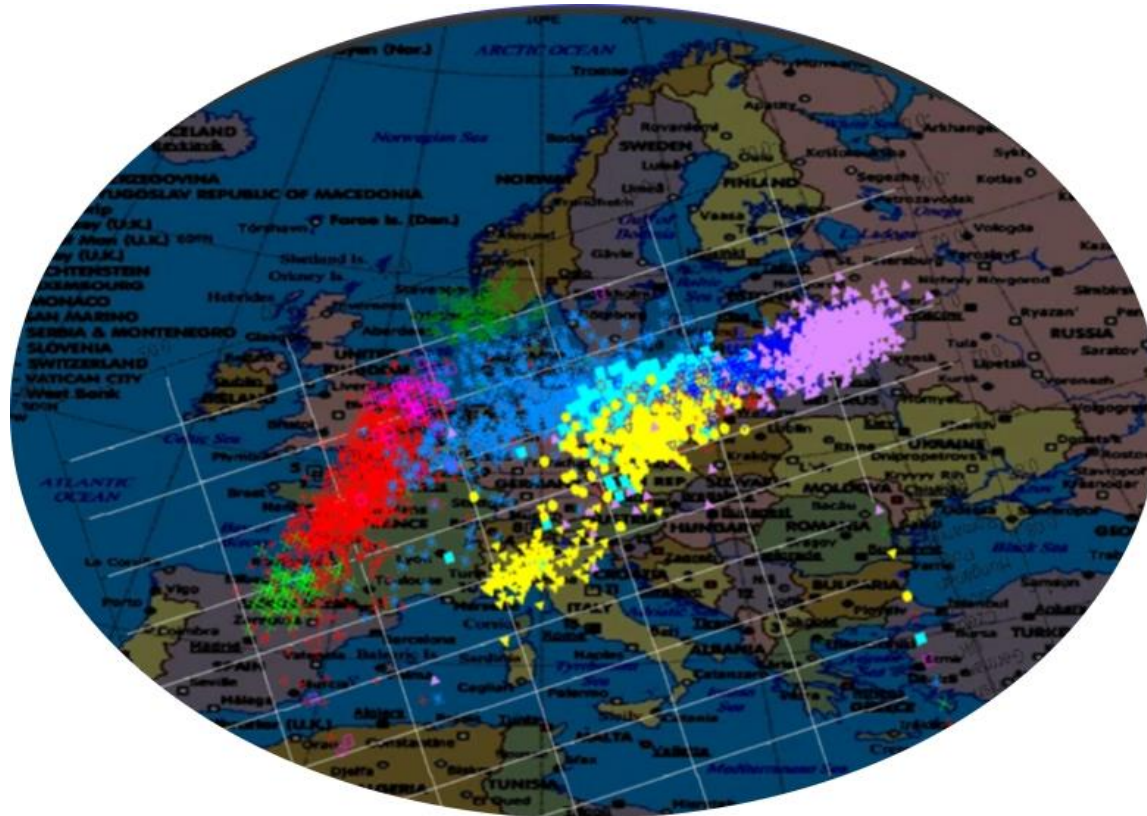
- In European data, the first 2 principal components “nicely” reflect the N-S and E-W axes !



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 Lynn, *Brigham and Women's Hospital, Harvard Medical School, Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115, USA*

Dawn L. DeMeo, *Brigham and Women's*

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's*

Alex MacGillivray, *Twin Research & Genetic Epidemiology Unit, St Thomas' Hospital, London SE17 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!



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, Kridsakorn 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

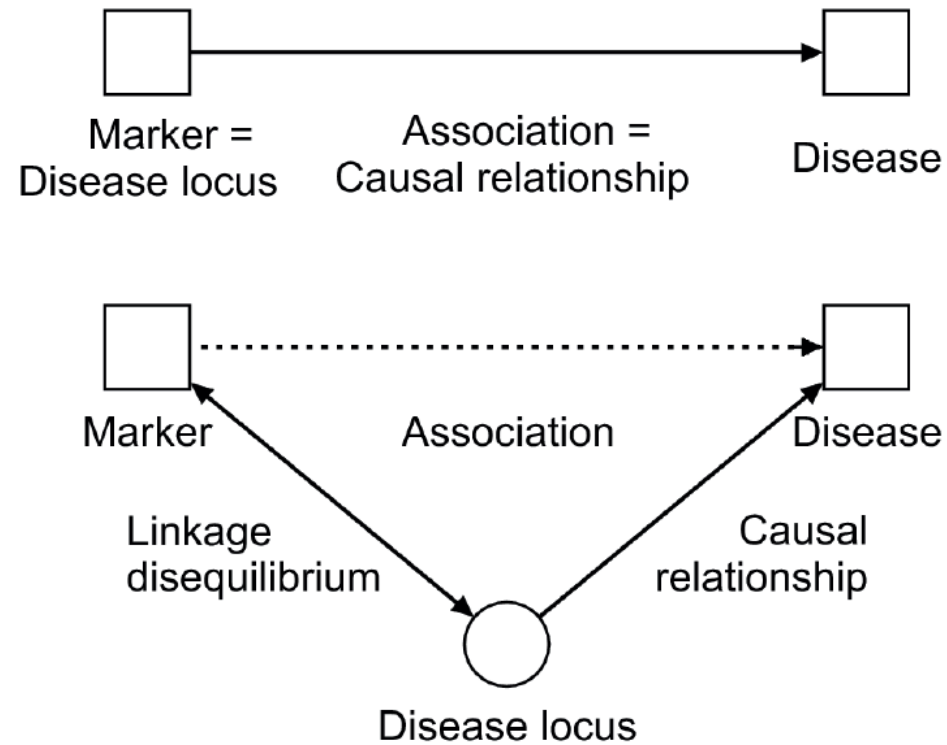
Abstract

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)

The linear regression model

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon$$

- y : response variable.
- x_1, \dots, x_k : regressor variables, independent variables.
- $\beta_0, \beta_1, \dots, \beta_k$: regression coefficients.
- ϵ : model error.
 - ▶ Uncorrelated: $\text{cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$.
 - ▶ Mean zero, Same variance: $\text{var}(\epsilon_i) = \sigma^2$. (homoscedasticity)
 - ▶ Normally distributed.

Linear vs non-linear

Linear Models Examples:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

$$y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \epsilon$$

$$\log y = \beta_0 + \beta_1 \left(\frac{1}{x_1} \right) + \beta_2 \left(\frac{1}{x_2} \right) + \epsilon$$

Nonlinear Models Examples:

$$y = \beta_0 + \beta_1 x_1^{\gamma_1} + \beta_2 x_2^{\gamma_2} + \epsilon$$

$$y = \frac{\beta_0}{1 + e^{\beta_1 x_1}} + \epsilon$$

Regression inference

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon$$

- Least square estimation of the regression coefficients.
 $b = (X^T X)^{-1} X^T y.$
- Variance estimation for σ^2 (see later)
- Coefficient of Determination. R^2 .
- Partial F test or t-test for $H_0 : \beta_j = 0$.

What is R-squared?

- R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the **coefficient of determination, or the coefficient of multiple determination for multiple regression**.
- The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model:

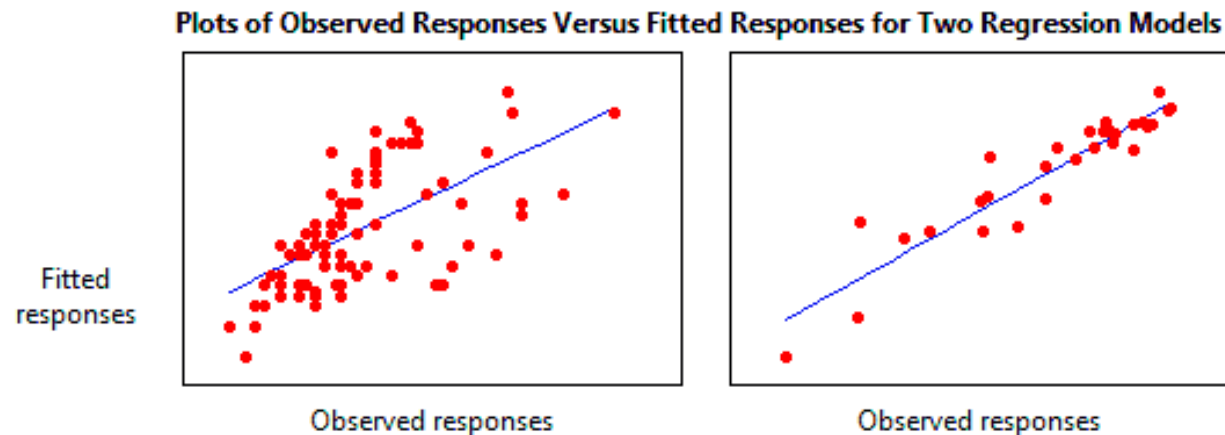
$$\text{R-squared} = \text{Explained variation} / \text{Total variation}$$

(compare with well-known formula for $\text{cor}(X,Y)$)

- R-squared is always between 0 and 100%:
 - 0% indicates that the model explains none of the variability of the response data around its mean; 100% indicates that the model explains all the variability of the response data around its mean.

