Genetics and Bioinformatics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.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”

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 (incl multiple testing/ pop struct.)

5.b Causation

6 Post-analysis Steps

6.a Replication and validation

6.b GWA Interpretation and follow-up

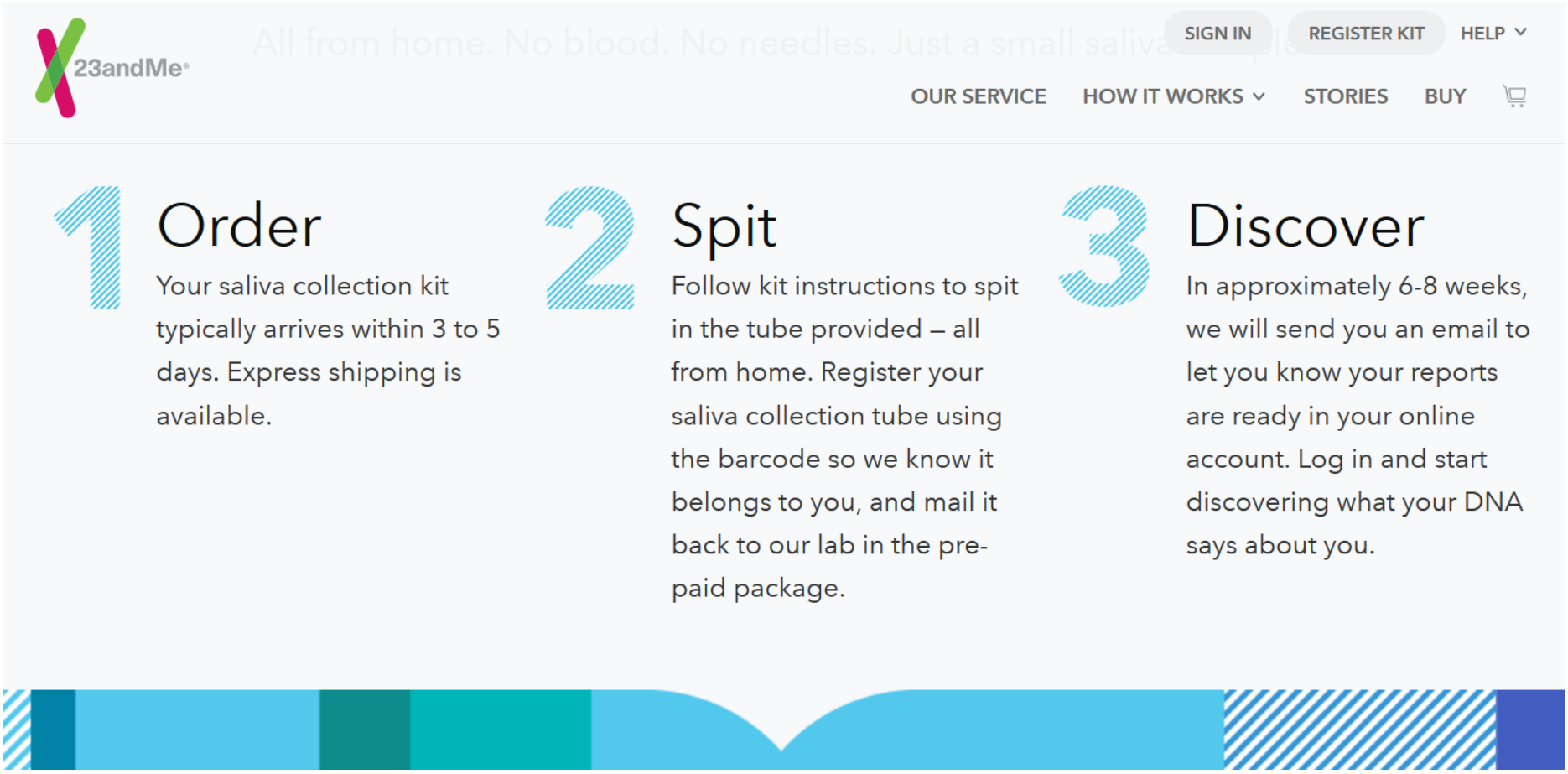
1 Setting the pace

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)


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

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

Example: 23andMe

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

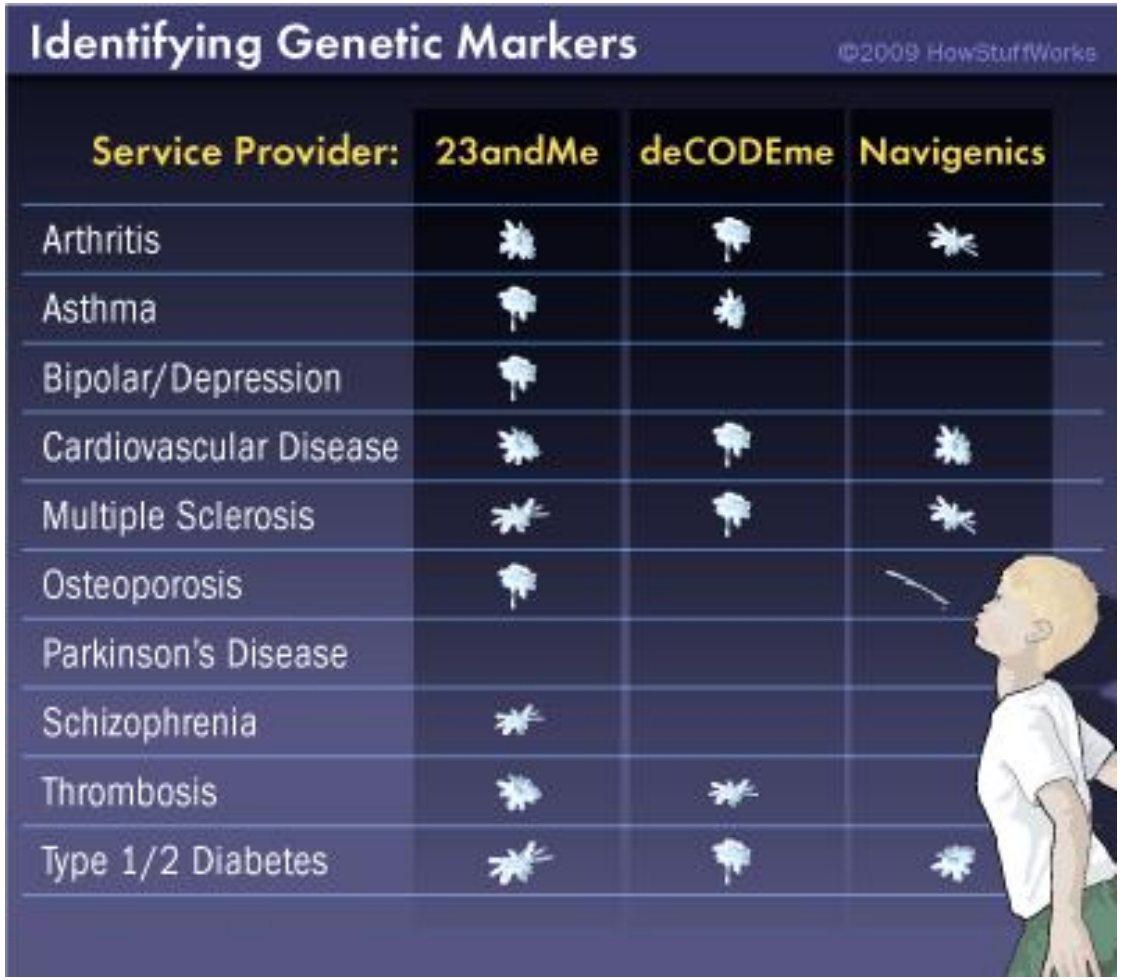
Collect the recommended volume of saliva. The recommended volume of saliva to provide is 2 mL, or about ½ teaspoon. Your saliva sample should be just above the fill line.

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.


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

Can you handle the truth?



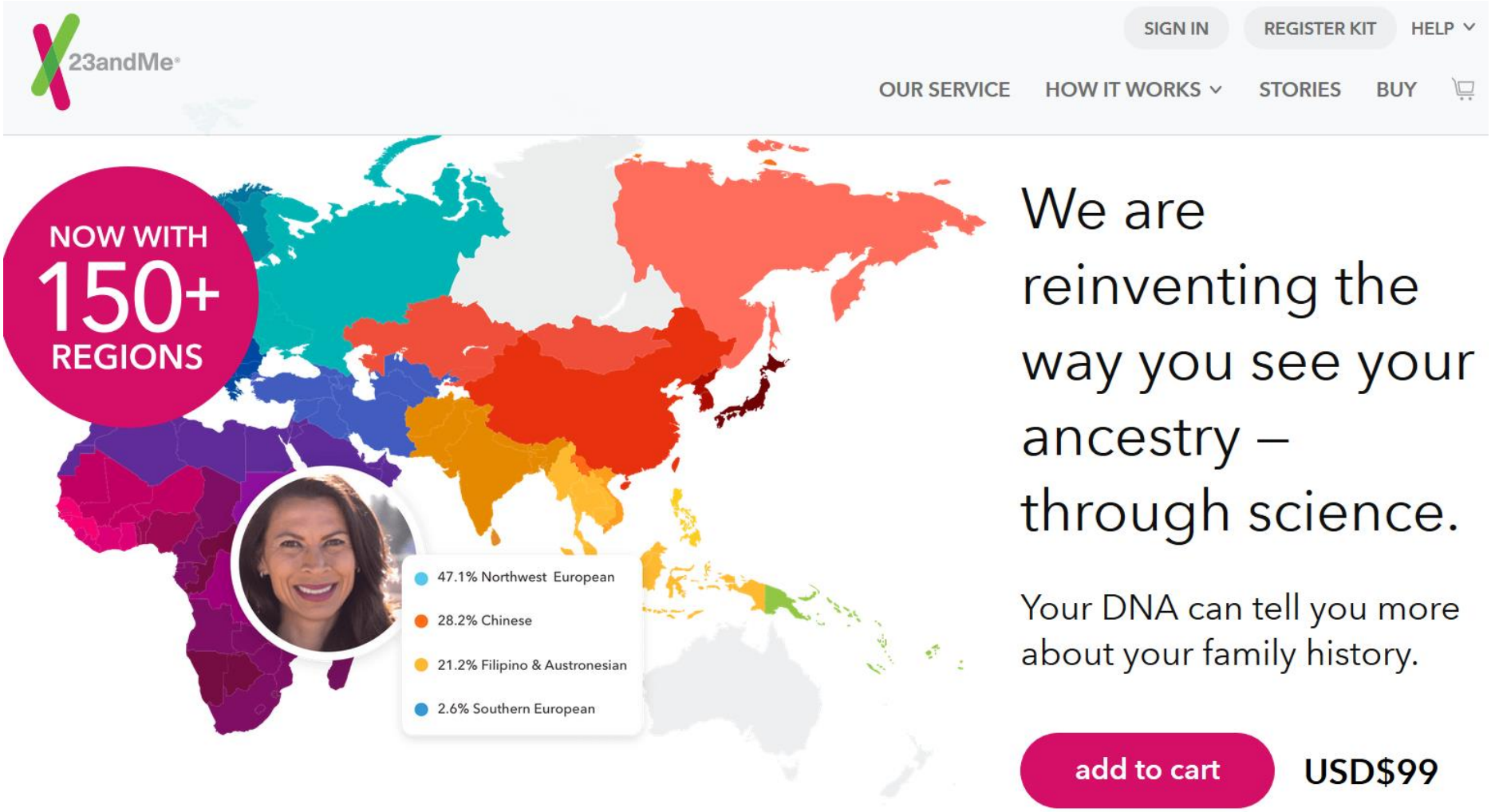
NewStatesman

23ANDME CO-FOUNDER ANNE WOJCICKI. PHOTO: KIMBERLY WHITE/GETTY IMAGES

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.


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

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

Routledge
Taylor & Francis Group

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]

[a]Department of Genetics, University of Leicester, Leicester, UK; [b]School 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.

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
  - presymptomatic testing for **estimating risk** developing disease
  - presymptomatic testing for **predicting** disease
  - **prenatal screening**
  - newborn screening
  - preimplantation genetic diagnosis
  - carrier screening
  - **forensic testing**
  - **paternal testing**

How is genetic testing used clinically?

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

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

How is genetic testing used clinically?

- **Pharmacogenomics**: classifies subtle variations in an individual's genetic makeup to determine whether a drug is suitable for a particular patient, and if so, what would be the safest and most effective dose. Learn more about pharmacogenomics. → DNA passports … are no 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. See also “The promise and challenges of next-generation genome sequencing for clinical care” (JAMA Intern Med. 2014)

The basics of 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.

Reference genome

- 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

Reference genome

PLOS ONE

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-Lleal[3], Daniel I. Chasman[4,5], Stella Trompet[6,7], Tarunveer S. Ahluwalia[8,9], Alexander Teumer[10], Marcus E. Kleber[11], Ming-Huei Chen[12,13], Jie Jin Wang[14], John R. Attia[15,16], Riccardo E. Marioni[17,18,19], Maristella Steri[20], Lu-Chen Weng[21], Rene Pool[22,23], Vera Grossmann[24], Jennifer A. Brody[25], Cristina Venturini[26,27], Toshiko Tanaka[28], Lynda M. Rose[4], Christopher Oldmeadow[15,16], Johanna Mazur[29], Saonli Basu[30], Mattias Frånberg[3,31], Qiong Yang[13,32], Symen Ligthart[1], Jouke J. Hottenga[22], Ann Rumley[33], Antonella Mulas[20], Anton J. M. de Craen[7], Anne Grotevendt[34], Kent D. Taylor[35,36], Graciela E. Delgado[11], Annette Kifley[14], Lorna M. Lopez[17,37,38], Tina L. Berentzen[39], Massimo Mangino[27,40], Stefania Bandinelli[41], Alanna C. Morrison[1], Anders Hamsten[3], Geoffrey Tofler[42], Moniek P. M. de Maat[43], Harmen H. M. Draisma[22,44], Gordon D. Lowe[45], Magdalena Zoledziewska[20], Naveed Sattar[46], Karl J. Lackner[47], Uwe Völker[48], Barbara McKnight[49], Jie Huang[50], Elizabeth G. Holliday[51], Mark A. McEvoy[16], John M. Starr[17,52], Pirro G. Hysi[27], Dena G. Hernandez[53], Weihua Guan[30], Fernando Rivadeneira[1,54], Wendy L. McArdle[55], P. Eline Slagboom[56], Tanja Zeller[57,58], Bruce M. Psaty[59,60], André G. Uitterlinden[1,54], Eco J. C. de Geus[22,23], David J. Stott[61], Harald Binder[62], Albert Hofman[1,63], Oscar H. Franco[1], Jerome I. Rotter[64,65], Luigi Ferrucci[28], Tim D. Spector[27], Ian J. Deary[17,66], Winfried März[11,67,68], Andreas Greinacher[69], Philipp S. Wild[70,71,72], Francesco Cucca[20], Dorret I. Boomsma[22], Hugh Watkins[73], Weihong Tang[21], Paul M. Ridker[4,5], Jan W. Jukema[6,74,75], Rodney J. Scott[76,77], Paul Mitchell[14], Torben Hansen[78], Christopher J. O'Donnell[13,79], Nicholas L. Smith[60,80,81], David P. Strachan[82], Abbas Dehghan[1,83]*

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

Types of molecular markers (Schlötterer 2004)

OPINION

The evolution of molecular markers — just a matter of fashion?

Christian Schlötterer

In less than half a century, molecular markers have totally changed our view of nature, and in the process they have evolved themselves. However, all of the molecular methods developed over the years to detect variation do so in one of only three conceptually different classes of marker: protein variants (allozymes), DNA sequence polymorphism and DNA repeat variation. The latest techniques promise to provide cheap, high-throughput methods for genotyping existing markers, but might other traditional approaches offer better value for some applications?

Being able to distinguish between genotypes that are relevant to a trait of interest is a key goal in genetics. Often, this distinction is not based directly on the trait of interest, but on informative marker systems. A genetic marker provides information about allelic variation at a given locus. The first genetic map of *Drosophila melanogaster* was built by Sturtevant using phenotypic markers[1]. How-

continuous improvement in the way in which we assay genetic variation; that is, the latest marker systems are the most informative ones. Nevertheless, in reviewing the history of molecular markers and their pros and cons, I argue that there are only a few conceptually different classes of marker and that recently developed high-throughput methods might not be unconditionally superior to more traditional approaches.

Allozymes

The first true molecular markers to be established were allozymes (a term that originates from a contraction of the phrase 'allelic variants of enzymes'). The principle of allozyme markers is that protein variants in enzymes can be distinguished by native gel electrophoresis according to differences in size and charge caused by amino-acid substitutions. To visualize the allozyme bands, the electrophoretic gels are treated with enzyme-specific stains that contain substrate for the enzyme, cofactors and an oxidized salt (for example, nitro-blue tetra-

sample sizes are typically studied in allozyme surveys. Nevertheless, the number of informative marker loci is too small to use allozymes for mapping and ASSOCIATION STUDIES[8]. Furthermore, surveys of natural variation based on allozymes were often challenged by non-neutral evolution of some of the markers used (see, for example, REFS 9–11).

The arrival of DNA-based markers

One of the criticisms levelled at allozyme markers is that they are an indirect and insensitive method of detecting variation in DNA. A more direct molecular marker would survey DNA variation itself, rather than rely on variations in the electrophoretic mobility of proteins that the DNA encodes. Another important advantage that DNA-based markers have over allozymes is that they allow the number of mutations between different alleles to be quantified. Given these unambiguous advantages, the arrival of DNA manipulation techniques promoted a shift from enzyme-based to DNA-based markers.

"…the arrival of DNA manipulation techniques promoted a shift from enzyme-based to DNA-based markers."

Types of molecular markers

- Enzyme based:
  - Enzymes are biological molecules (typically proteins) that act as catalysts and help complex reactions occur everywhere in life.
- **DNA sequence-based**:
  - Nowadays, **genetic markers represent sequences of DNA** which have been traced to specific locations on the chromosomes and associated with particular traits (i.e., coded phenotype = coded subject's/object's characteristic).
  - They demonstrate **polymorphism**, which means that the genetic markers in different organisms of the same species are different.

Marker	Advantages	Disadvantages
SNPs	• Low mutation rate • High abundance • Easy to type • New analytical approaches are being developed at present • Cross-study comparisons are easy; data repositories already exist	• Substantial rate heterogeneity among sites • Expensive to isolate • Ascertainment bias • Low information content of a single SNP
Microsatellites	• Highly informative (large number of alleles, high heterozygosity) • Low ascertainment bias • Easy to isolate	• High mutation rate • Complex mutation behaviour • Not abundant enough • Difficult to automate • Cross-study comparisons require special preparation
Allozymes	• Cheap • Universal protocols	• Requirement for fresh or frozen material • Some loci show protein instability • Limited number of available markers • Potentially direct target of selection
RAPDs and derivatives	• Cheap • Produces a large number of bands, which can then be further characterized individually (for example, converted into single locus markers)	• Low reproducibility • Mainly dominant • Difficult to analyse • Difficult to automate • Cross-study comparisons are difficult
DNA sequencing	• Highest level of resolution possible • Not biased • Cross-study comparisons are	• Still significantly more expensive than the other techniques

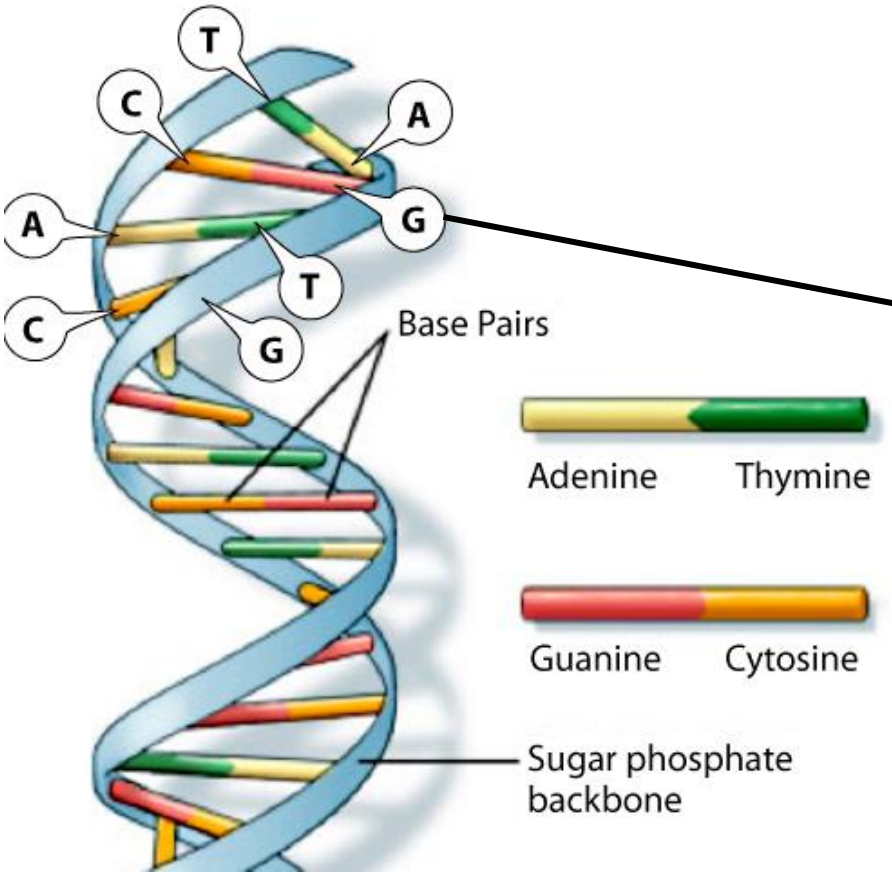
A microsatellite is a tract of repetitive DNA in which certain DNA motifs (ranging in length from one to six or more base pairs) are repeated, typically 5–50 times

Be critical

(date of publication = 2004)

Hence, it is important to keep the historical time lines and achievements in mind

Types of genetic markers: single nucleotide polymorphisms



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

Types of genetic markers: single nucleotide polymorphisms or **SNPs**

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

(Ziegler and Van Steen, Brazil 2010)

Types of genetic markers: single nucleotide polymorphisms

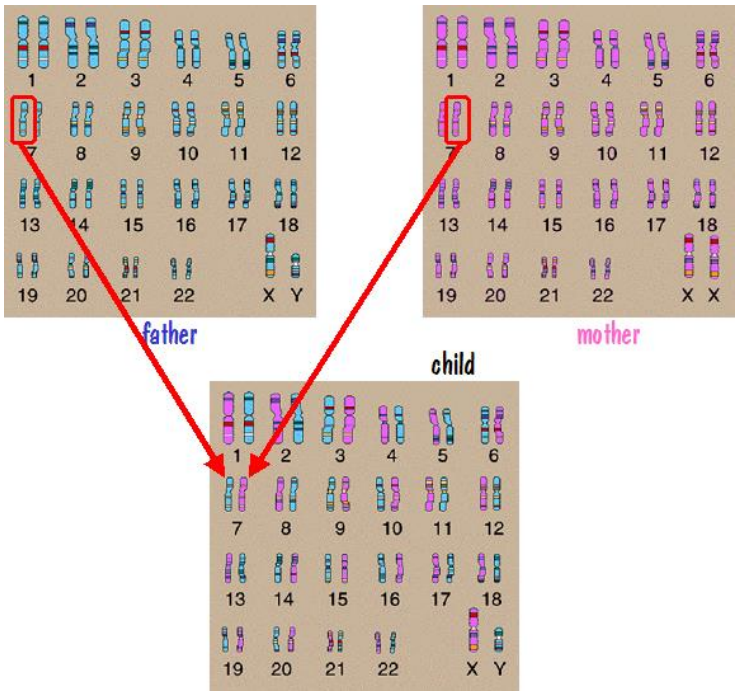
**Submissions received after reclustering of current build will appear as new rs# clusters in the next build.*

BUILD STATISTICS:

Organism	dbSNP Build	Genome Build	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (rs#'s) in gene	Number of (ss#'s) with genotype	Number of (ss#'s) with frequency	Number of weight 1 SNPs	Number of weight 2+ SNPs
Homo sapiens	150	38.3	907,237,763	325,658,303 (135,967,291)	191,585,061	73,917,935	129,875,536		
Bos taurus	150	7.2	332,061,559	104,286,568 (12,102,319)	46,308,631	10,202	968		
Mus musculus	150	38.5	189,214,027	84,152,707 (6,466,270)	40,278,667	24,843,897	77	DIV:9911312 MNV:452 Named:6779 SNV:67883617	DIV:180165 MNV:2259 SNV:1647286
Sus scrofa	150	5.1	195,656,177	67,116,509 (8,107,358)	36,126,981	52	184		
Ovis aries	150	2.1	147,584,937	63,745,118 (3,570,277)	30,029,327	65	173		
Macaca mulatta	150	2.1	95,808,453	53,929,680 (2,760,325)	23,087,008	29	8,072	DIV:9 SNV:32798877	SNV:38416
Zea mays	150	1.1	86,608,237	58,915,360 (14,672,946)	13,436,128	90			
Gallus gallus	150	4.1	73,244,003	24,277,657 (15,305,602)	14,926,051	3,624,831	203		
Bos indicus	150	1.1	30,533,959	17,758,946 (621)	5,131,669		223		
Arabidopsis thaliana	150	9.2	15,307,574	13,412,809 (5,947)	9,174,636	299		DIV:4 MNV:5 SNV:1069121	MNV:1 SNV:338

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 structures called chromosomes.
- A chromosome contains a single, long DNA molecule, only a portion of which corresponds to a single gene.



(Figure : Human chromosomes)

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

→ links to giving an “interpretation”

Alleles

- Allele: one of several alternative forms of DNA sequence at specific chromosomal location
- Polymorphism: often used to indicate the existence of at least 2 alleles at a single “locus”
- **Homozygosity** (homozygous): both alleles identical at locus
- **Heterozygosity** (heterozygous): different alleles at locus

- Genetic marker (in this course): polymorphic DNA sequence at single locus

[Mutations ~polymorphisms (see later)]

Hunting 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?

- 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 of bases seen only in family members who have the disease or trait. These characteristic molecular patterns are referred to as polymorphisms, or markers.
- Before researchers identify the gene responsible for the disease or trait, DNA markers can tell them roughly where the gene is on the chromosome.

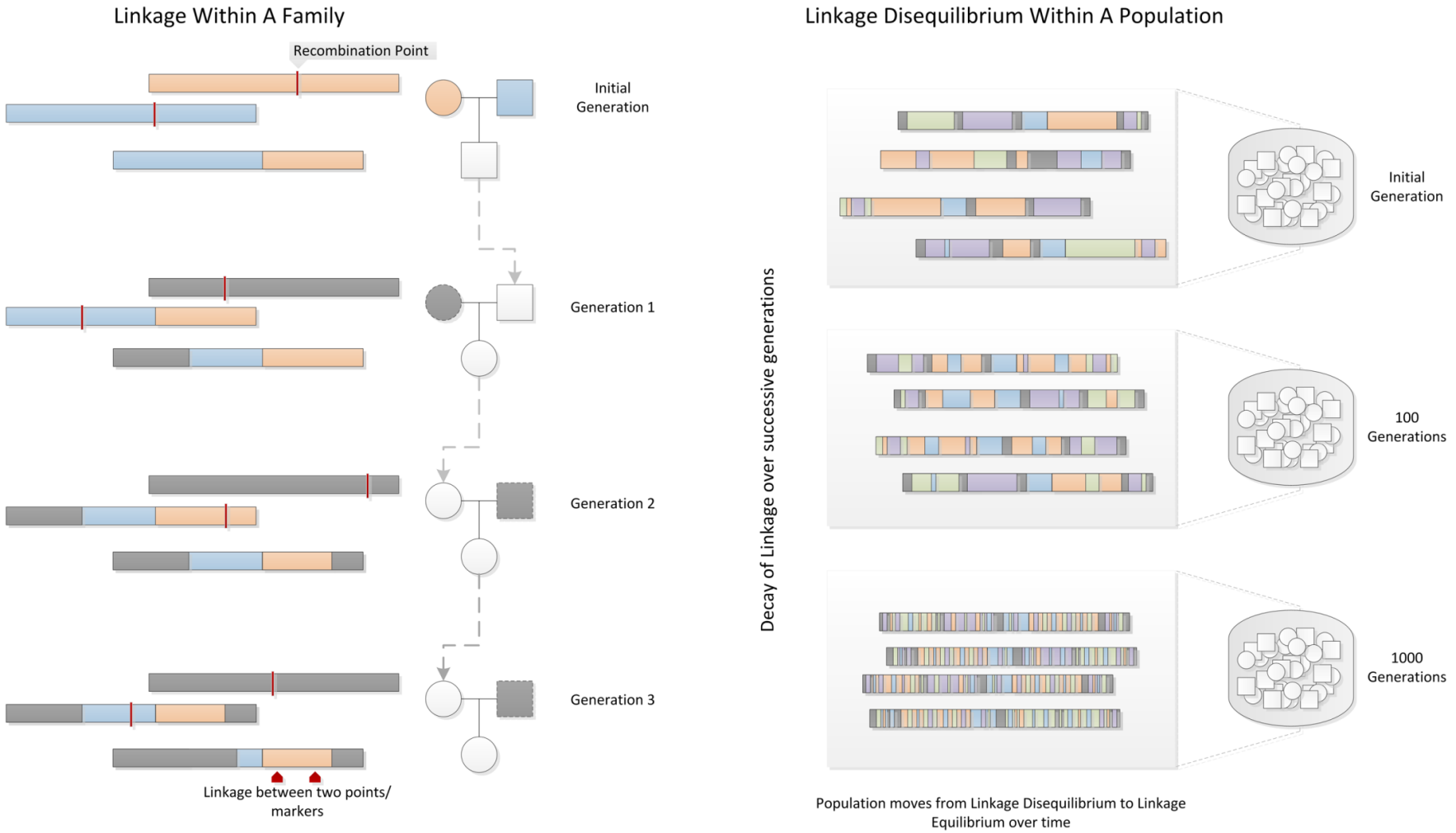
How is this possible?

How to generate a genetic map? (continued)

- This is possible because of a genetic process known as recombination.

> *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? (continued)



(Bush et al. 2012)

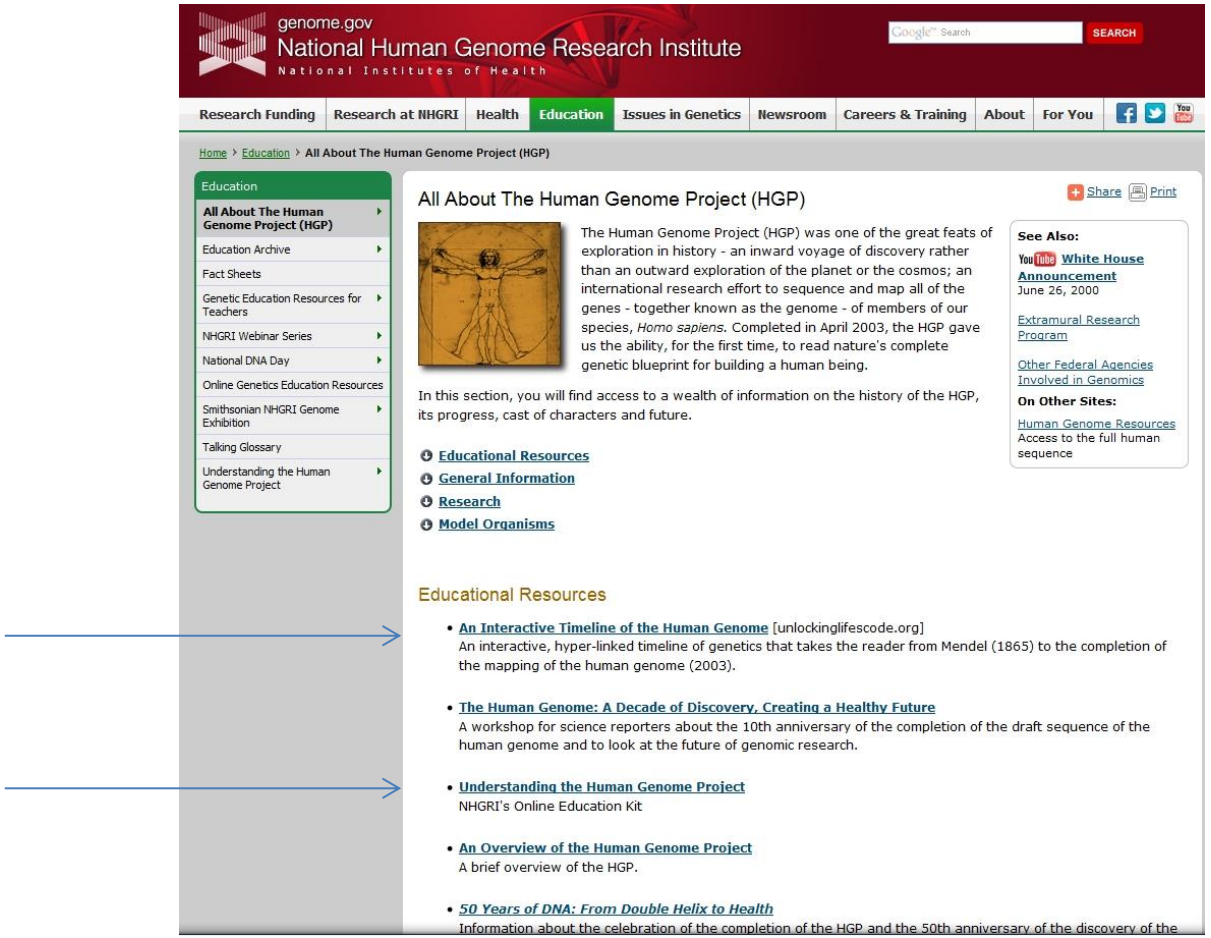
How to generate a genetic map? (continued)

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

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

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

1.c “The Human Genome Project”