Graphical representation of R-squared

- Plotting fitted values by observed values graphically illustrates different R-squared values for regression models.



- The regression model on the left accounts for 38.0% of the variance while the one on the right accounts for 87.4%. The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line.

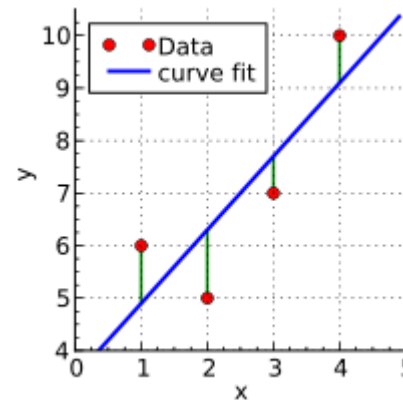
General linear test approach

- The full model (continuous response, say “BMI”):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Fit the model by f.i. the method of least squares (this leads to estimations b for the beta parameters in the model)
- It will also lead to the **error sums of squares** (SSE): the sum of the squared deviations of each observation Y around its estimated expected value
- The error sums of squares of the full model $SSE(F)$:

$$\begin{aligned} \sum [Y - b_0 - b_1 X_1 - b_2 X_2]^2 \\ = \sum (Y - \hat{Y})^2 \end{aligned}$$



General linear test approach

- Next we consider a null hypothesis H_0 of interest:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

- The model when H_0 holds is called **the reduced or restricted model**. When $\beta_1 = 0$, then the regression model reduces to

$$Y = \beta_0 + \beta_2 X_2 + \varepsilon$$

- Again we can fit this model with f.i. the least squares method and obtain an error sums of squares, now for the reduced model: $SSE(R)$

Which error sums of squares will be smaller? $SSE(F)$ or $SSE(R)$

General linear test approach

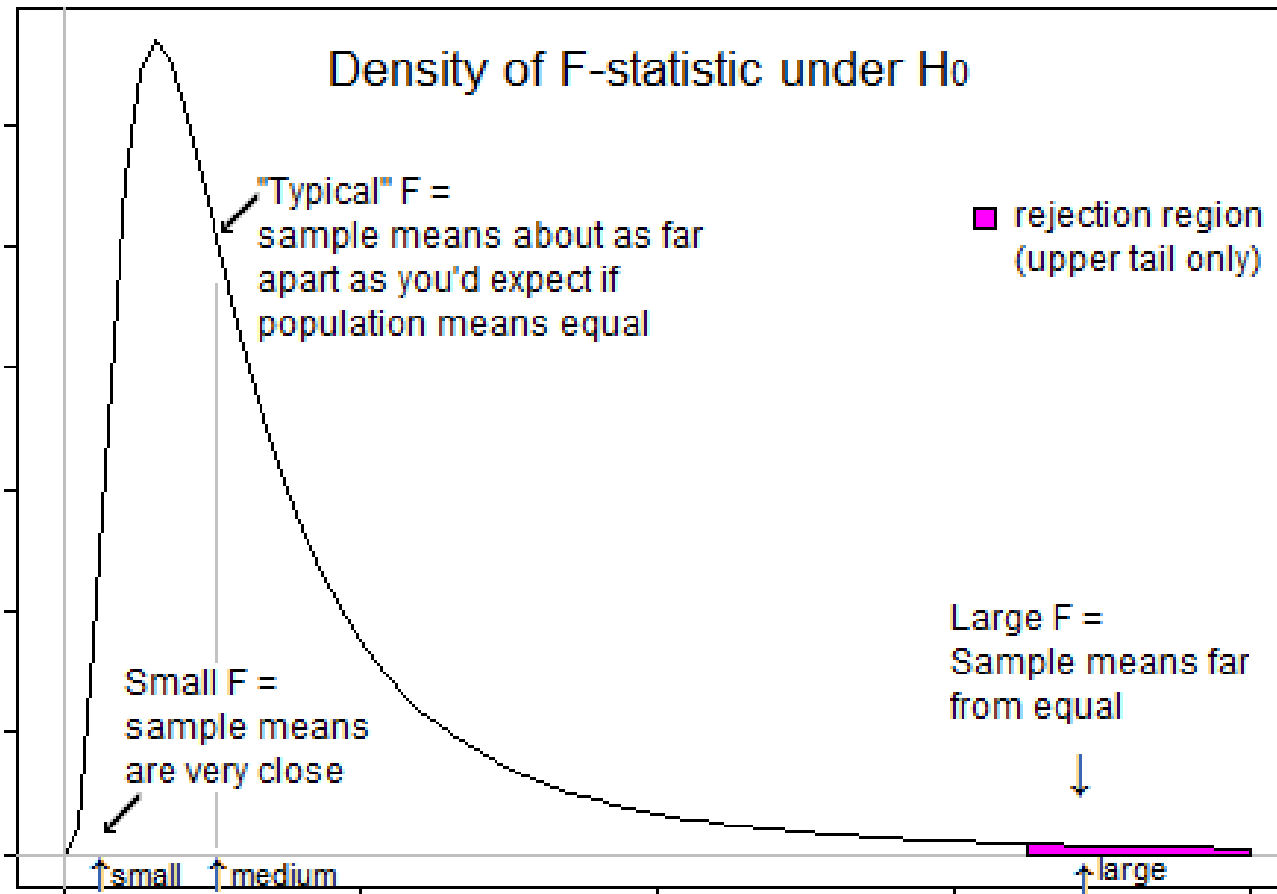
- The logic now is to compare both SSEs. The actual test statistic is a function of $SSE(R)$ - $SSE(F)$:

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

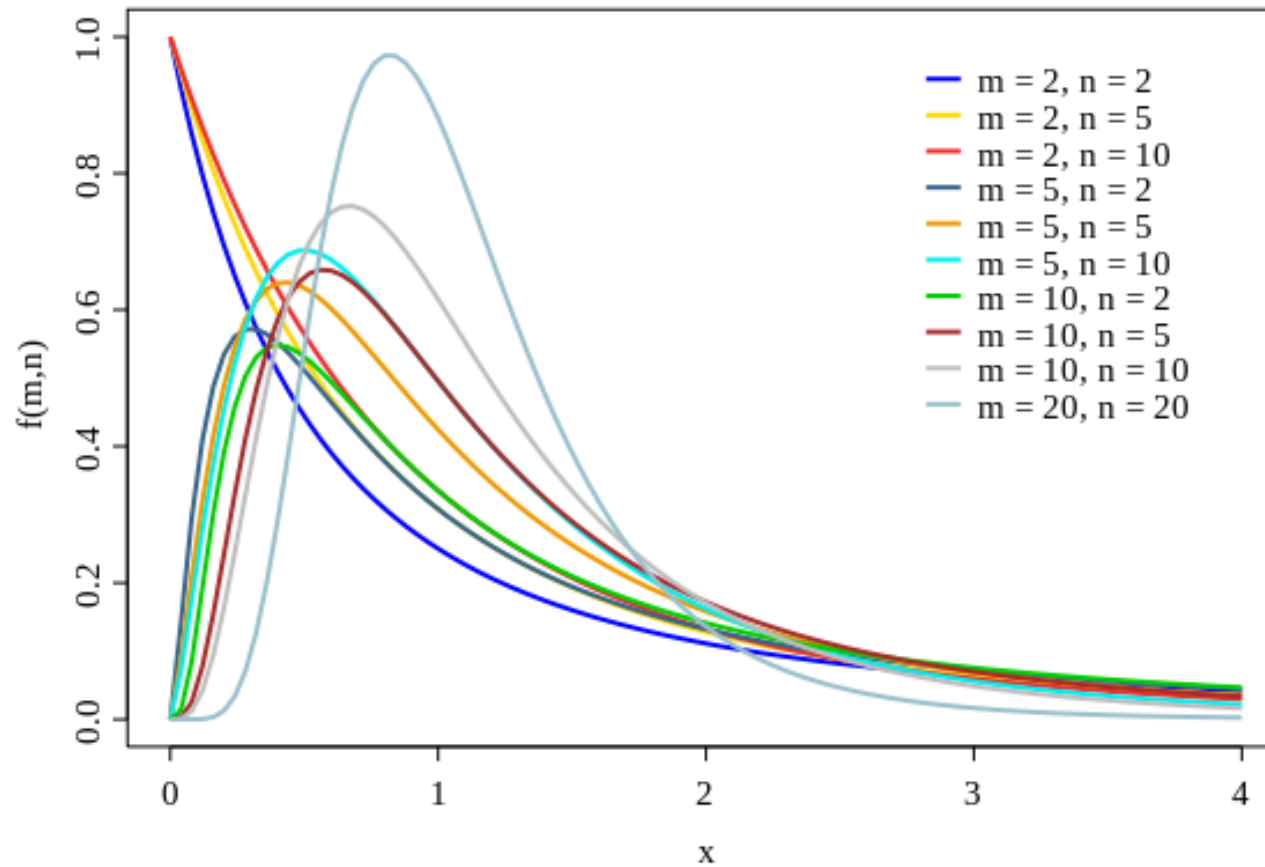
which follows an F distribution when H_0 holds

- The decision rule (for a given alpha level of significance) is:
If $F^* \leq F(1 - \alpha; df_R - df_F, df_F)$, you cannot reject H_0
If $F^* > F(1 - \alpha; df_R - df_F, df_F)$, conclude H_1

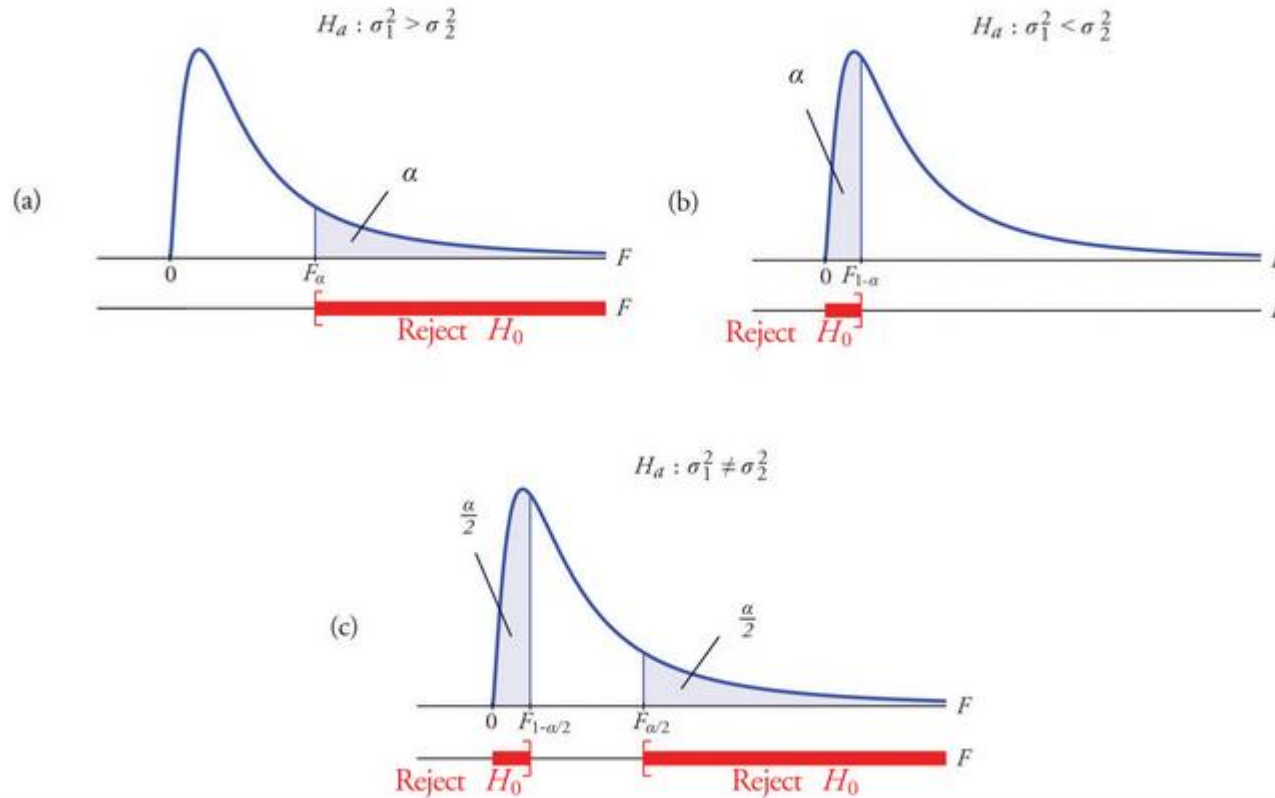
Recall: rejection and non-rejection regions



What are different shapes of an F distribution (parameters m, n)?



Why F test?



Terminology	Alternative Hypothesis	Rejection Region
Right-tailed	$H_a : \sigma_1^2 > \sigma_2^2$	$F \geq F_\alpha$
Left-tailed	$H_a : \sigma_1^2 < \sigma_2^2$	$F \leq F_{1-\alpha}$
Two-tailed	$H_a : \sigma_1^2 \neq \sigma_2^2$	$F \leq F_{1-\alpha/2}$ or $F \geq F_{\alpha/2}$

Special cases of an F distribution

normal distribution = $F(1, \infty)$

t distribution = $F(1, n_2)$

chi-square
distribution = $F(n_1, \infty)$

Tests in GWAS using the regression framework

• Example 1:

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

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

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

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

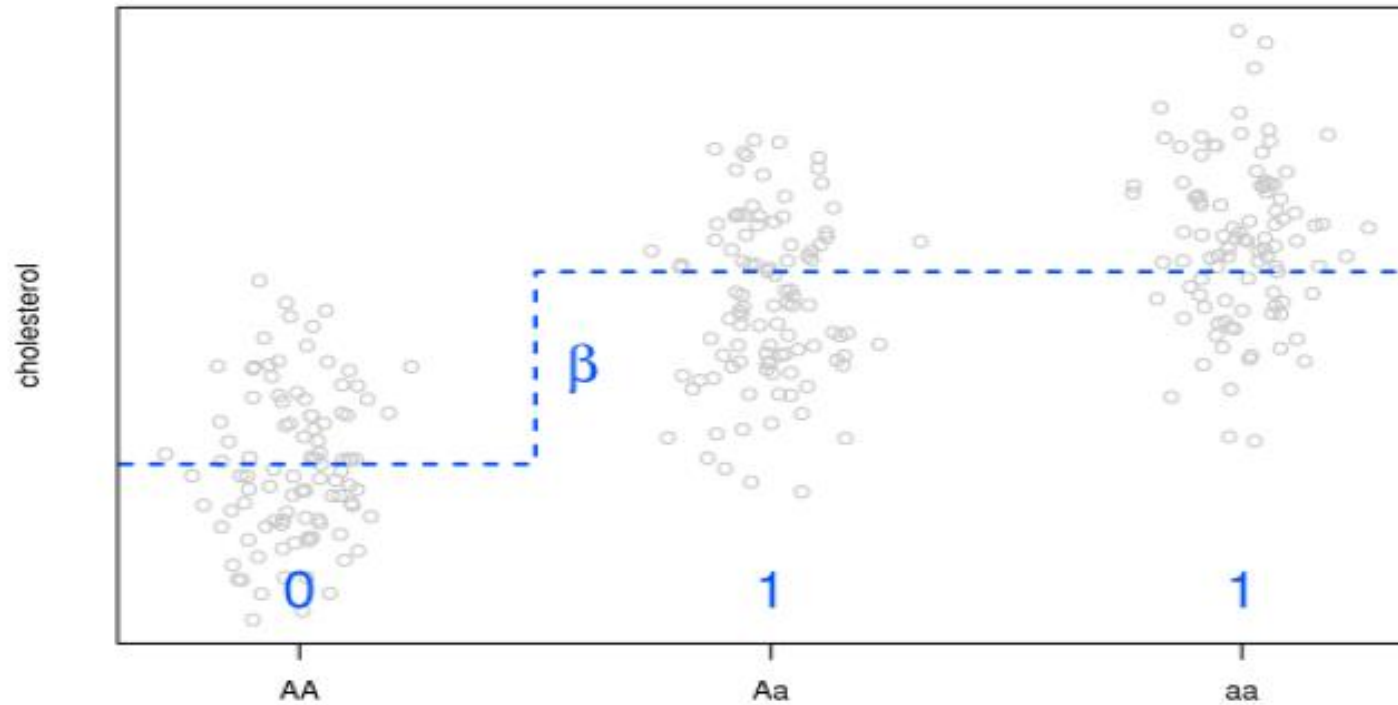
How many dfs would the corresponding F-test have?