Historical overview

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

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

1865

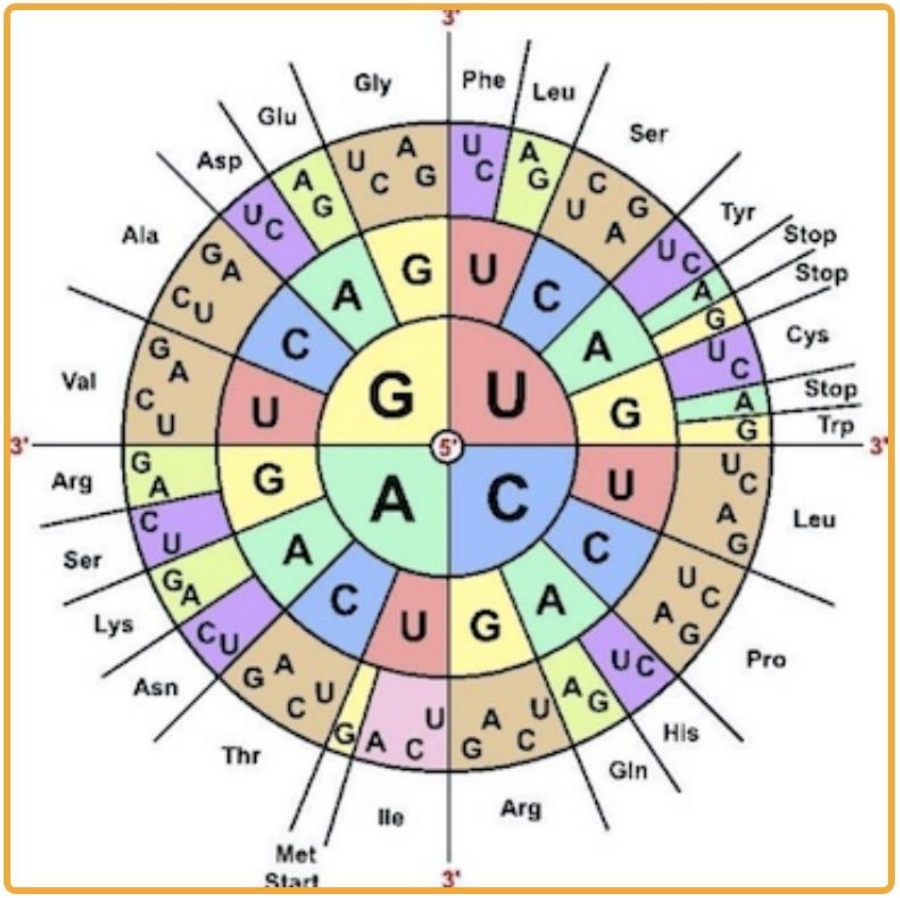
Historical overview

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

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

1953

Historical overview

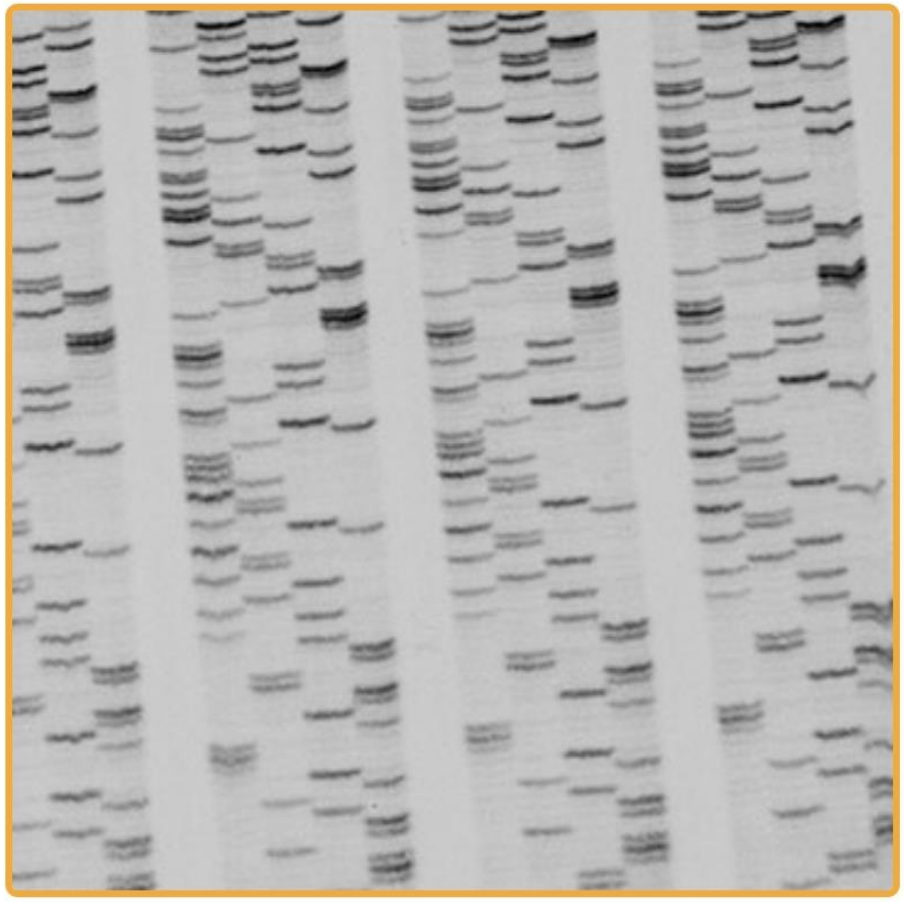


Marshall Nirenberg cracks the genetic code for protein synthesis

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

1961

Historical overview

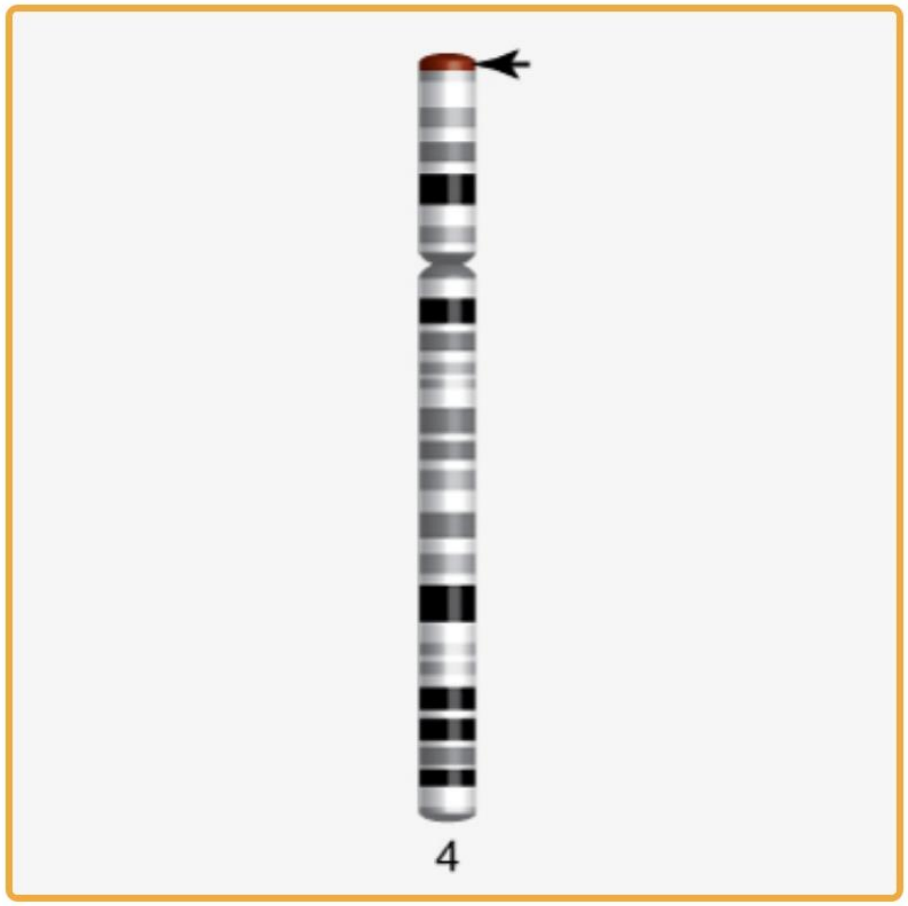


Frederick Sanger develops rapid DNA sequencing technique

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

1977

Historical overview

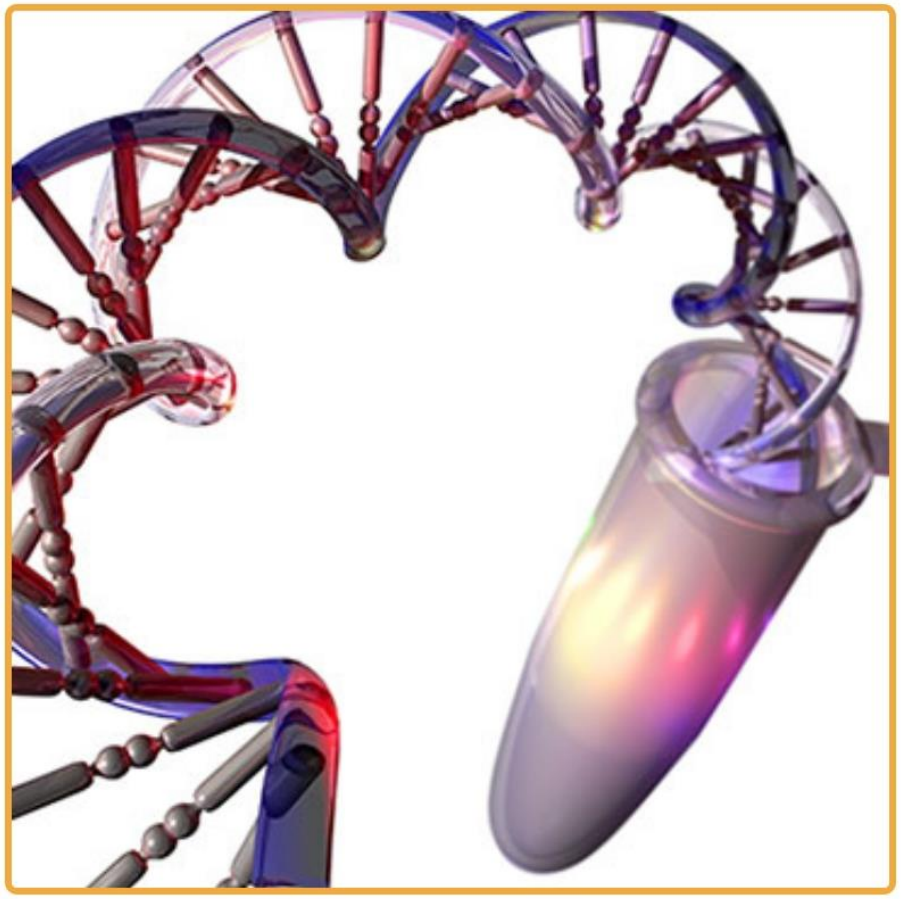


First genetic disease mapped, Huntington's Disease

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

1983

Historical overview

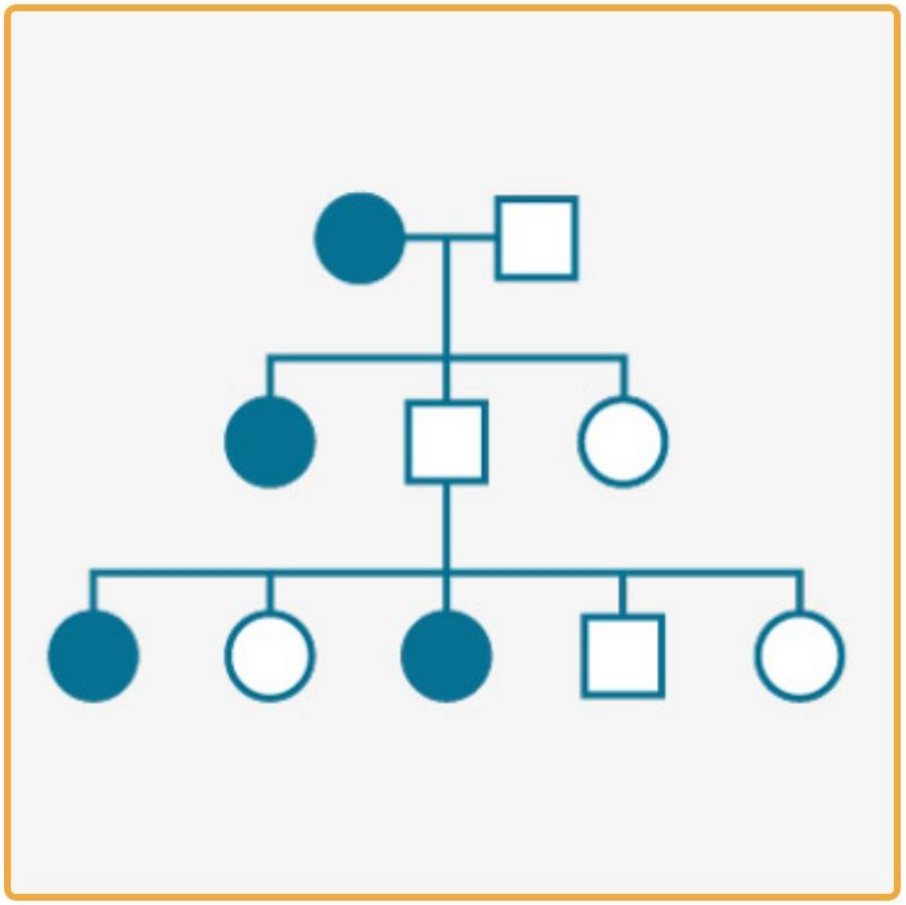


Invention of polymerase chain reaction (PCR) technology for amplifying DNA

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

1983

Historical overview



First evidence provided for the existence of the BRCA1 gene

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

1990

Historical overview

The Human Genome Project begins

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

1990

Historical overview

THE HUMAN GENOME

The Sequence of the Human Genome

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

16 FEBRUARY 2001 VOL 291 SCIENCE www.sciencemag.org

- 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**;

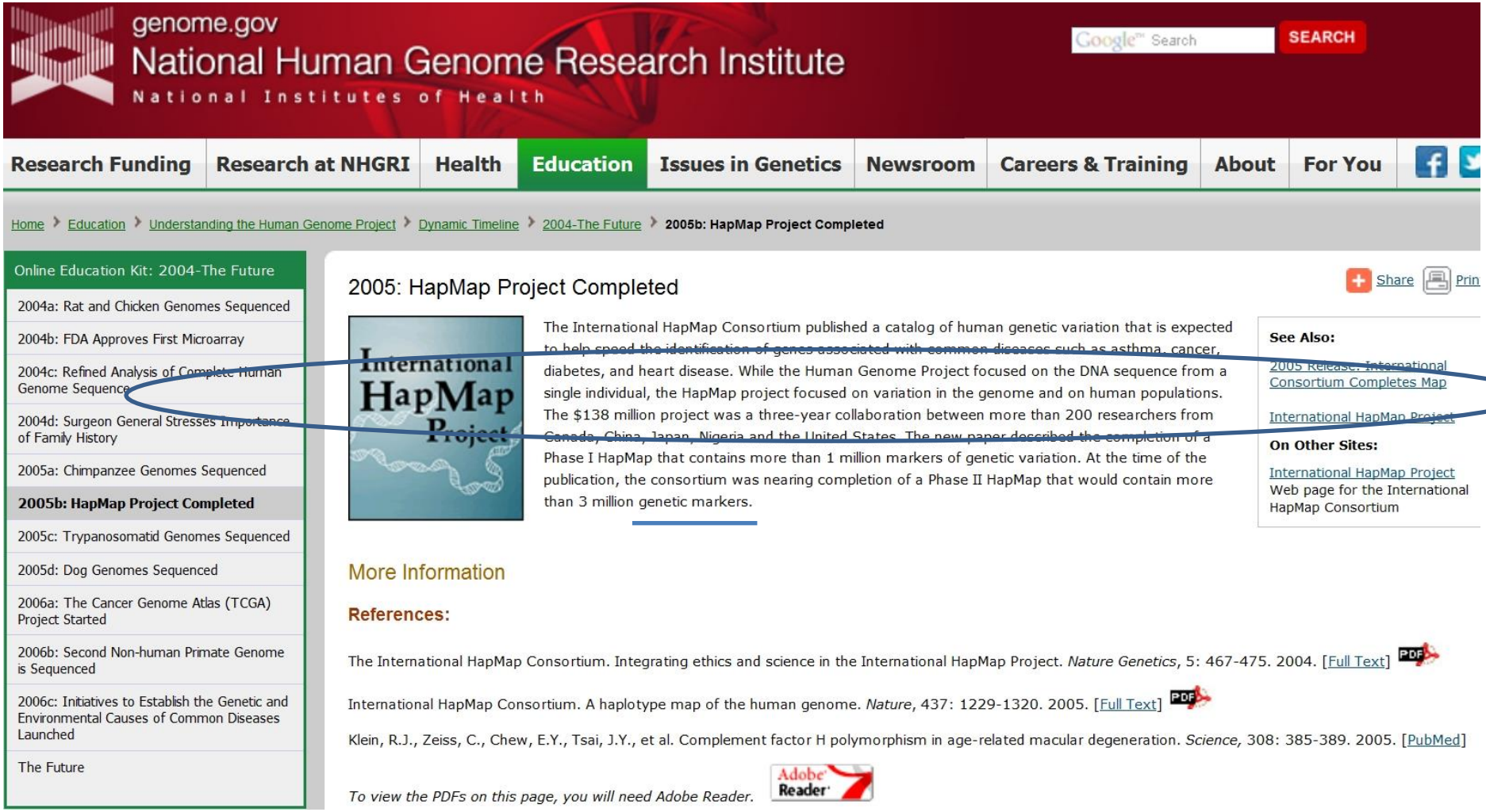
    Note: 100,000 → 30000-35000 → 24000 → 19000-20000

  - 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

Historical overview



Historical overview

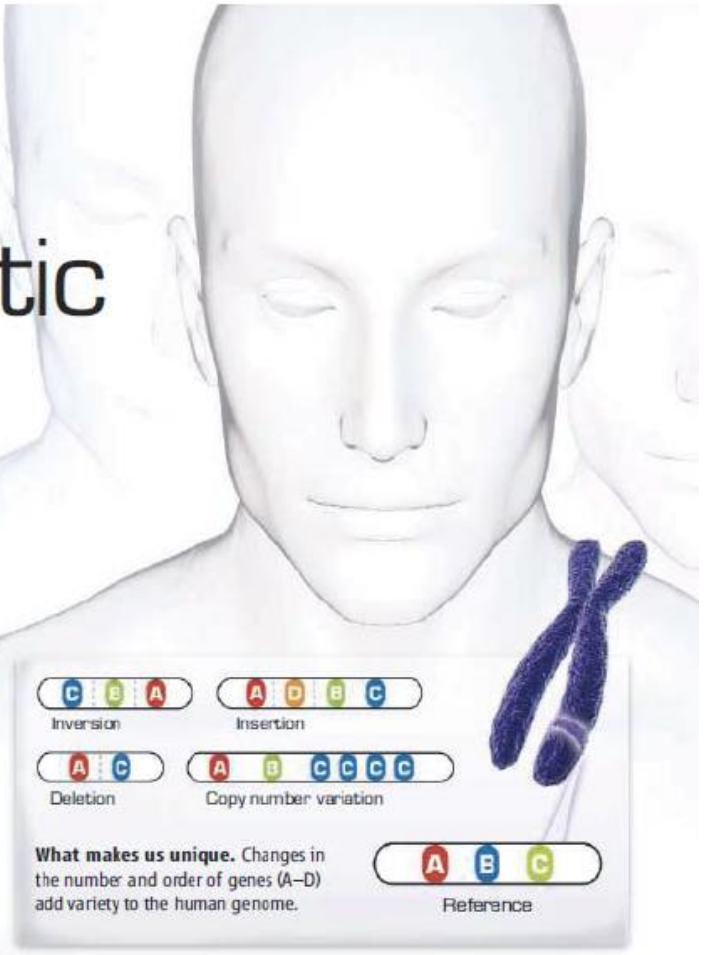
BREAKTHROUGH OF THE YEAR

Human Genetic Variation

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

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

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



What makes us unique. Changes in the number and order of genes (A–D) add variety to the human genome.