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

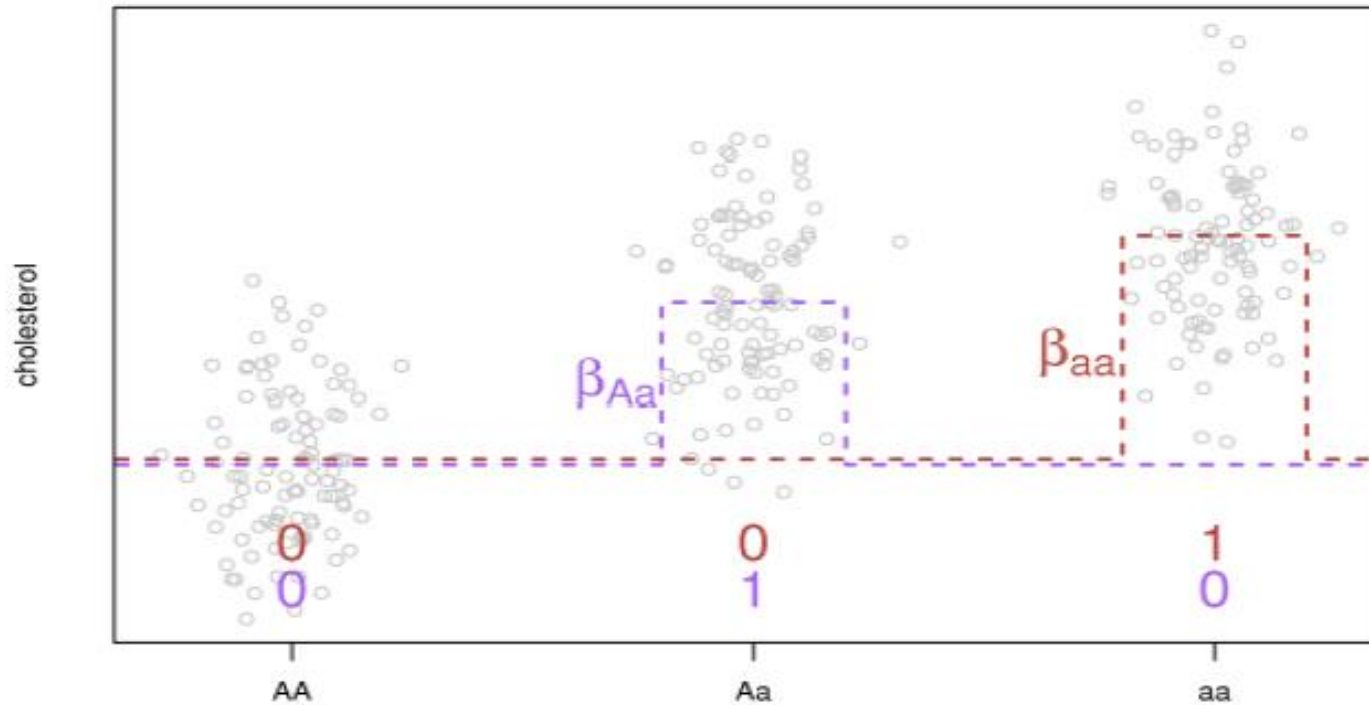
The impact of different encoding schemes for SNPs

	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?

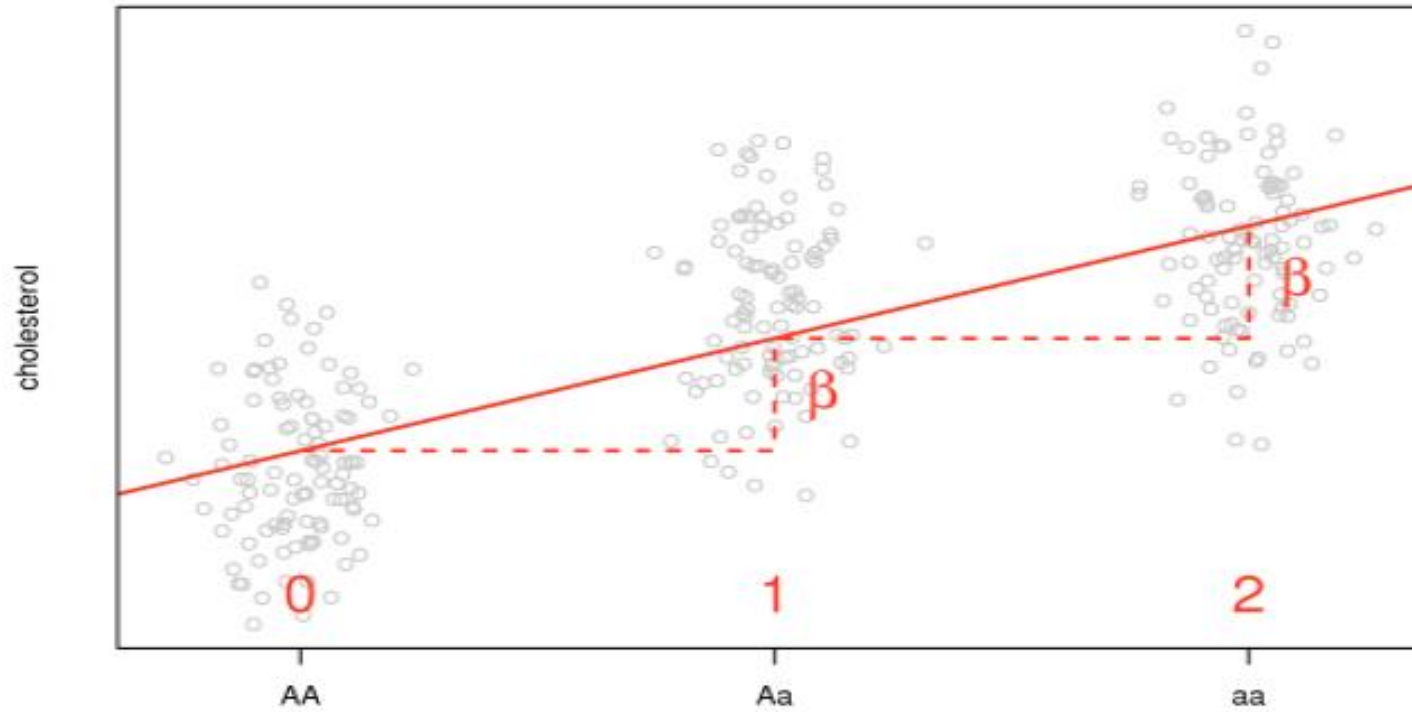


Which encoding scheme provides a good fit to the data?



Robust vs overkill ?

Which encoding scheme provides a good fit to the data?



Most commonly used

Regression analysis in R

Syntax	Model	Comments
$Y \sim A$	$Y = \beta_0 + \beta_1 A$	Straight-line with an implicit y-intercept
$Y \sim -1 + A$	$Y = \beta_1 A$	Straight-line with no y-intercept; that is, a fit forced through (0,0)
$Y \sim A + I(A^2)$	$Y = \beta_0 + \beta_1 A + \beta_2 A^2$	Polynomial model; note that the identity function $I()$ allows terms in the model to include normal mathematical symbols.
$Y \sim A + B$	$Y = \beta_0 + \beta_1 A + \beta_2 B$	A first-order model in A and B without interaction terms.
$Y \sim A:B$	$Y = \beta_0 + \beta_1 AB$	A model containing only first-order interactions between A and B.
$Y \sim A*B$	$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB$	A full first-order model with a term; an equivalent code is $Y \sim A + B + A:B$.
$Y \sim (A + B + C)^2$	$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 BC$	A model including all first-order effects and interactions up to the n^{th} order, where n is given by $()^n$. An equivalent code in this case is $Y \sim A*B*C - A:B:C$.

Special cases of an F distribution

normal distribution = $F(1, \text{infinite})$

t distribution = $F(1, n_2)$

chi-square
distribution = $F(n_1, \text{infinite})$

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/riagnostics.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.

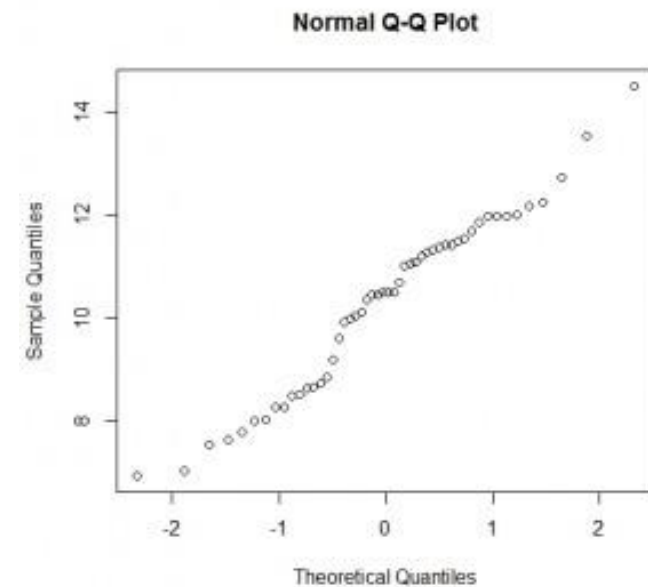
The number of quantiles is selected to match the size of your sample data.

The quantile function in R offers 9 different quantile algorithms!