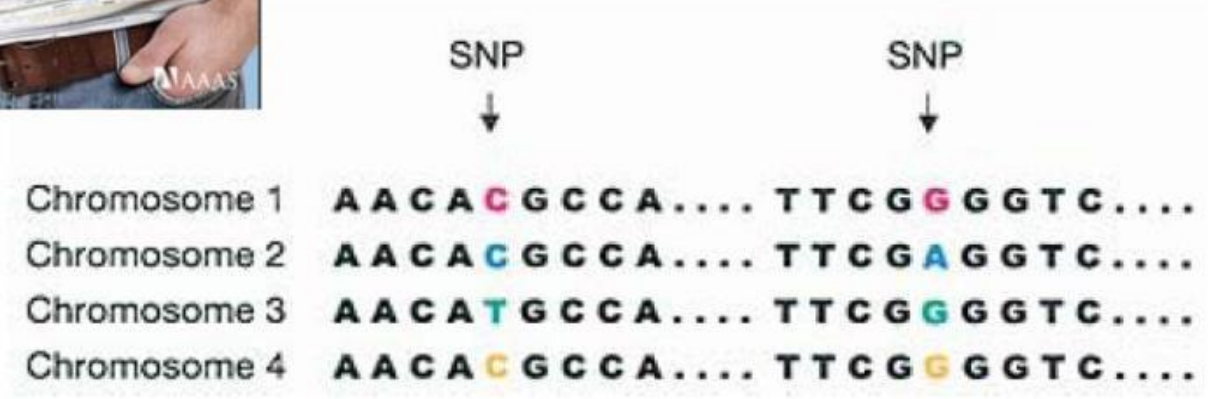
Pennisi 2007 Science 318:1842-3

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

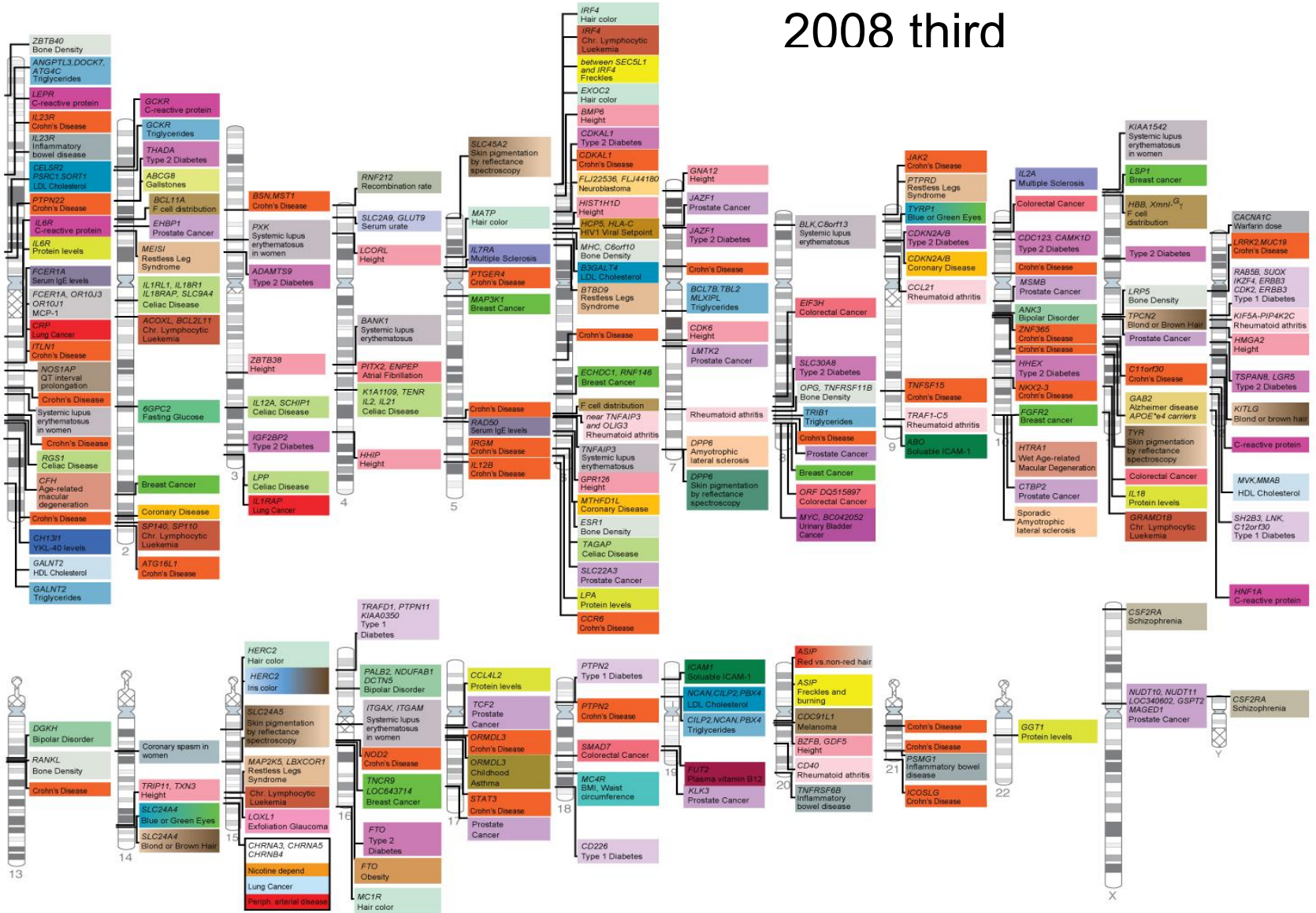
BREAKTHROUGH OF THE YEAR: The Runners-Up

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

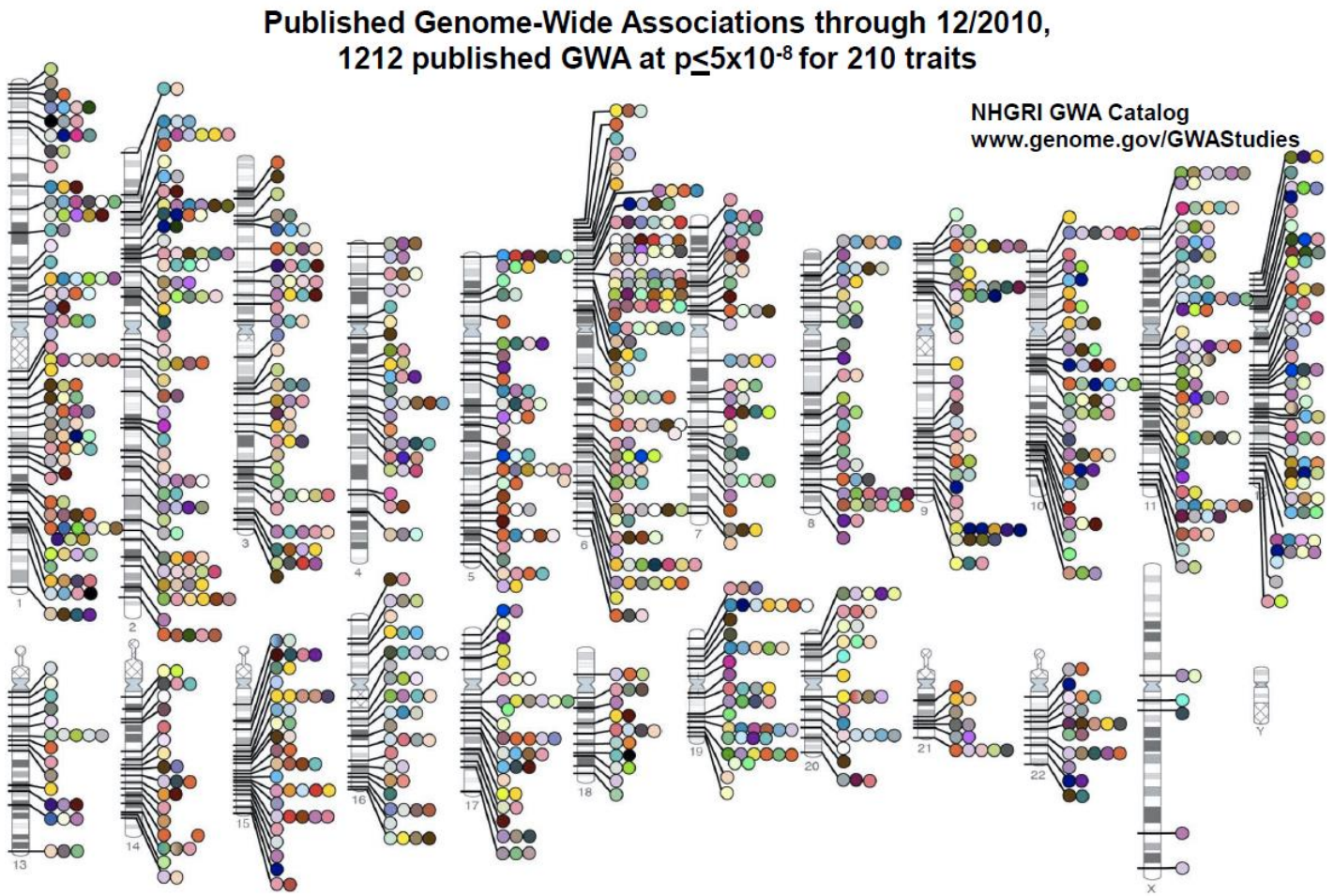
Areas to Watch in 2007

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?

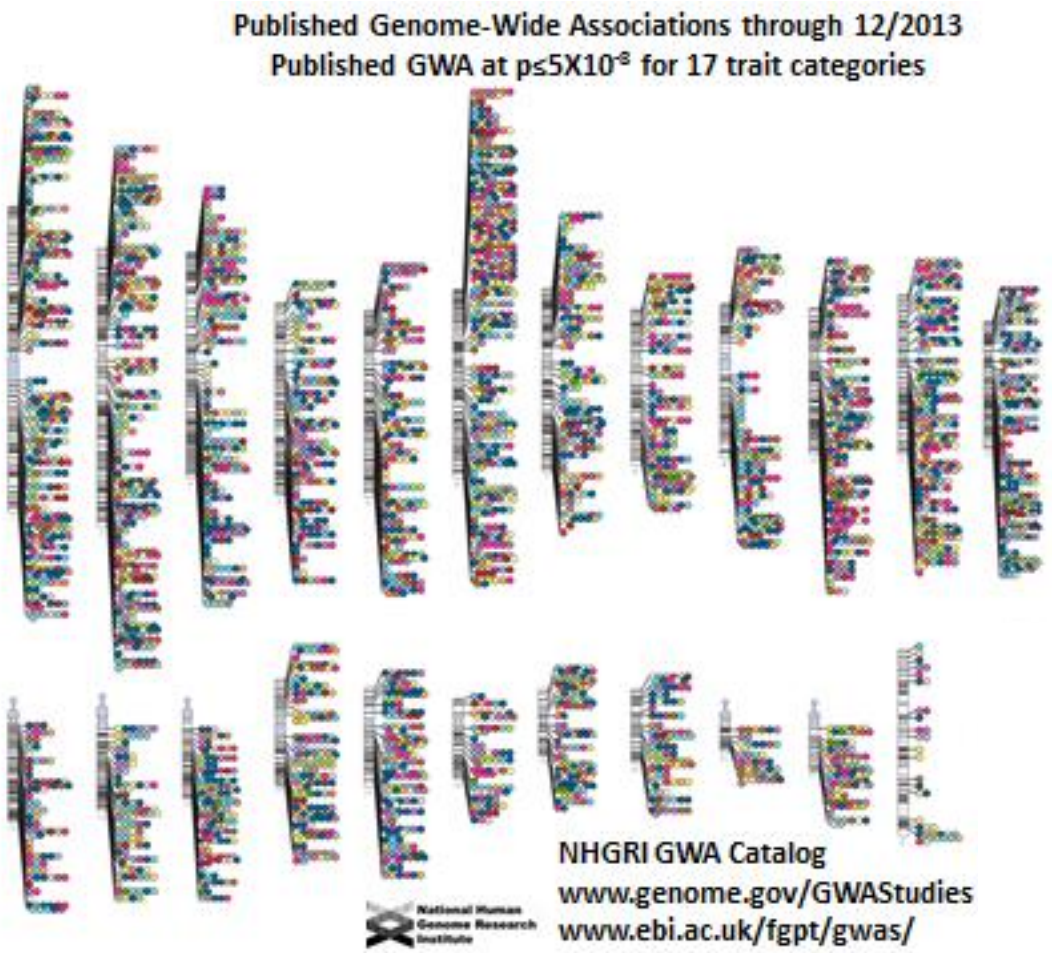
Historical overview: GWAs as a tool to “map” diseases



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

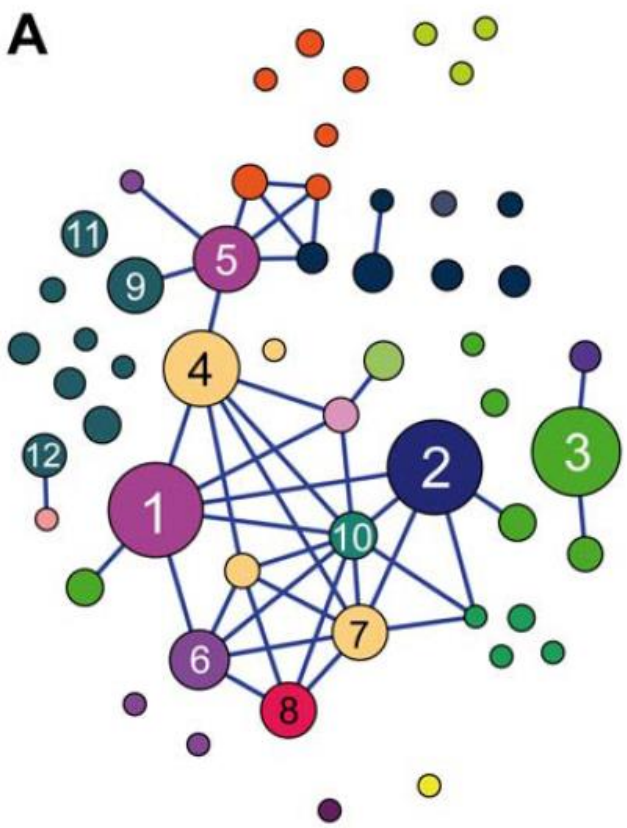


Historical overview: trait categories



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

Historical overview: inter-relationships (networks)

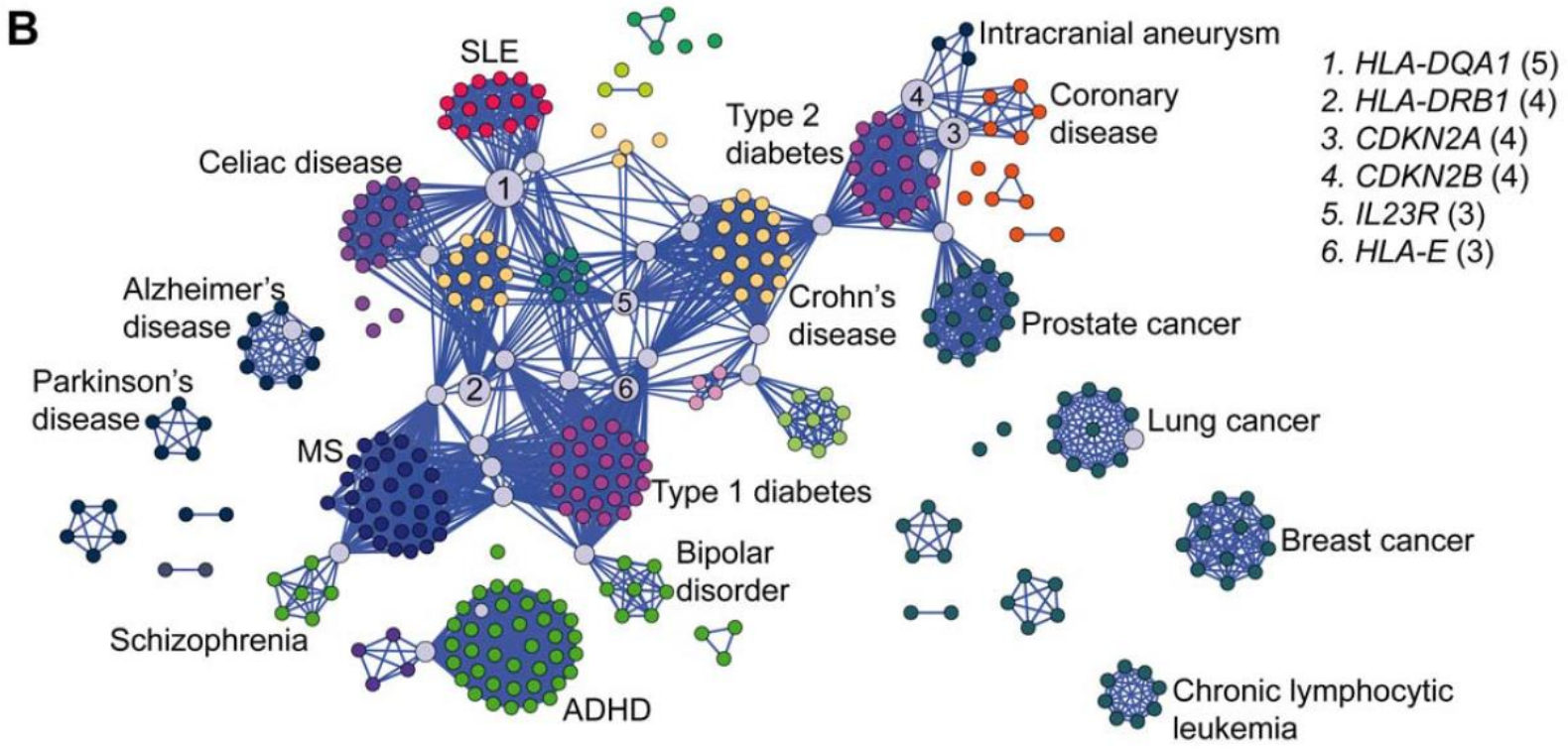


1. Type 1 diabetes (36)
2. Multiple sclerosis (36)
3. ADHD and conduct disorder (33)
4. Crohn's disease (27)
5. Type 2 diabetes (22)
6. Celiac disease (19)
7. Ulcerative colitis(17)
8. Systemic lupus erythematosus (17)
9. Prostate cancer (17)
10. Rheumatoid arthritis (13)
11. Breast cancer (12)
12. Lung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