See `help(quantile)`

QQ plots for model diagnostics – Q for Quantile

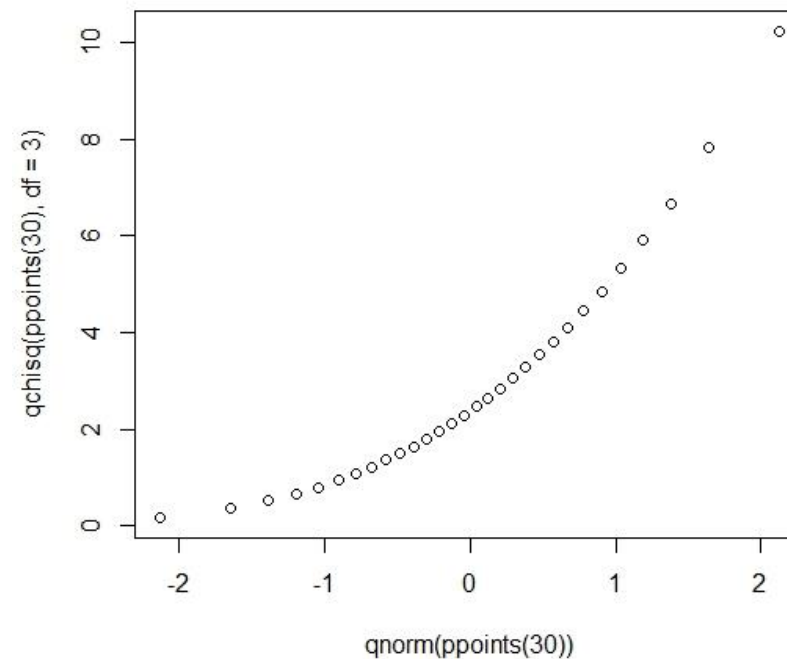
- A Q-Q plot is a scatterplot created by plotting **two sets of quantiles** against one another.
- If both sets of quantiles come from the same distribution, we should see the points forming a line that's roughly straight.
- Here's an example of a Normal Q-Q plot when both sets of quantiles truly come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



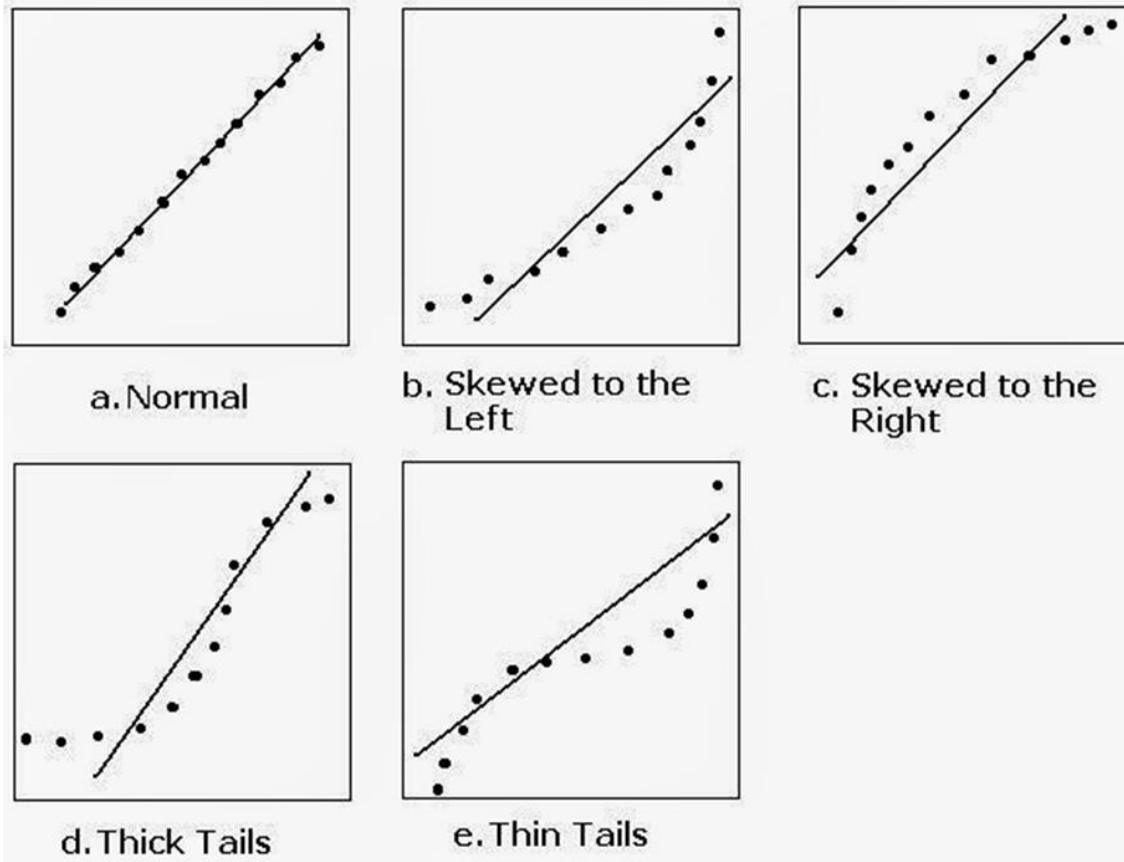
Examples of QQ plots: no straight line

- 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



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

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

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

What is the df ?

The flexible regression framework

Instead of

$$Y = \beta_0 + \beta_1 SNP + \varepsilon; Y \text{ continuous}$$

and modelling

$$E[Y|SNP] = \beta_0 + \beta_1 SNP \text{ (without error term!)}$$

consider

$\beta_0 + \beta_1 SNP = \boldsymbol{\eta}$ representing the linear combination as it can never be equal to a binary variable (0/1 response; control/case status)

and model

$$\boldsymbol{g}(E[Y|SNP]) = \beta_0 + \beta_1 SNP = \boldsymbol{\eta}$$

where $\boldsymbol{g}()$ is called a **link function** between response and linear predictor

and thus

$$E[Y|SNP] = \boldsymbol{g_inv}(\boldsymbol{\eta})$$

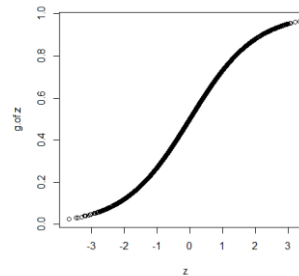
For a binary trait Y:

$$E[Y|SNP] = Prob(Y = 1|SNP) \\ = \frac{\exp(\eta)}{(1 + \exp(\eta))} = \frac{1}{(1 + \exp(-\eta))} = g_inv(\eta)$$

where

g_inv is the **logistic function (sigmoid function)**
(squashing the linear predictor to an acceptable range)

```
> Z <- rnorm(10000)
> ginv.of.Z <- (1/(1+exp(-Z)))
> plot(Z,ginv.of.Z)
```



Since

$$Prob(Y = 1|SNP) = \frac{\exp(\eta)}{(1+\exp(\eta))}$$

we have

$$\frac{Prob(Y = 1|SNP)}{1 - Prob(Y = 1|SNP)} = \exp(\eta)$$

and thus

$$\mathbf{g}(E[Y|SNP]) = \beta_0 + \beta_1 SNP = \log\left(\frac{Prob(Y = 1|SNP)}{1 - Prob(Y = 1|SNP)}\right) = \boldsymbol{\eta}$$

g is called **the logit function**

Which encoding scheme? Same as before; independent from trait type but dependent on the nature of the marker

- Analyses based on phased haplotype data rather than unphased genotypes may be *quite powerful*...

M1	1	1	2	2
DSL	D	d	d	d
M2	1	2	1	2

Test 1 vs. 2 for M1:

D + d vs. d

Test 1 vs. 2 for M2:

D + d vs. d

Test haplotype H1 vs. all others:

D vs. d

- If the **Disease Susceptibility Locus** (DSL) is located at a marker, haplotype testing can be *less powerful*

5.b Causation

“Association does not imply causation”

- Meaning:

In all observational epidemiologic studies, findings of an association between a substance or exposure and a health effect do not necessarily imply causation.

For example, a study might show that the habit of carrying matches is associated with an increased likelihood of later developing lung cancer.

"Correlation (as a measure of association) is not causation"

- Meaning:

Just because two things correlate does not necessarily mean that one causes the other.

As a seasonal example, just because people in Belgium tend to spend more in the shops when it's cold and less when it's hot doesn't mean cold weather causes high street spending.

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.