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

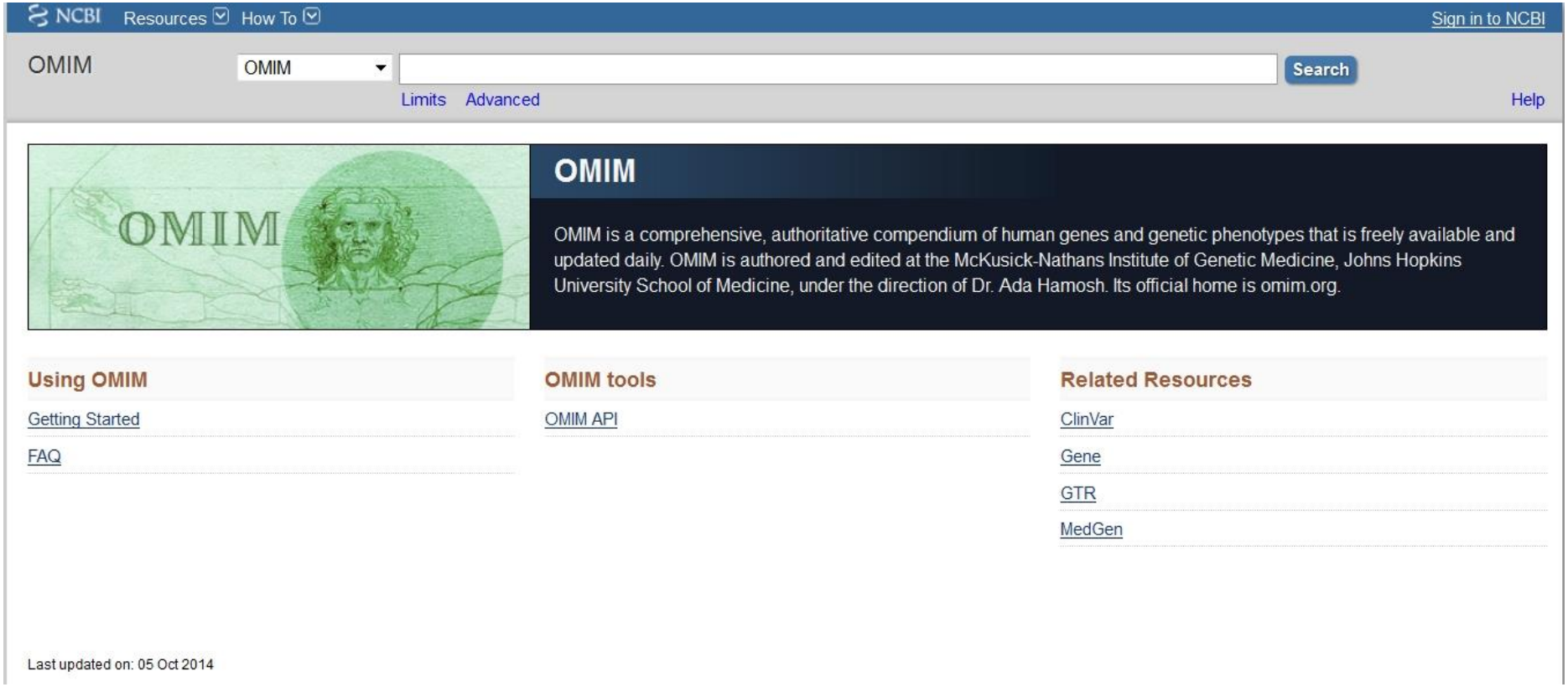
(Barrenas et al 2009: complex disease network – nodes are diseases)

Historical overview: inter-relationships (networks)



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

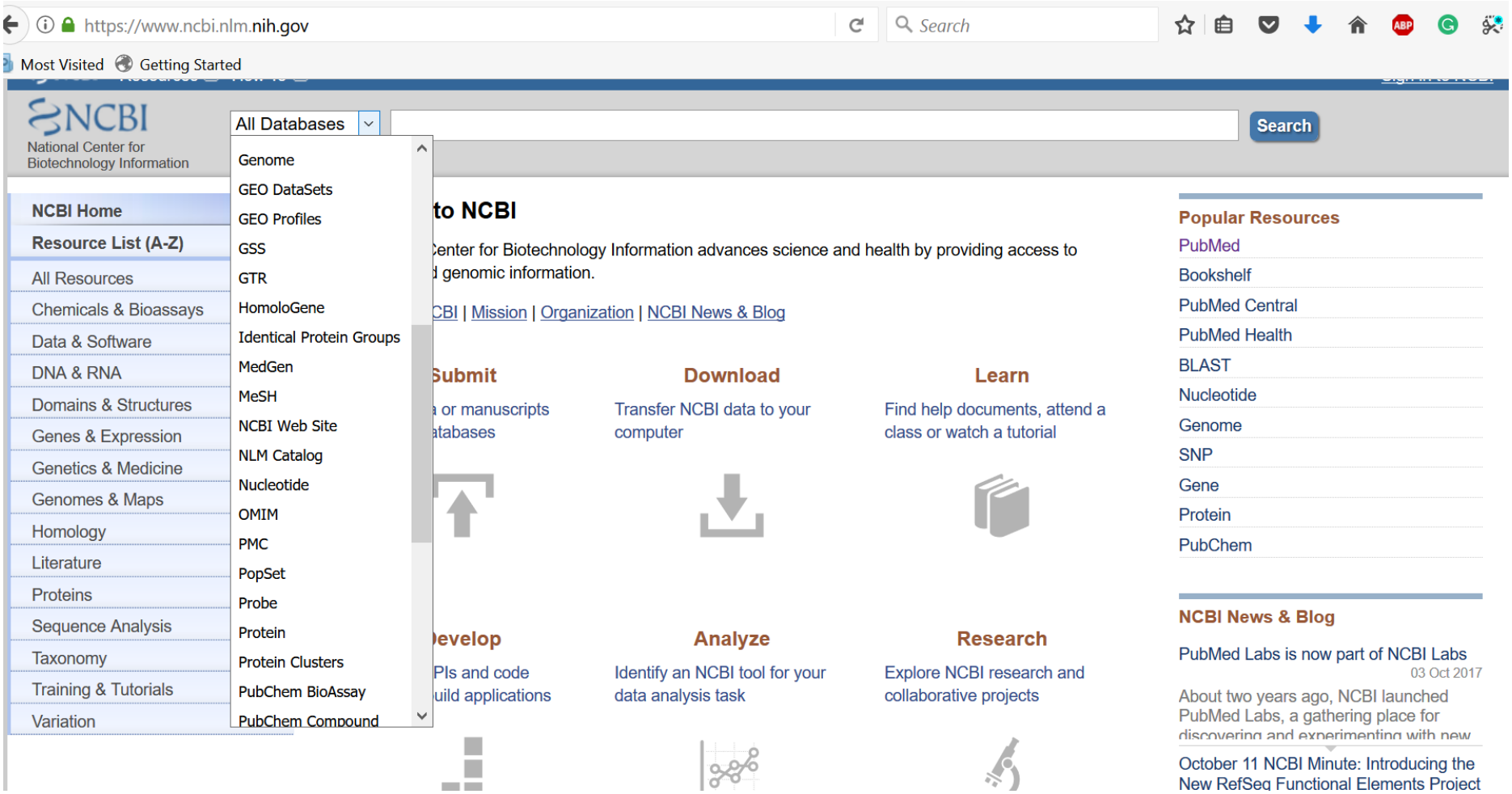
Historical overview: monitoring the progress



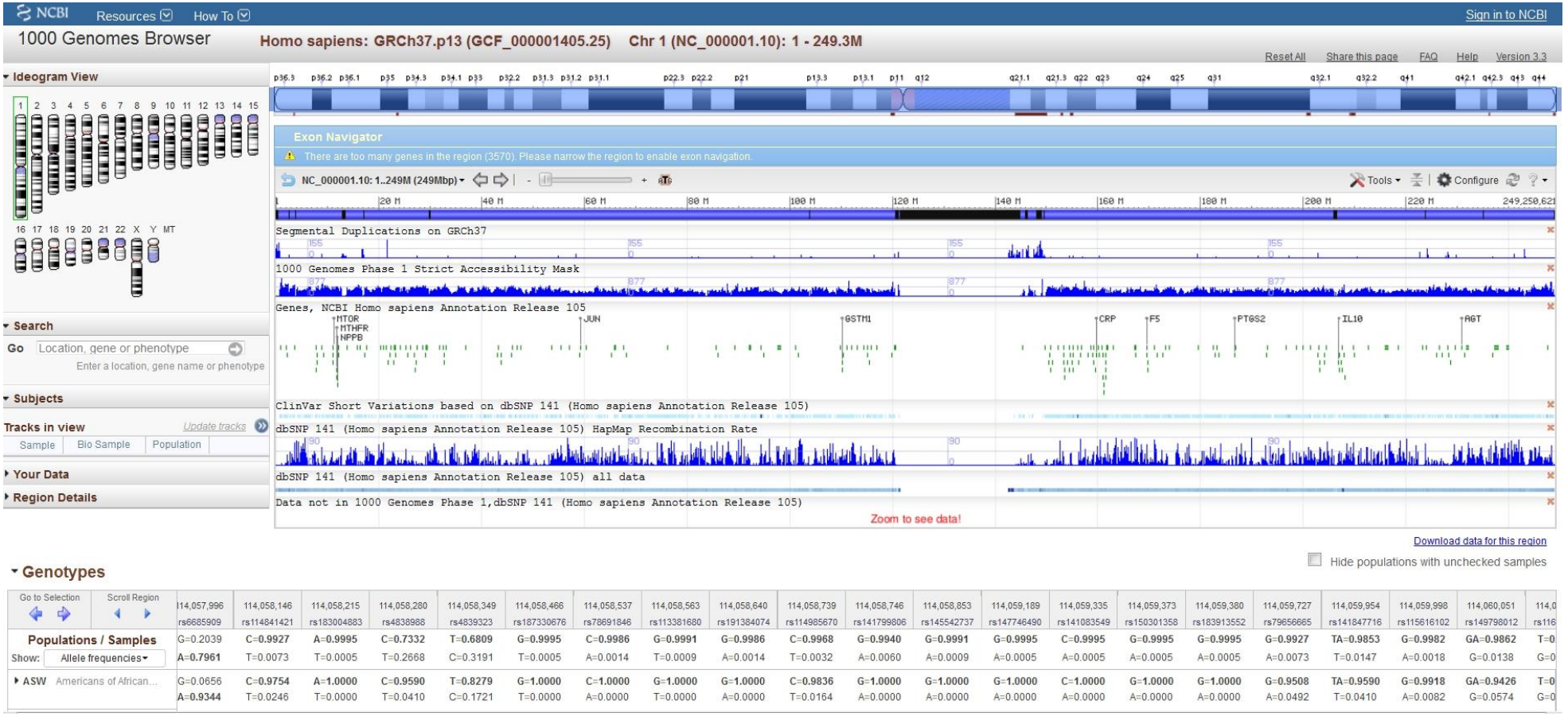
OMIM: molecular dissection of human disease

- 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



Historical overview: exome sequencing, full genome sequencing



2 The rise of GWAs

(slide Doug Brutlag 2010)

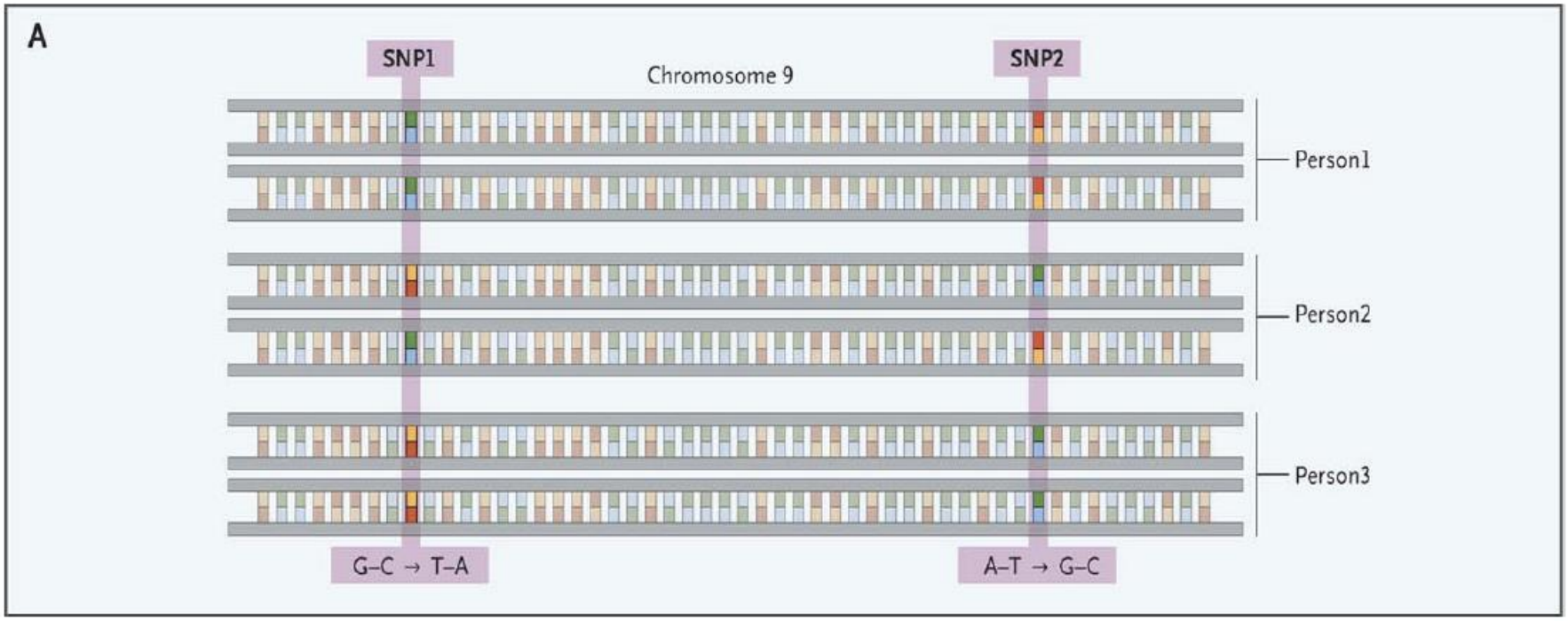
What are GWAs?

- 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.

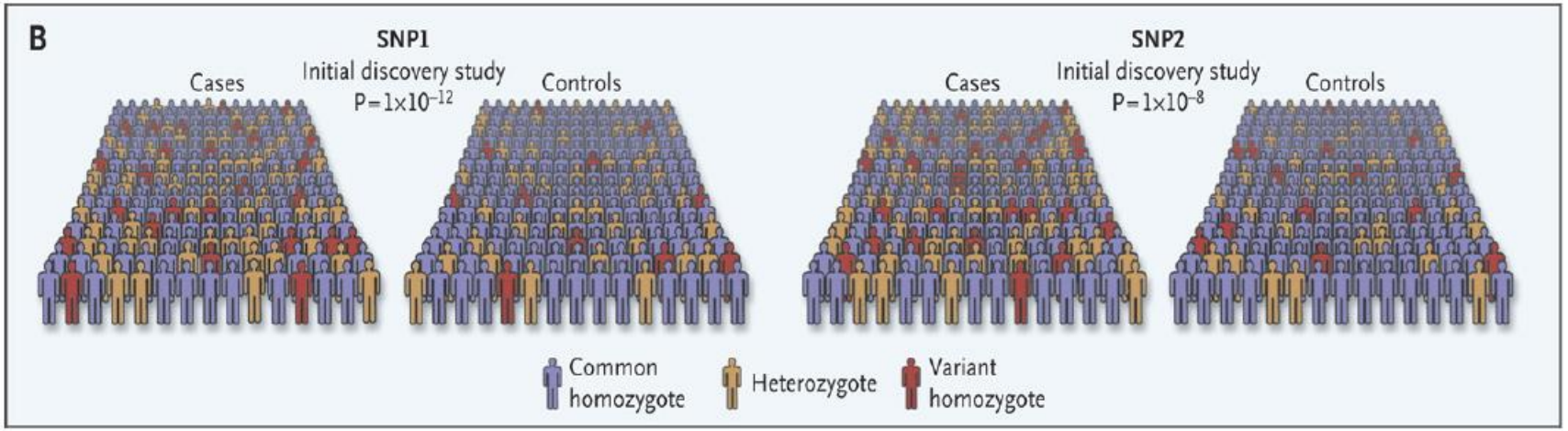
- **Recall**: a **trait** can be defined as a coded phenotype, a particular characteristic such as hair color, BMI, disease, gene expression intensity level, …

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)



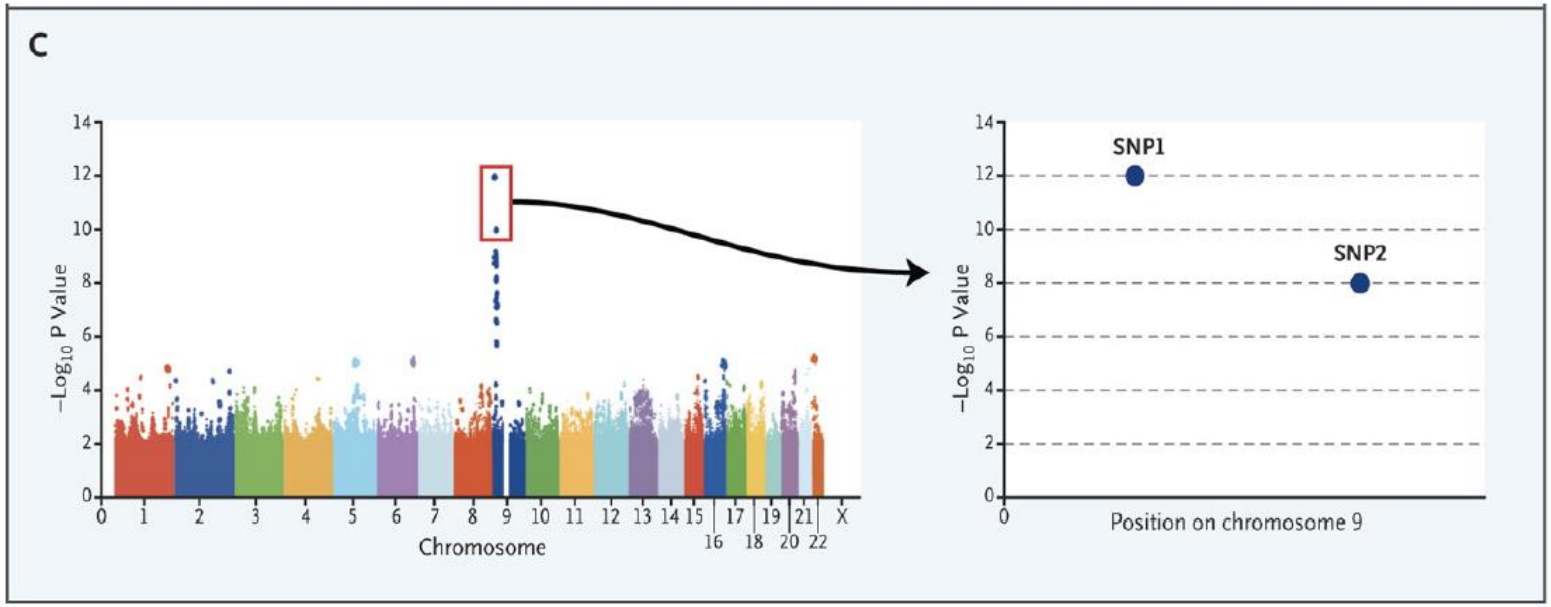
Genome-wide association studies: basic principles



- 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)

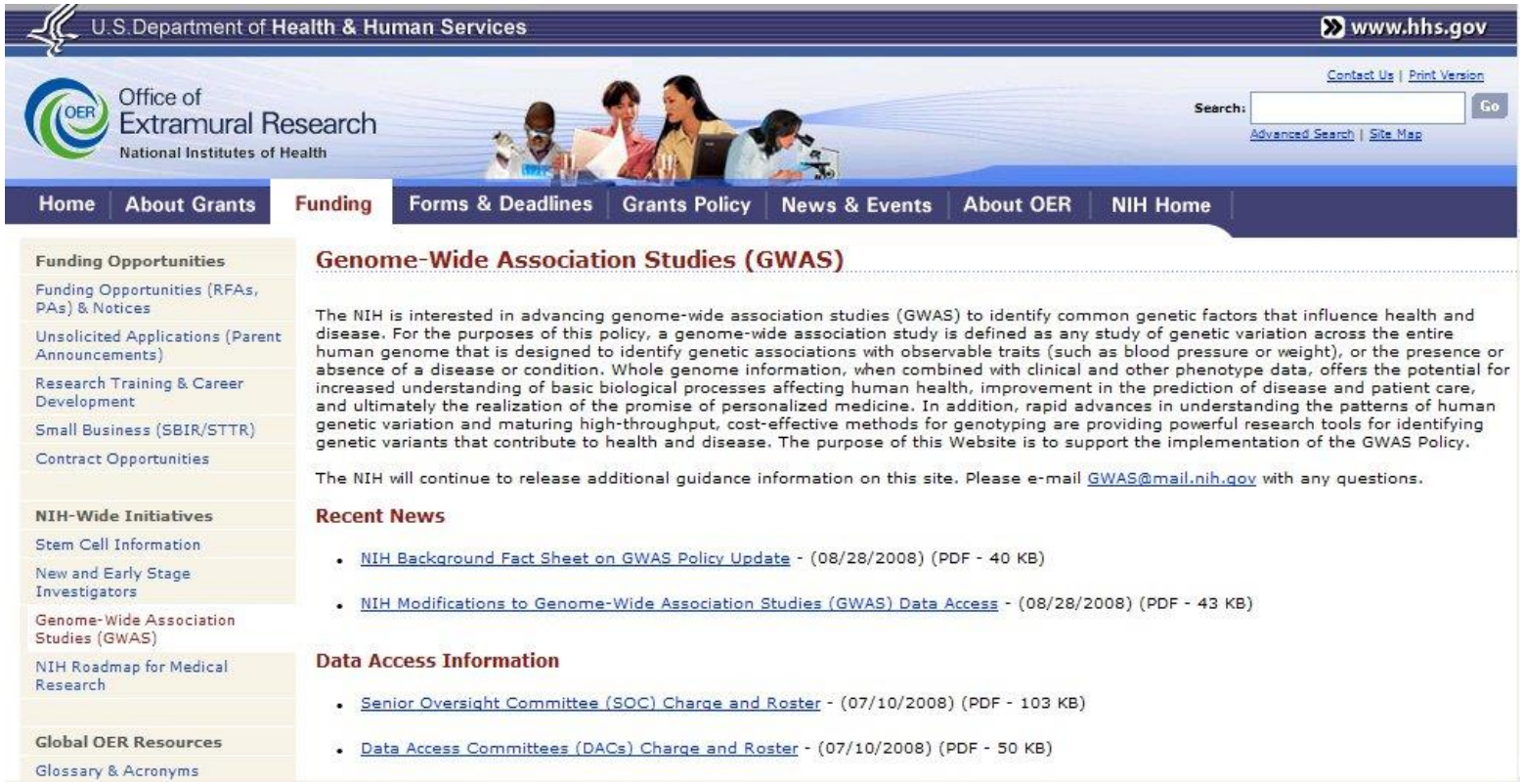
Genome-wide association studies: basic principles



- 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)

How can we use genome-wide association studies results?

- Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease.



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

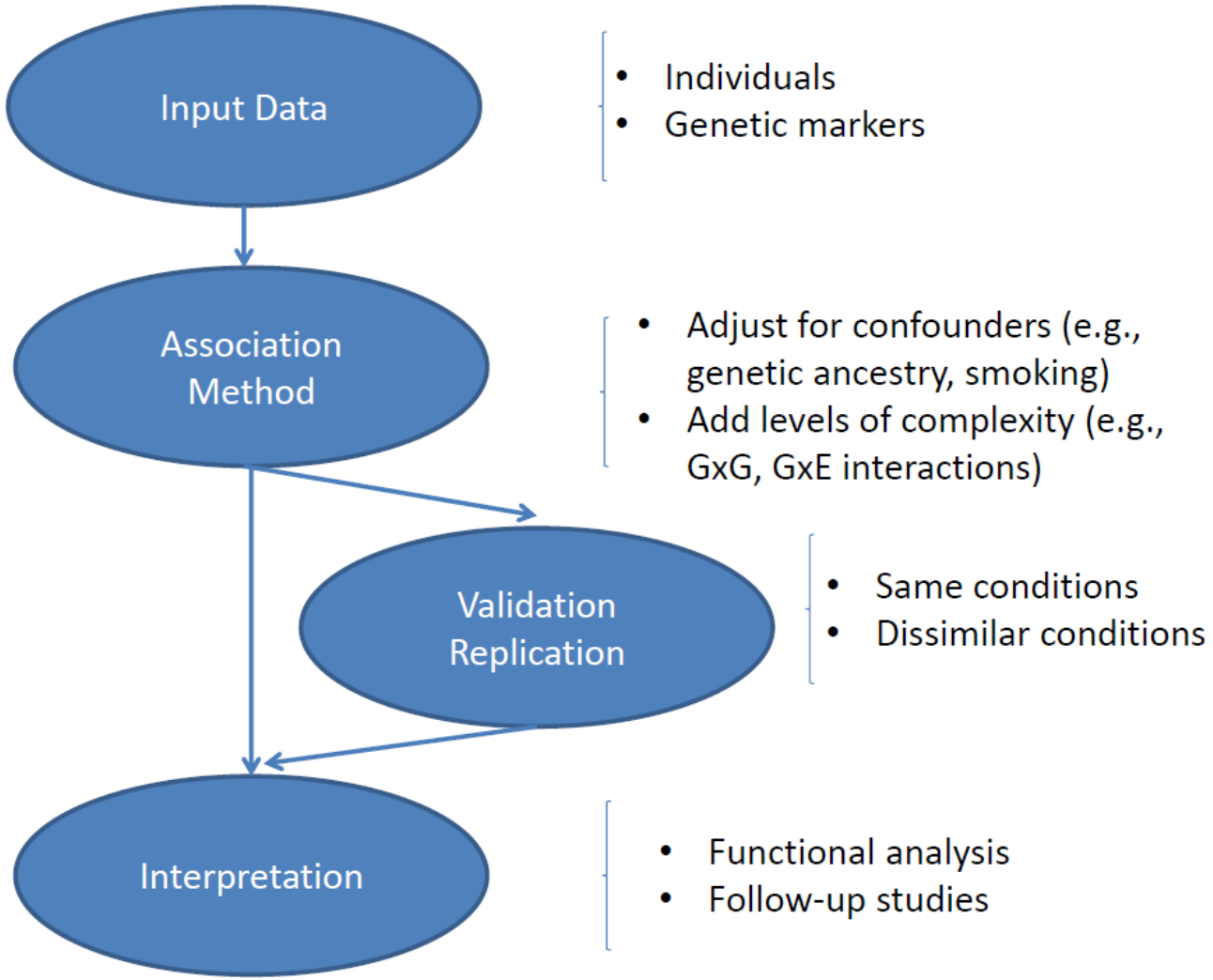
2317 studies (6/10/2014)

(Entries 1-50 of 2317)

Page 1 of 47 Next > Last >>

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

Genome-wide association studies: key components



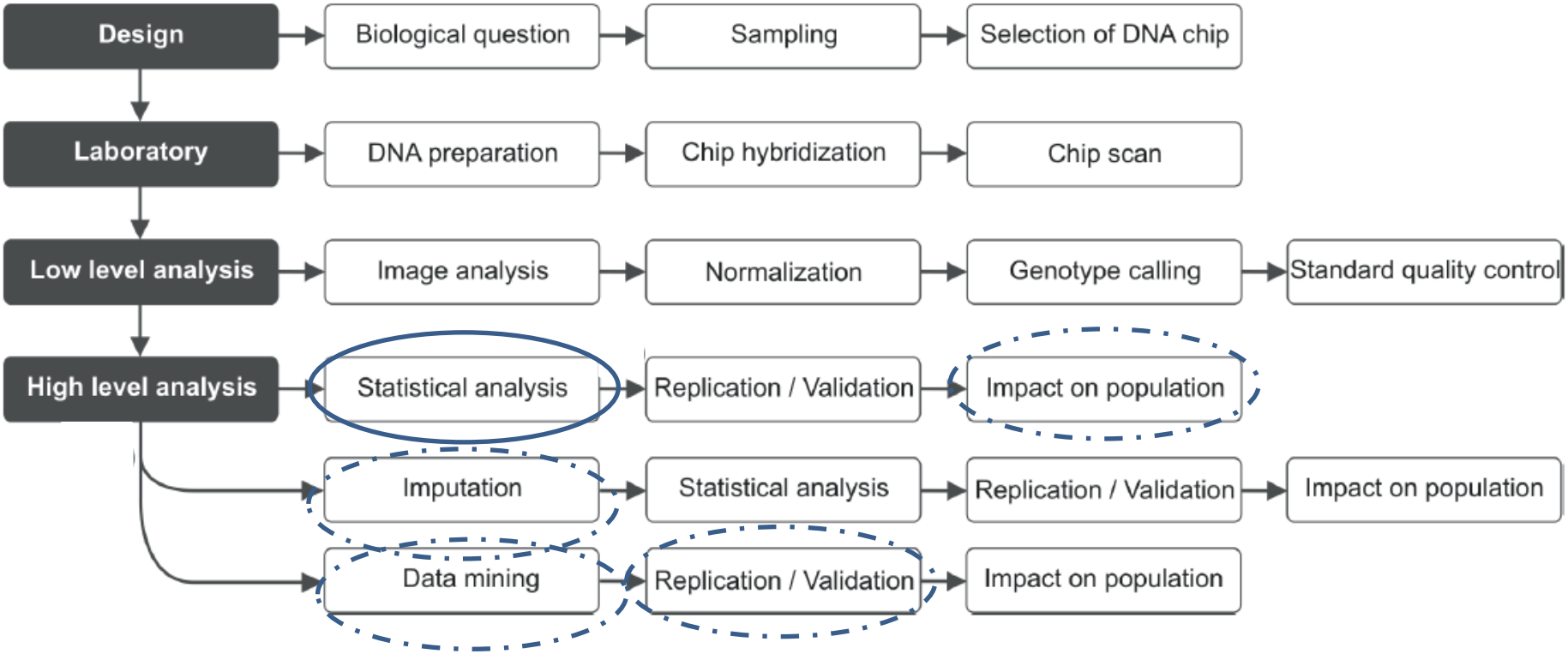
Genome-wide association studies: key components

- 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)

Rise of bioinformatics determines rise of GWAs (1)

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

Genetics and population analysis

GenABEL: an R library for genome-wide association analysis

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

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

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

Advance Access publication March 23, 2007

Associate Editor: Martin Bishop

ABSTRACT

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

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

Contact: i.aoultchenko@erasmusmc.nl

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

2 IMPLEMENTATION

2.1 Objective (1)

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

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 Li[2]

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

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

Advance Access publication November 15, 2007

Associate Editor: Martin Bishop

ABSTRACT

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

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

Contact: chun.li@vanderbilt.edu

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

2 METHODS

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

2.1 Phased input data and control file

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

2.2 Determination of disease model

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

Rise of bioinformatics determines rise of GWAs (4)

BIOINFORMATICS APPLICATIONS NOTE

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

Genome analysis

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

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

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

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

Advance Access publication January 25, 2009

Associate Editor: John Quackenbush

ABSTRACT

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

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

Contact: brian.stevenson@licr.org

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

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

2 PROGRAM OVERVIEW

2.1 Cache and memory management

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

2.2 Data import and export

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

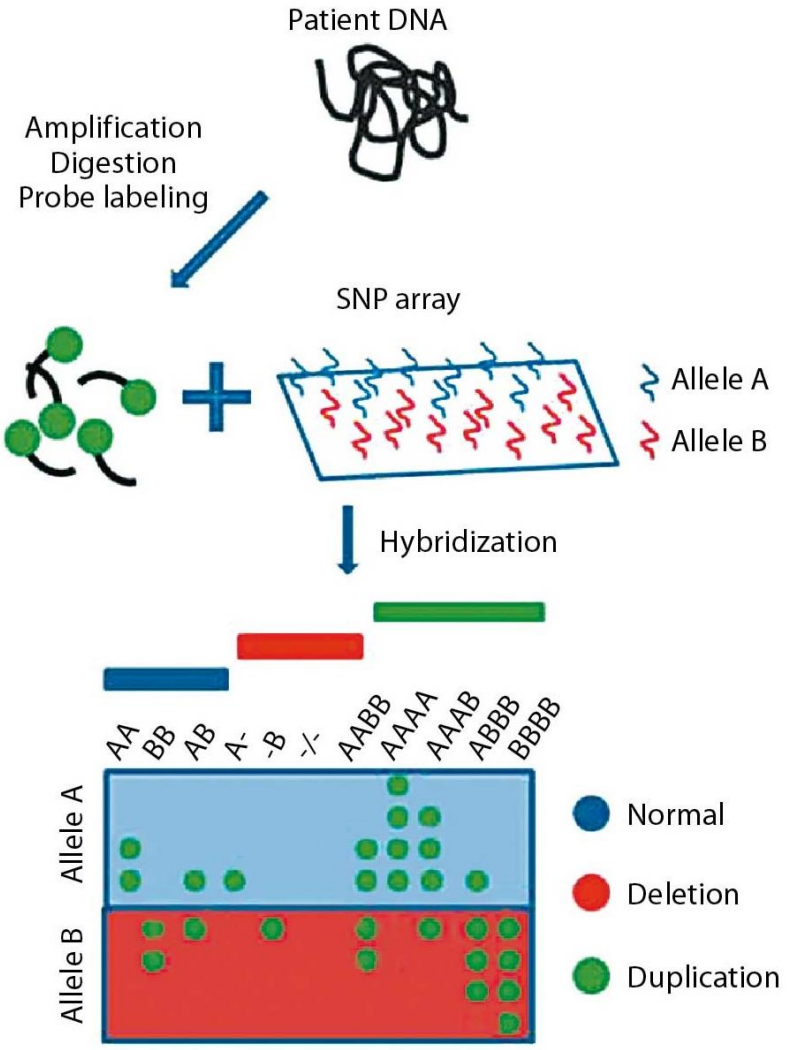
3 Study Design

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?

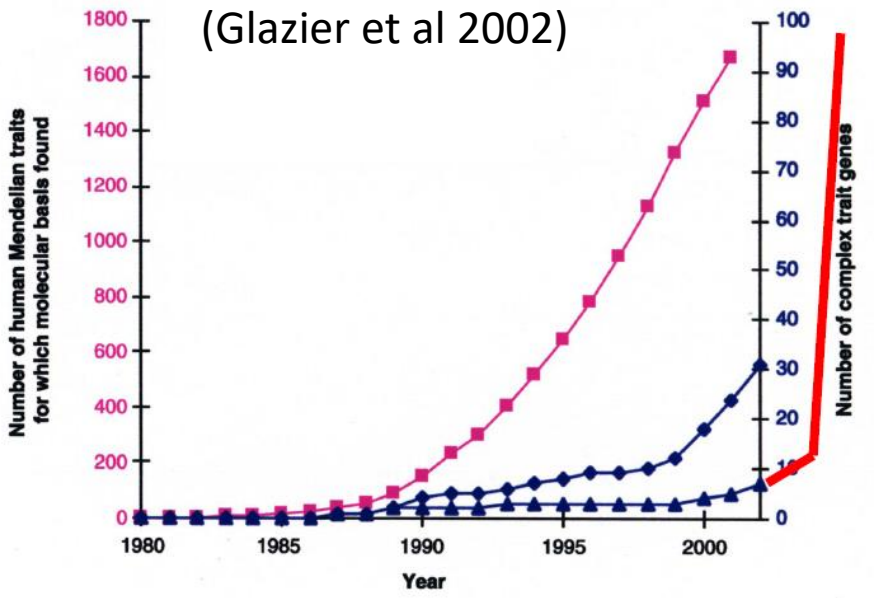
From common variants towards including rare variants

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

  - This hypothesis argues that **rare DNA sequence variations**, each with **relatively high 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

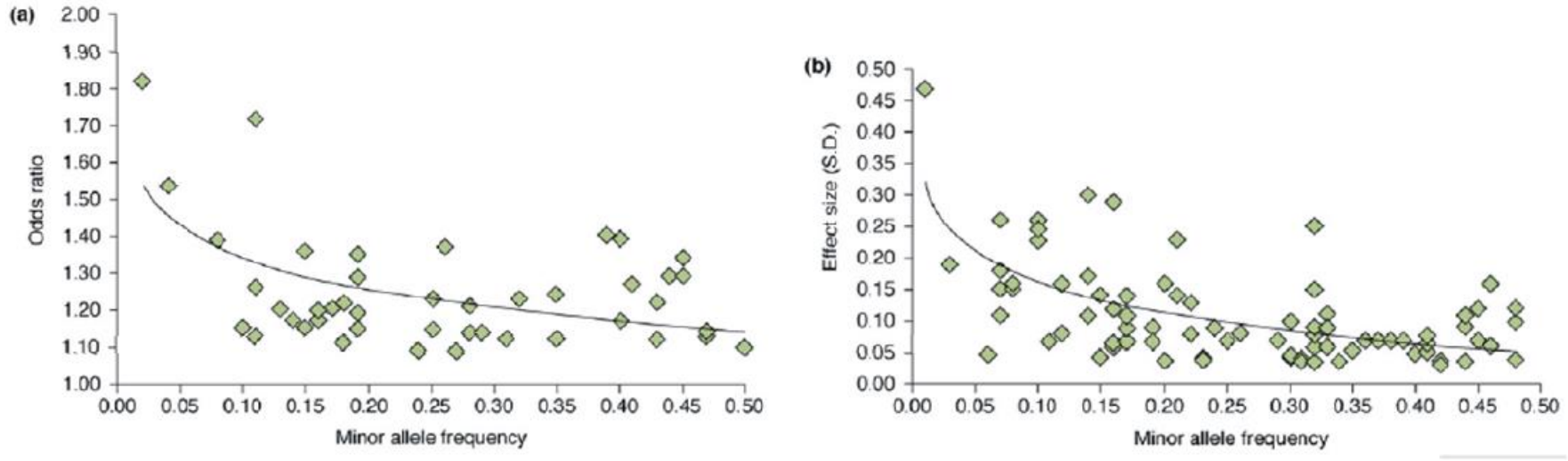
does NOT exhibit classic Mendelian inheritance attributable to a single gene;

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:

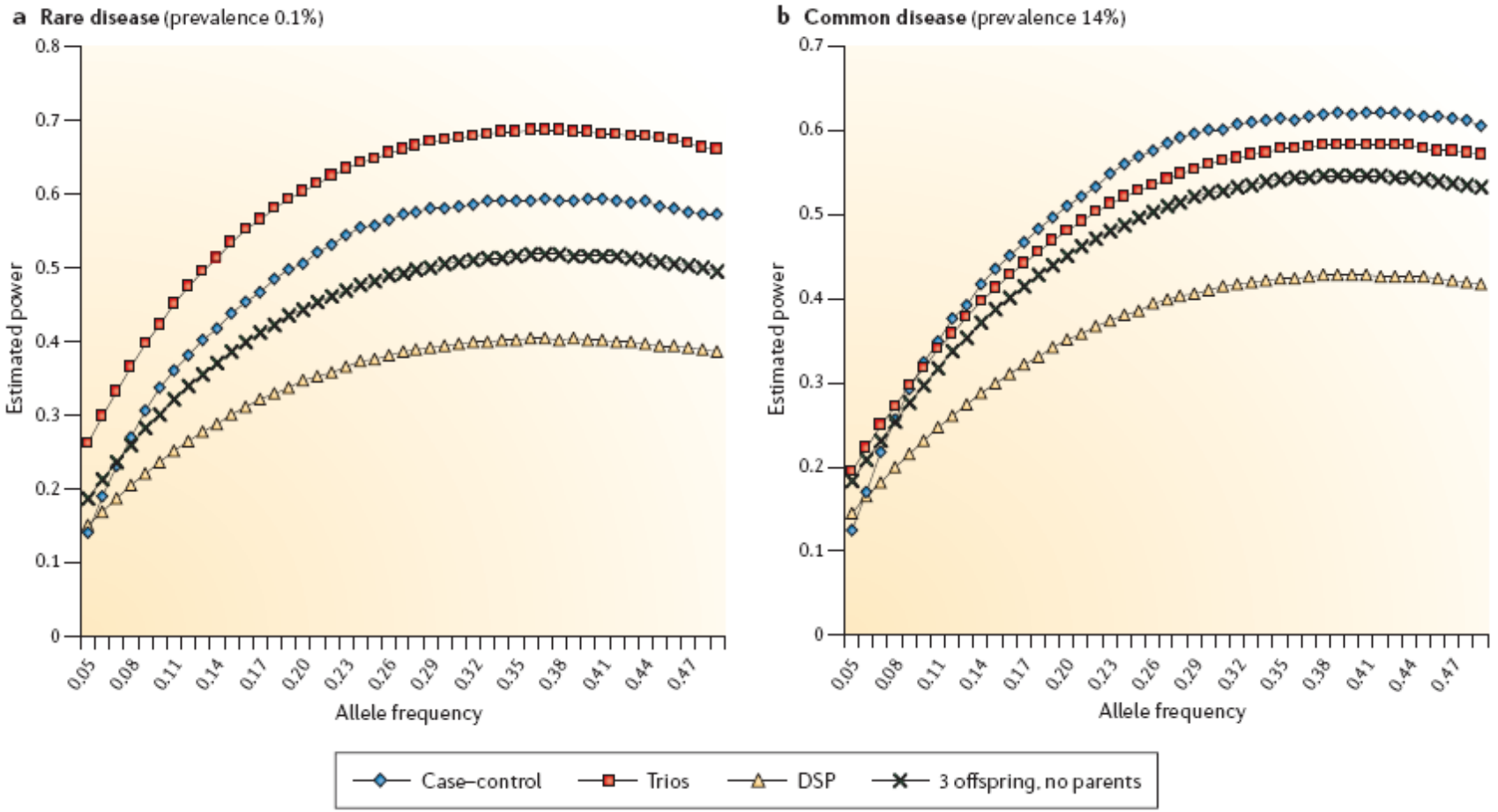
1. Controls immune to population stratification (no association without linkage, no "spurious" (false positive) association)
2. Checks for Mendelian inheritance possible (fewer genotyping errors)
3. Parental phenotyping not required (late onset diseases)
4. Simple logistics for diseases in children
5.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

tped file

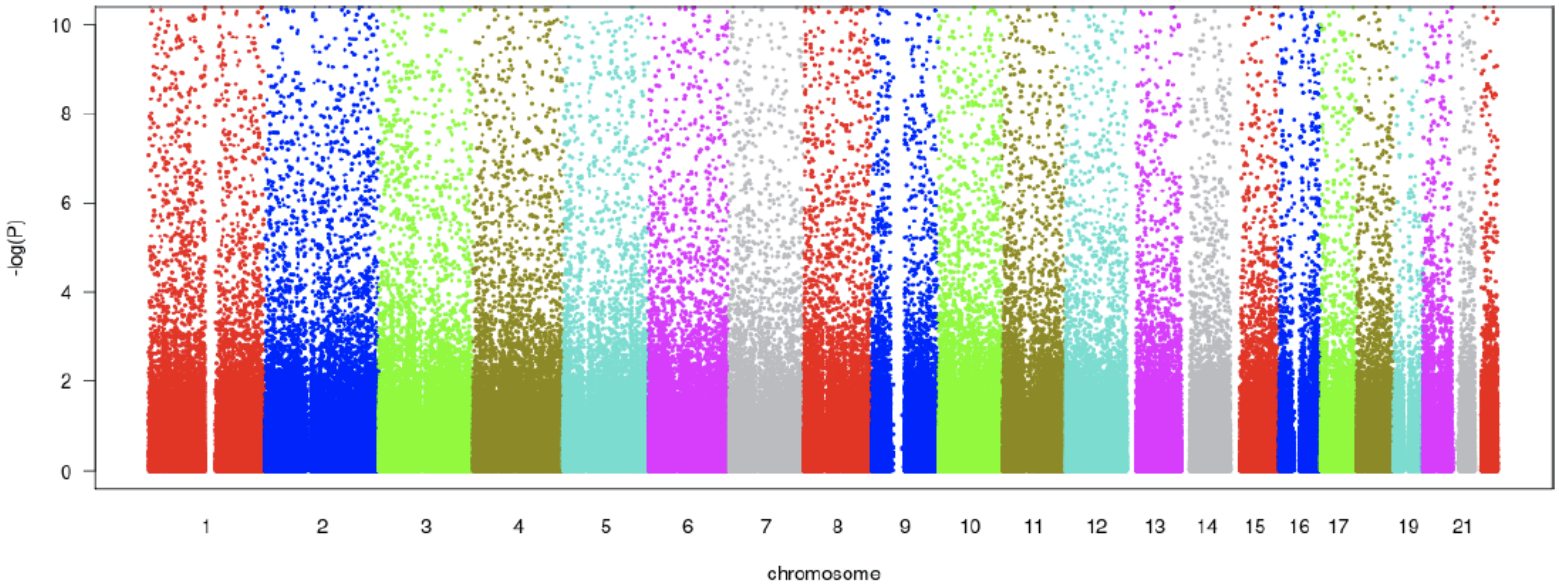
tfam file: First 6 columns of standard ped file

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

tfam file

Why is quality control (QC) important?

BEFORE QC → true signals are lost in false positive signals

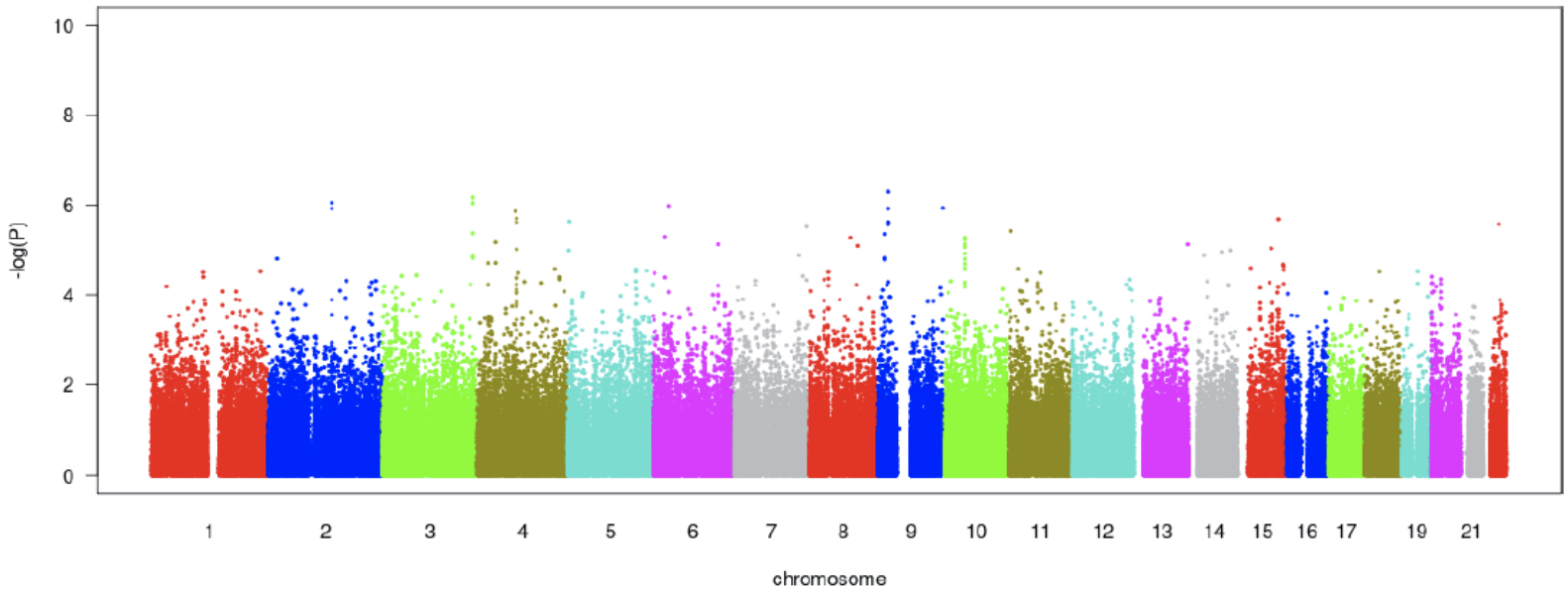


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(Ziegler and Van Steen 2010)

Why is quality control important?

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



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

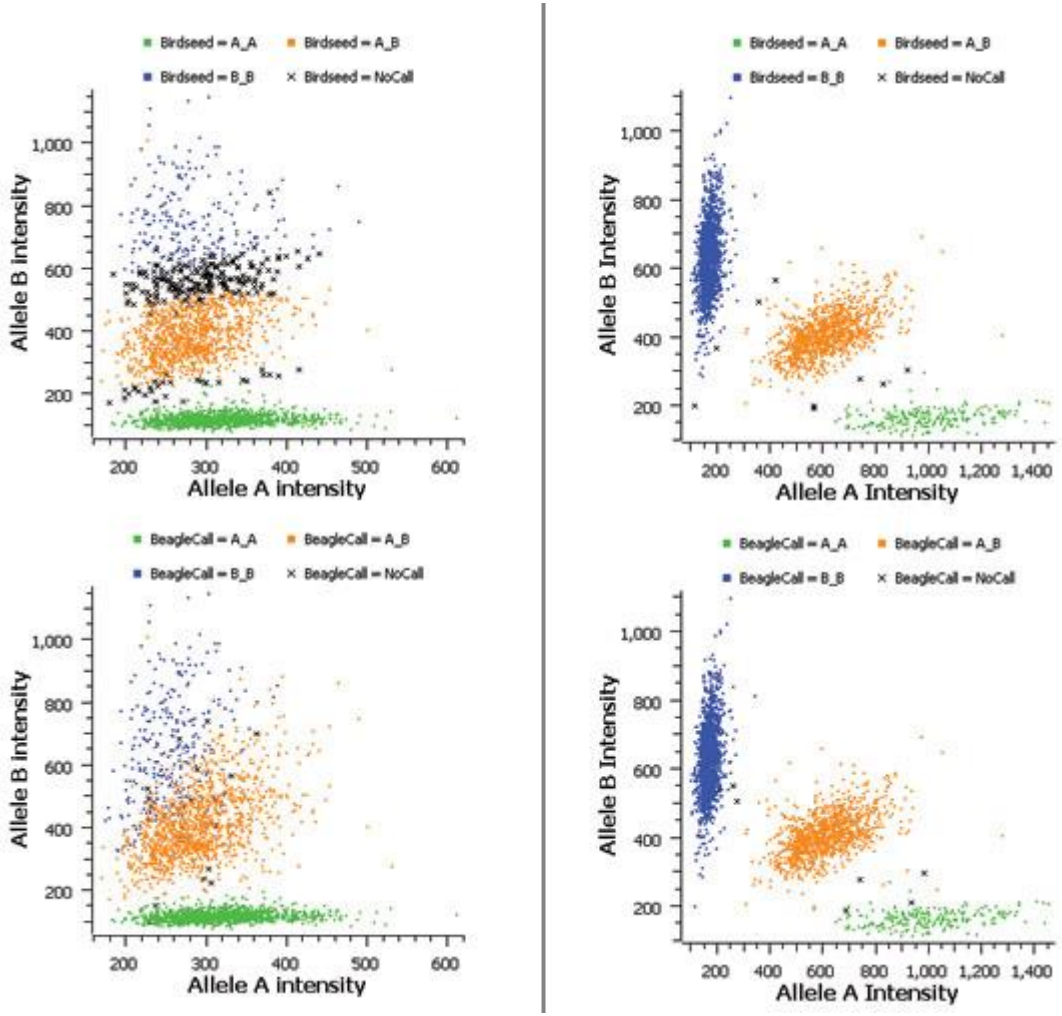
SNPs passing standard quality control: 270,701

(Ziegler and Van Steen 2010)

What is the standard quality control?

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

Marker level QC thresholds may be genotype calling algorithm dependent



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

Upper panels: Birdseed genotypes

Lower panels: BEAGLECALL genotypes.

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

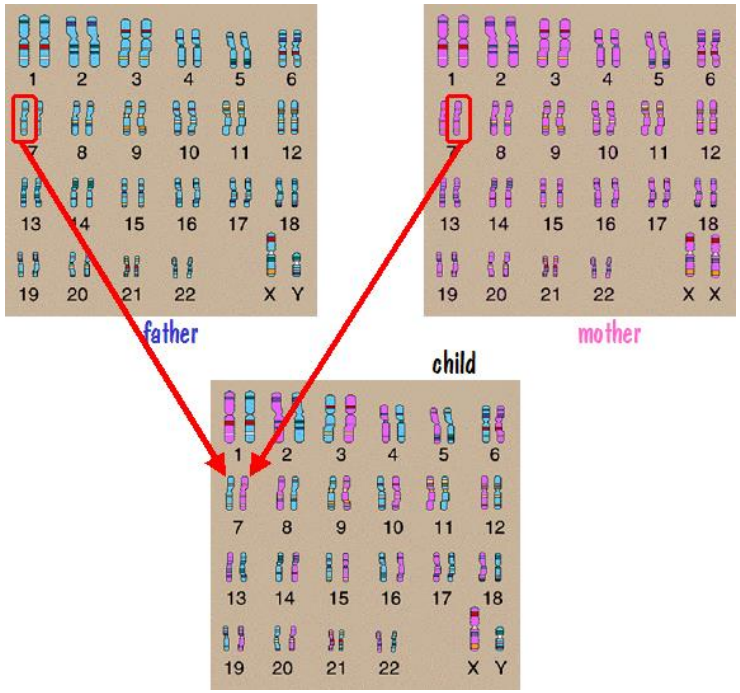
(Golden Helix Blog)

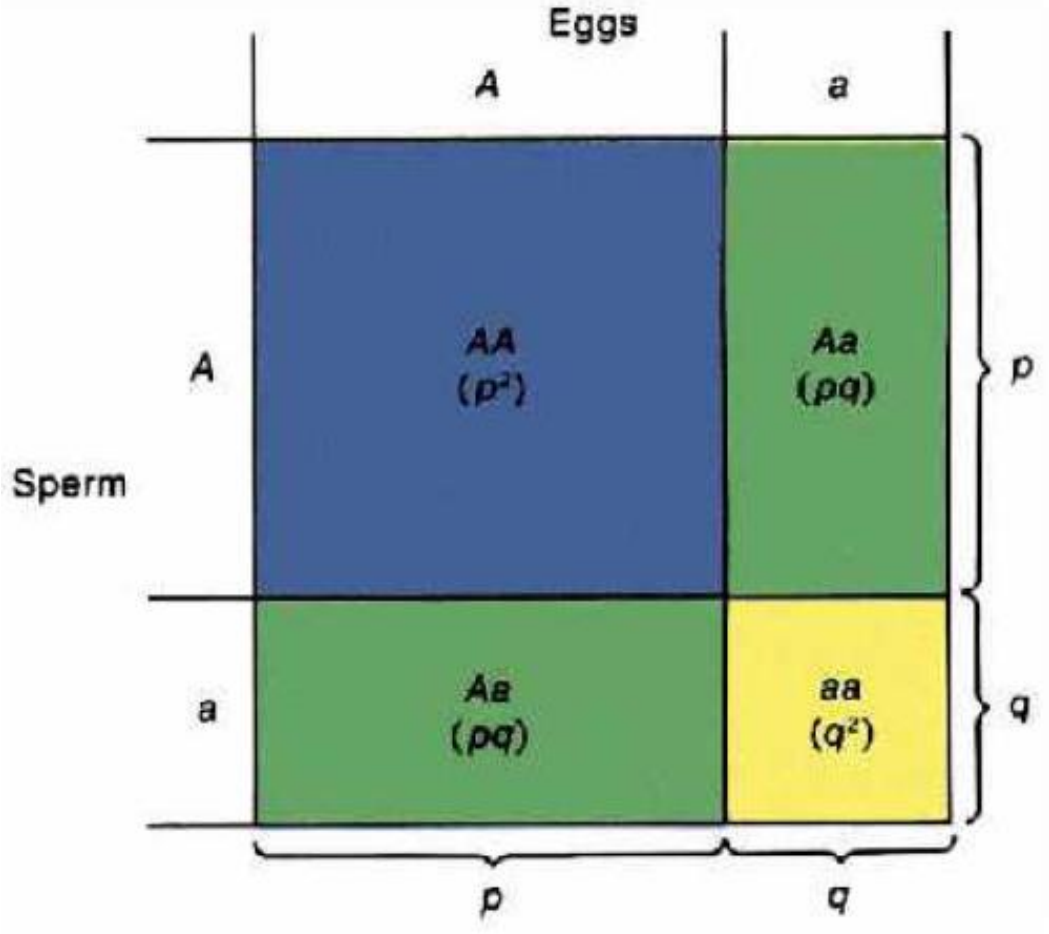
What entails quality control at the marker level?