Genome-wide Causation Studies of Complex Diseases

Rong Jiao¹, Xiangning Chen², Eric Boerwinkle³ & Momiao Xiong^{1*}

¹Department 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

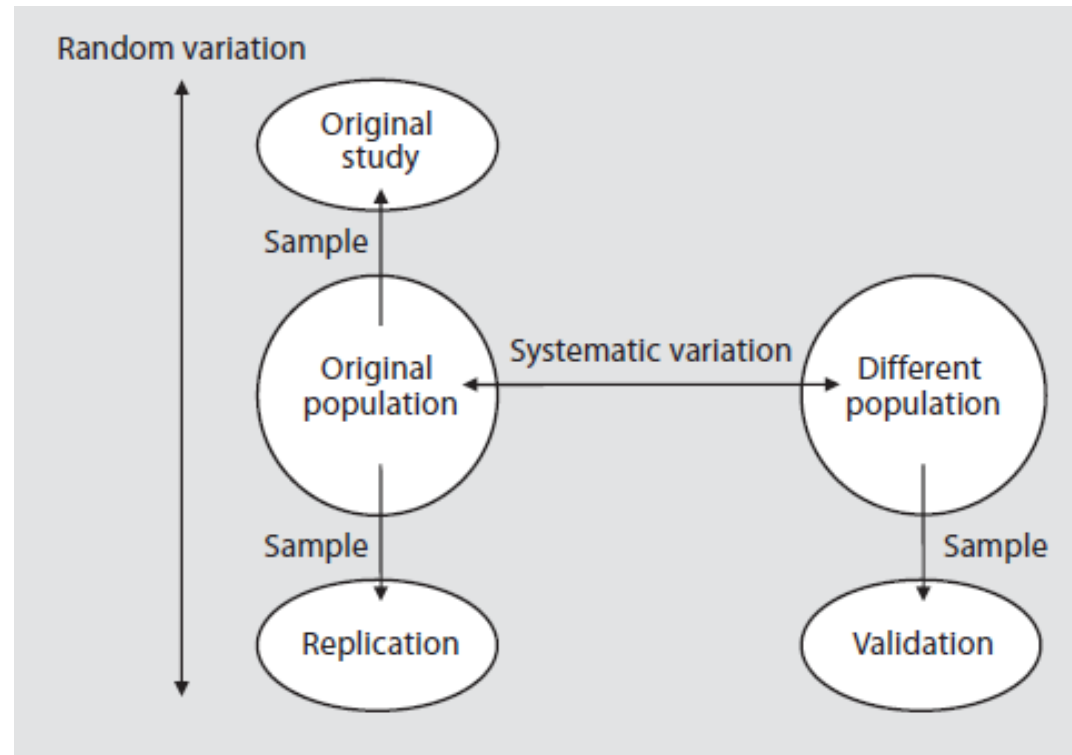
³Epidemiology, Human Genetics & Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA

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

6 Post Association Analysis Steps

6.a Replication and Validation

The difference



(Igl et al. 2009)

Guidelines for replication studies

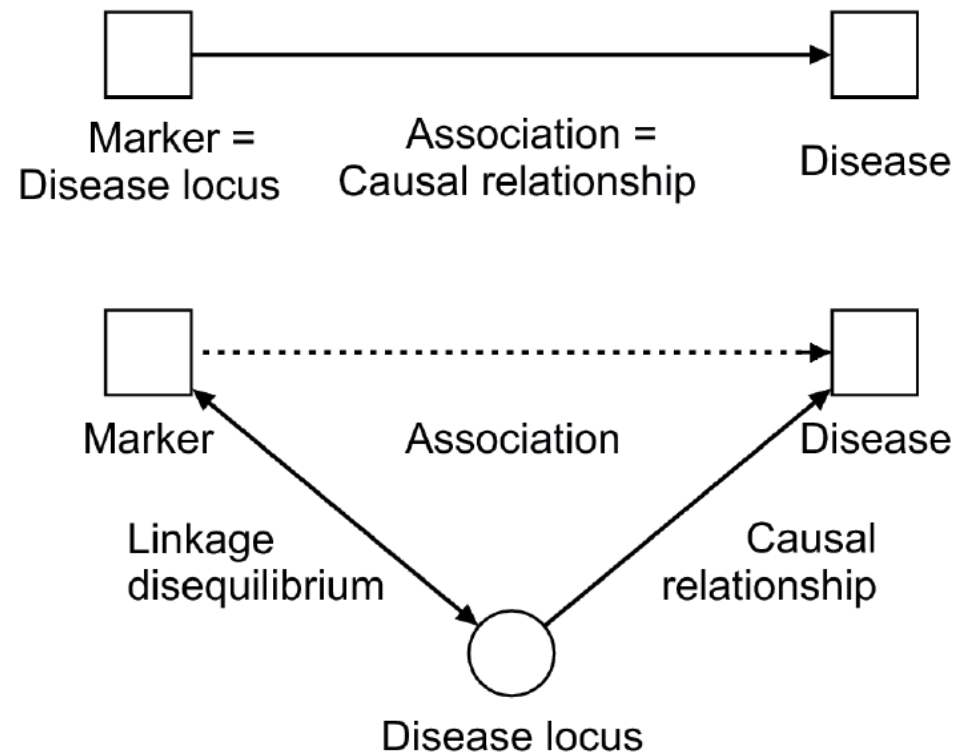
- Replication studies should be of sufficient size to demonstrate the effect
- Replication studies should be conducted in independent datasets
- Replication should involve the same phenotype
- Replication should be conducted in a similar population
- The same SNP should be tested
- The replicated signal 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)

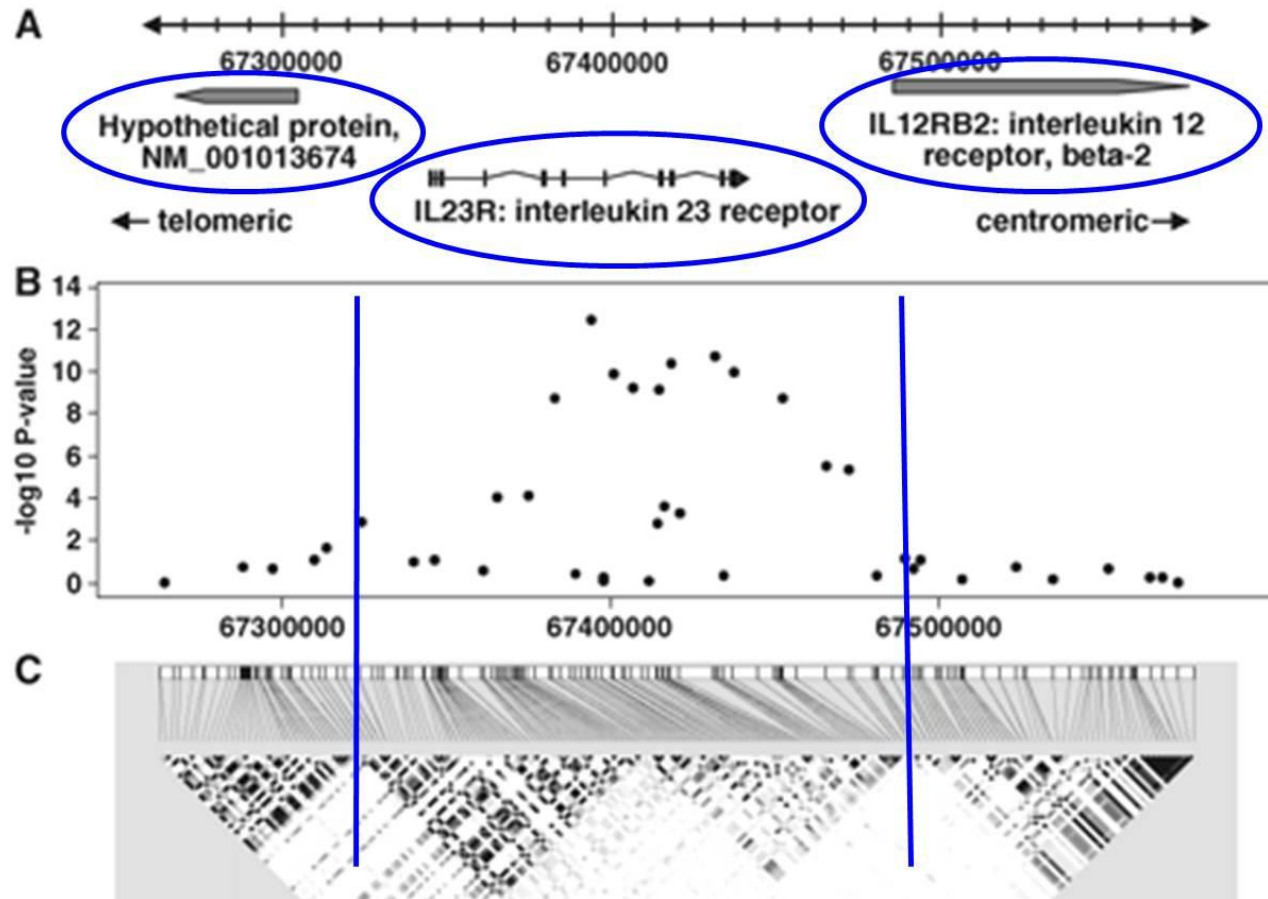
6.b GWA Interpretation and follow-up

Entering the field of functional genomics



(Ziegler and Van Steen, Brazil 2010)

Finding the “relevant” loci – naïve approach



(Duerr et al 2006)

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”



BROAD
INSTITUTE

DEPICT

Home Documentation Citation Contact Feedback

"DEPICT" your association study

DEPICT is an integrative tool that based on predicted gene functions systematically prioritizes the most likely causal genes at associated loci, highlights enriched pathways, and identifies tissues/cell types where genes from associated loci are highly expressed

Download DEPICT (2.9 GB) today

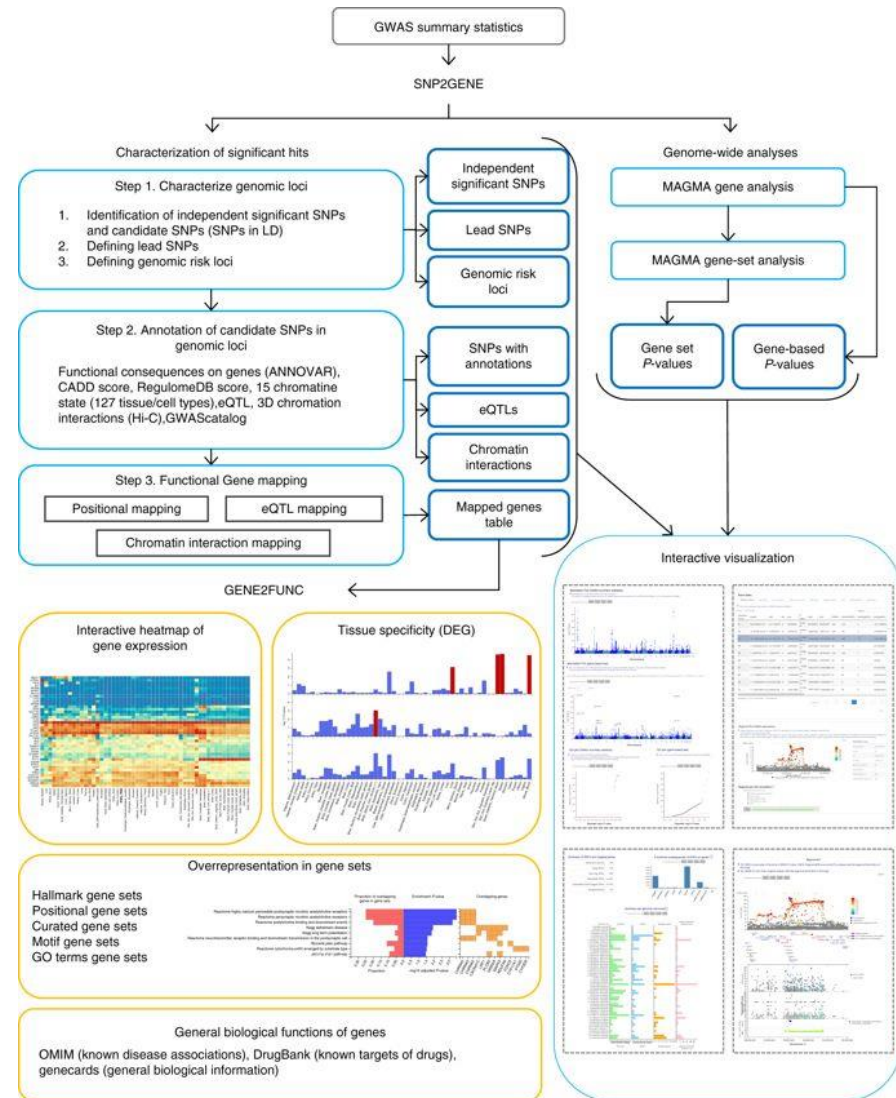
(<https://data.broadinstitute.org/mpg/depict/>)

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

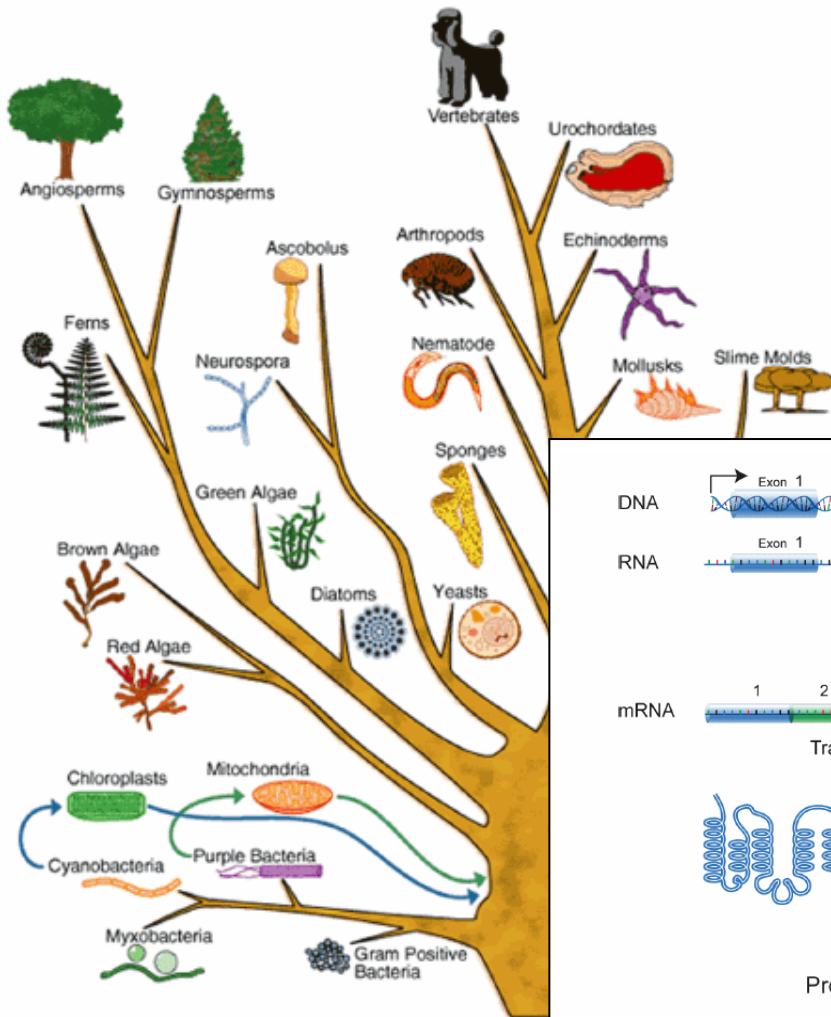
FUMA on GWAS summary statistics. SNP2GENE prioritizes functional SNPs and genes, outputs tables (blue boxes), and creates Manhattan, quantile–quantile (QQ) and interactive regional plots (box at right bottom).

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/>)

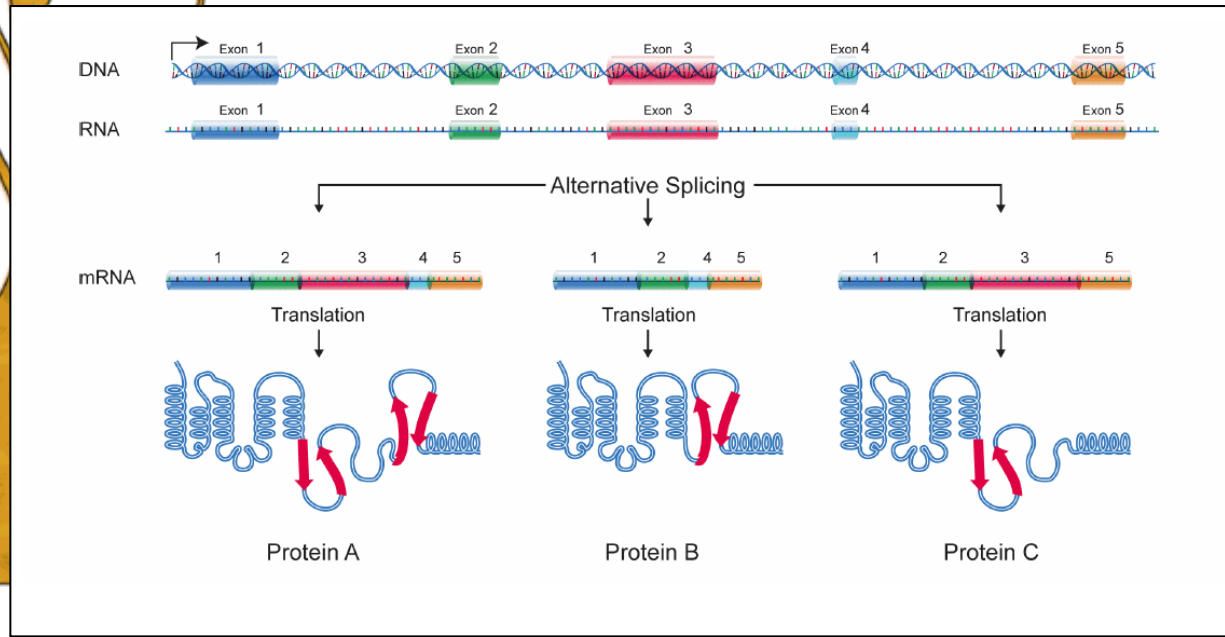


Natural to look at gene expression (in its full complexity!)



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**:



From SNPs to genes as units of (follow-up pathway) analysis

- Pathway analysis allows the interpretation of variants with respect to the biological processes in which the affected genes and proteins are involved.
- Examining the **cumulative effects** of numerous variants and visualizing them at the pathway level, can empower detection of genetic risk factors for complex diseases.
- Visualizing tools can largely aid in making sense of GWAS data!

Next plot:

Highlighted green are the tools in which the specific feature described is present, red highlights indicate features that are either not present or partially present in the tools reviewed.

(Cirillo et al 2017)

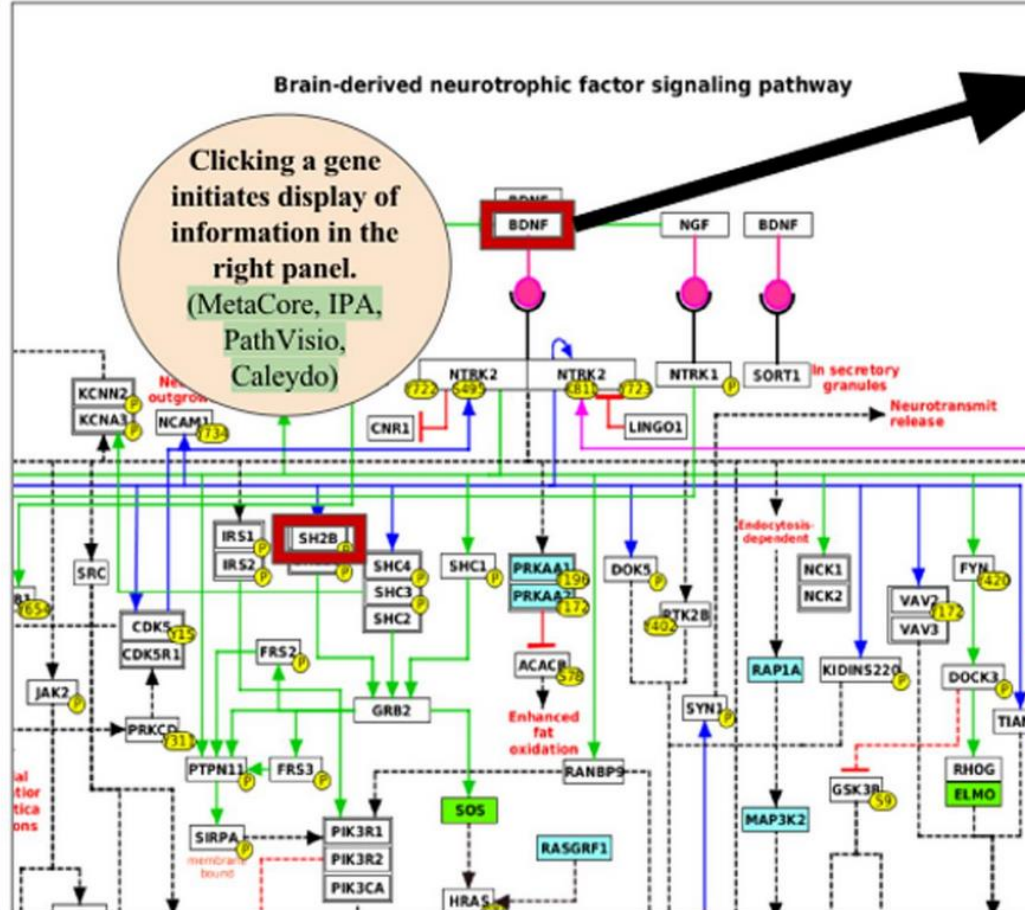
A Pathways list

-Pathways would be listed based on output of a specific GWAS pathway-based algorithm. (None)

- Pathway names are to be ranked according p-values and/or FDR, etc. (MetaCore, IPA and PathVisio)

-Pathways names upon click will appear in the central panel. (MetaCore and Caleydo)

B Pathways diagram



C Other information and hyperlinks

Specific information is shown related to the selected gene:

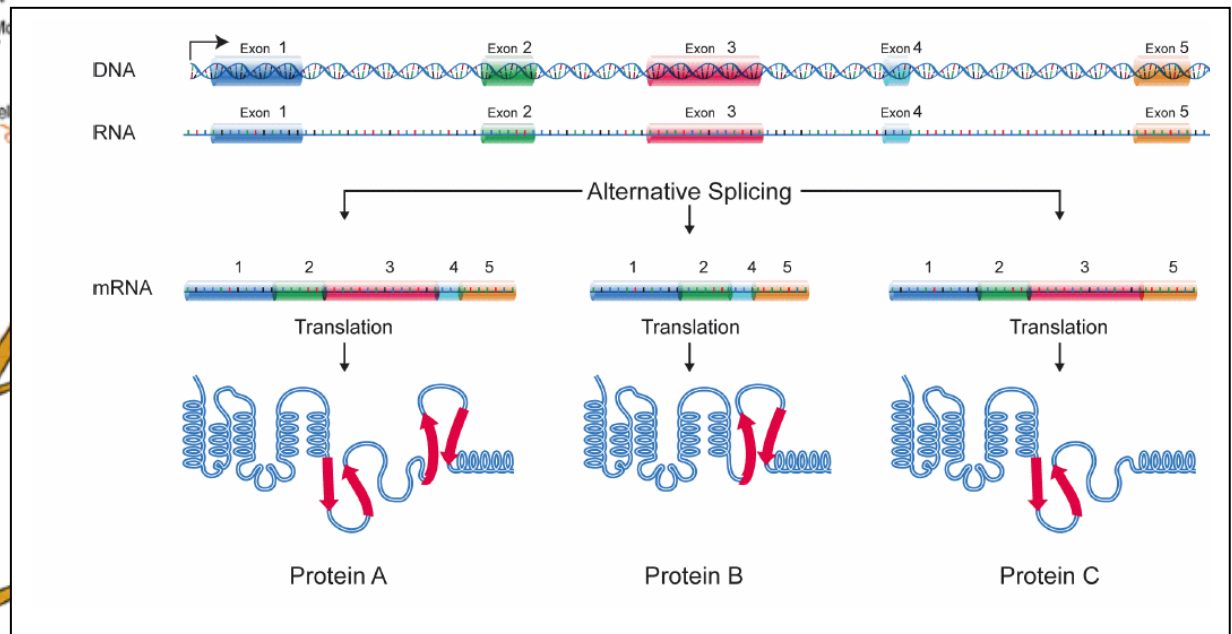
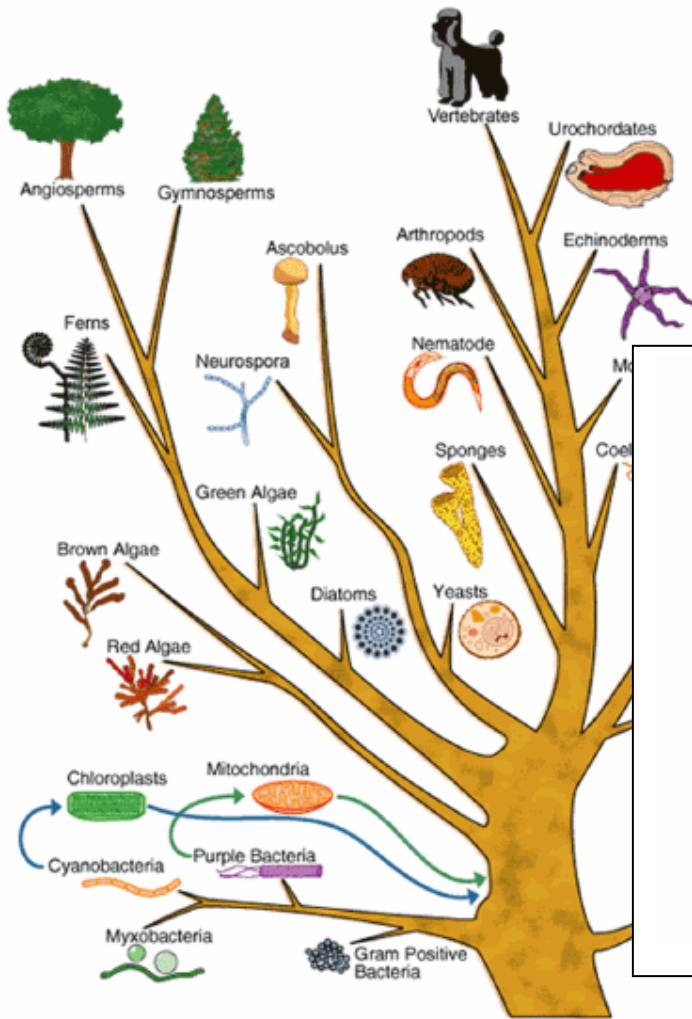
- Pathways
Pathways list of pathways (left panel) that contain the selected gene. (Caleydo)
- Gene
Hyperlinks to databases that contain gene information. (MetaCore, IPA PathVisio)
- SNP
The list of uploaded SNPs is displayed. (MetaCore, IPA PathVisio) SNPs IDs are hyperlinked to databases with added information: description, LD plot, GxE interaction, eQTLs, etc. (Pathvisio and Path cover some of these features)
- Other data
Other uploaded data related to the gene are shown. (MetaCore, IPA PathVisio)

(Cirillo et al 2017)

Gene expression as an extra source of information

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**:



Questions?

Main supporting doc to this class (complementing course slides)



OPEN ACCESS Freely available online



Education

Chapter 11: Genome-Wide Association Studies

William S. Bush^{1*}, Jason H. Moore²

¹ Department of Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University Medical School, Nashville, Tennessee, United States of America, ² Departments of Genetics and Community Family Medicine, Institute for Quantitative Biomedical Sciences, Dartmouth Medical School, Lebanon, New Hampshire, United States of America



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 I

+

Check out the document

“Critical evaluation of a paper/report”