- **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)

What are 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)**

What are “the Travemünde criteria”?

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

(Ziegler 2009)

The Travemünde criteria

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

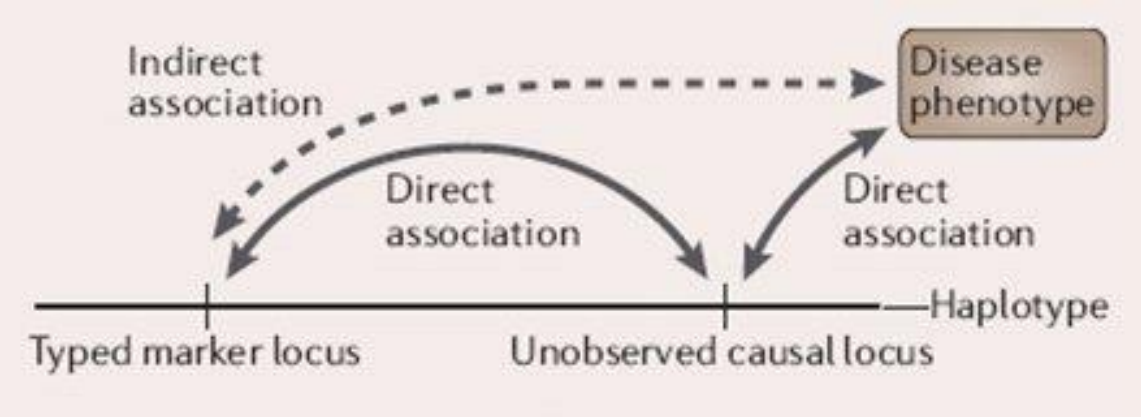
(Ziegler 2009)

4.b Linkage disequilibrium

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

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

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



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

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.

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).

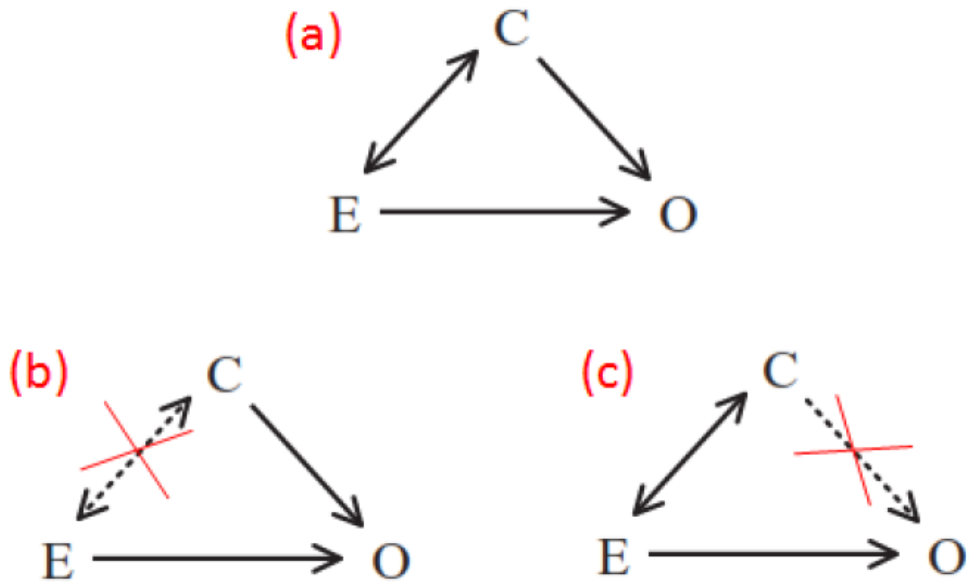


Figure 1: Schematic illustration of confounding control. Arrows represent causal effects, double arrows associations of any nature. E = exposure, C = confounder, O = outcome.

5 Analysis Steps

5.a Testing for Associations

The role of regression analysis

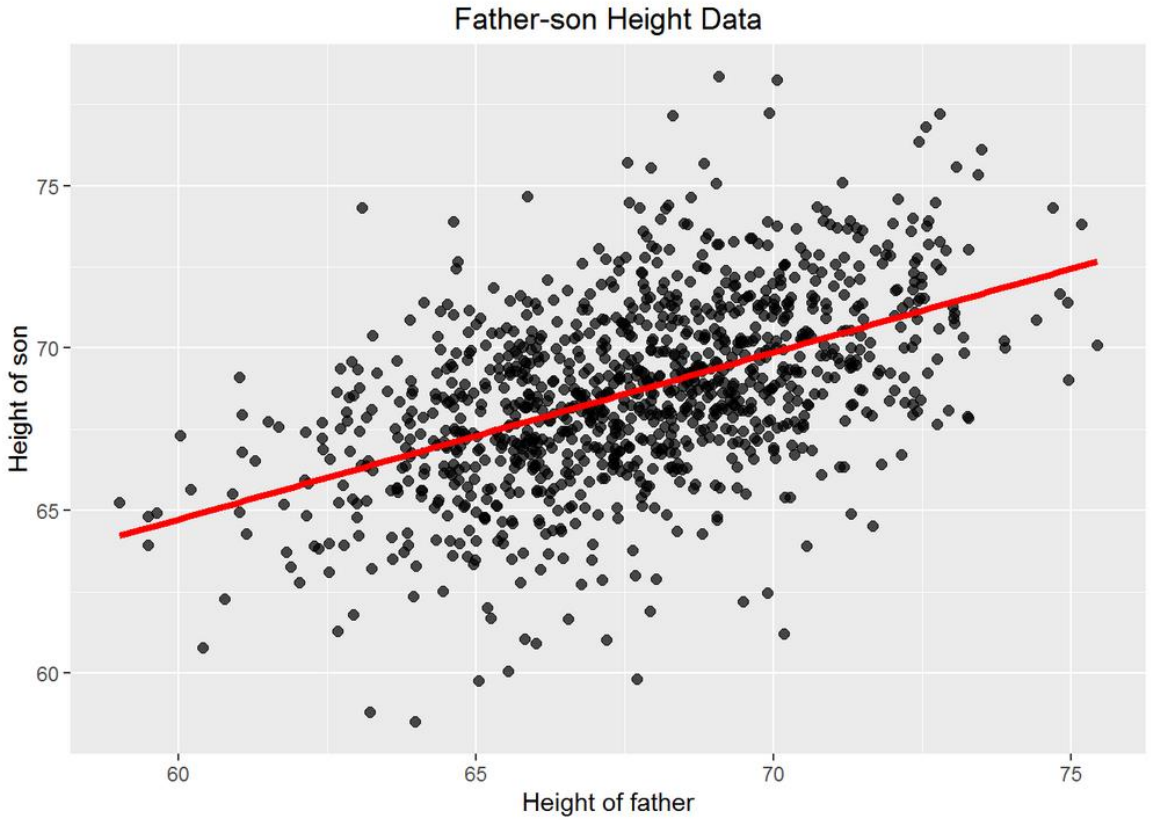
- Galton used the following equation to explain the phenomenon that sons of tall fathers tend to be tall but not as tall as their fathers while sons of short fathers tend to be short but not as short as their fathers:

$$\frac{y - \bar{y}}{SD_y} = r\frac{(x - \bar{x})}{SD_x}.$$

This effect is called the **regression effect**.

The use of regression analysis

- **regression line** goes through (mean Y, mean X)



(https://rstudio-pubs-static.s3.amazonaws.com/204984_dd2112475db84af2a03260c4a4f830ac.html)

The use of regression analysis

- **Regression analysis** is used for explaining or modeling the relationship between a single variable Y, called the response, output or dependent variable, and one or more predictor, input, independent or explanatory variables, X_1, …, X_p.
- When p=1 it is called **simple regression** but when p > 1 it is called **multiple regression**.
- When there is more than one Y, then it is called **multivariate** regression
- Regression analyses have several possible objectives including
  - Prediction of future observations.
  - Assessment of the effect of, or relationship between, explanatory variables on the response.
  - A general description of data structure

What is the difference between association and prediction?

The linear regression model

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

- y: response variable.
- $x_1, \ldots, x_k$: regressor variables, independent variables.
- $\beta_0, \beta_1, \ldots, \beta_k$: 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

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

- 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:

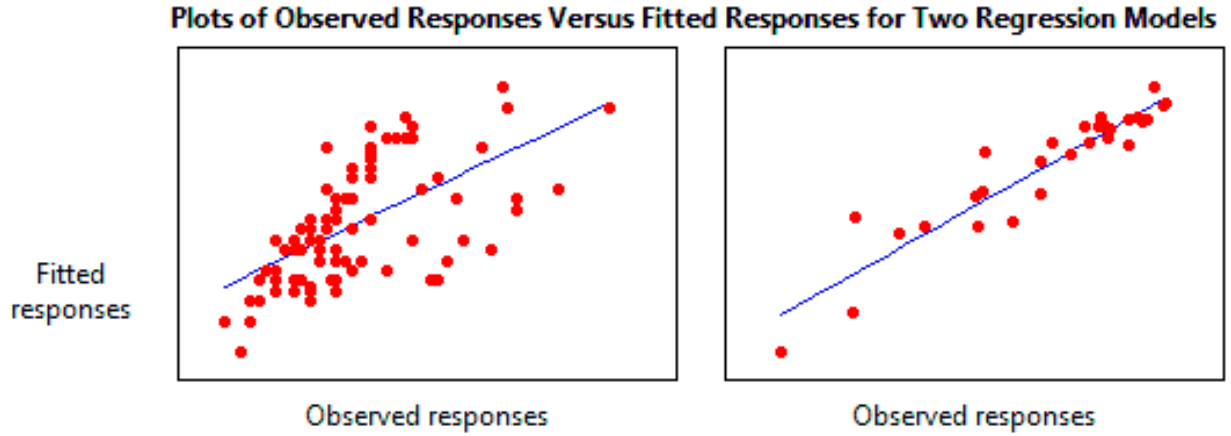
R-squared = Explained variation / Total variation

(compare with well-known formula for 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.

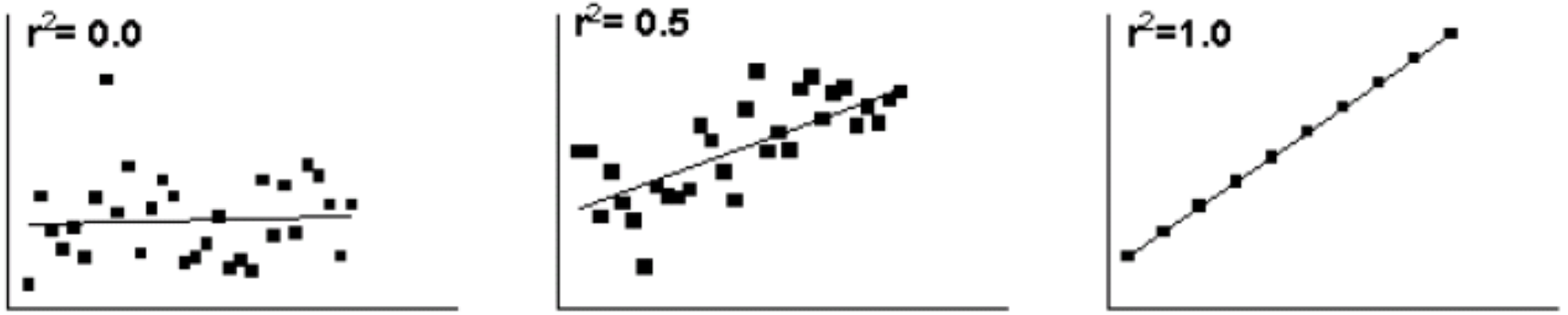


- The regression model on the left accounts for 38.0% of the variance while the one on the right accounts for 87.4%. The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line.

Coefficient of determination ~ squared correlation coefficient r^2

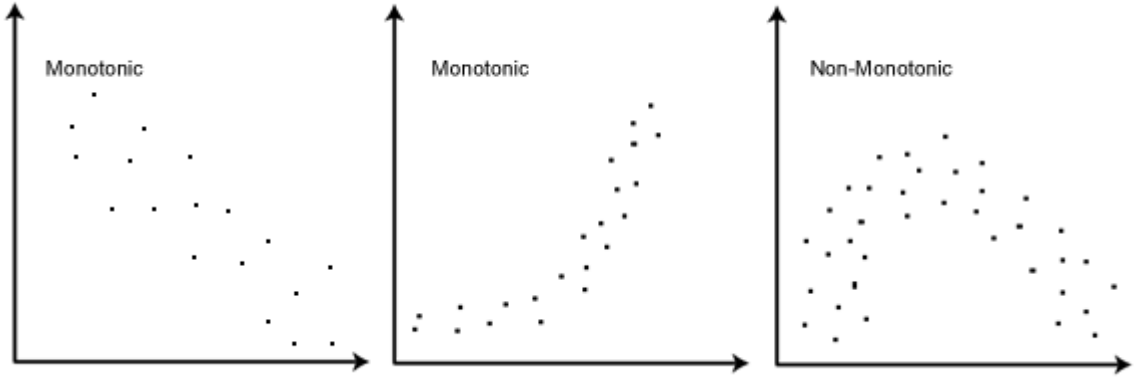
- An R^2 value of 0.0 means that knowing X does not help you predict Y. There is no linear relationship between X and Y, and the best-fit line is a horizontal line going through the mean of all Y values.

- When R^2 equals 1.0, all points lie exactly on a straight line with no scatter. Knowing X lets you predict Y perfectly.



Several formula exist for “correlation”?

- The **Spearman's rank-order correlation** is the nonparametric version of the **Pearson** product-moment **correlation**.
- Spearman's correlation determines the strength and direction of the **monotonic relationship** between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines.
- A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases.



How to compute Spearman's rank-order correlation?

- There are two methods to calculate Spearman's correlation depending on whether: (1) your data does not have tied ranks or (2) your data has tied ranks. The formula for when there are no tied ranks is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i = difference in paired ranks and n = number of cases.

- The formula to use when there are tied ranks is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where *i* = paired score.

General linear test approach

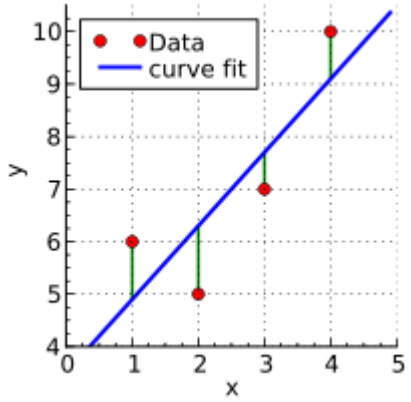
- The full model (continuous response, say “BMI”):

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

- 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):

$$\sum [Y - b_0 - b_1 X_1 - b_2 X_2]^2$$

$$= \sum (Y - \hat{Y})^2$$



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 H0 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

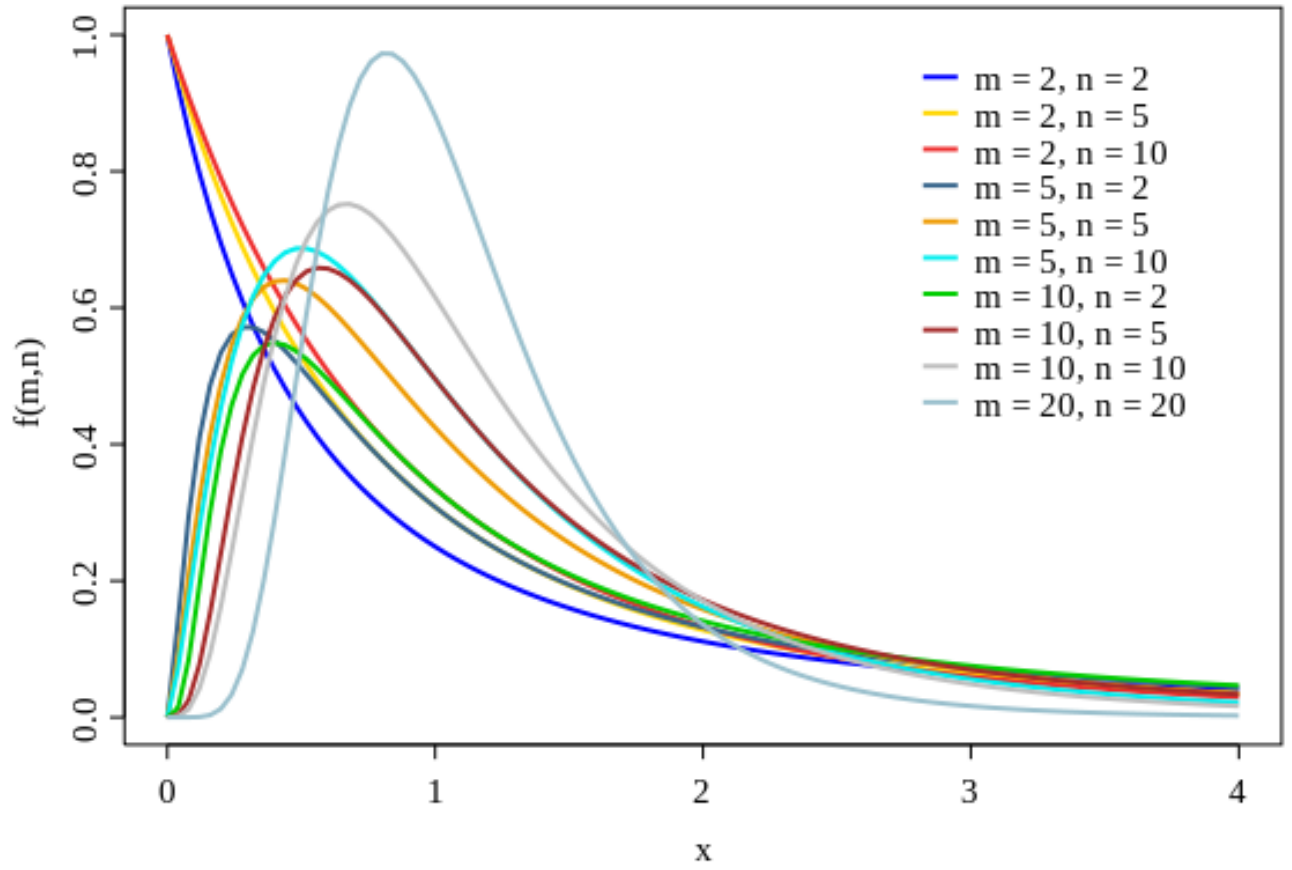
- The decision rule (for a given alpha level of significance) is:

  If $F^* \leq F(1-\alpha;\ df_R - df_F, df_F)$, *you cannot reject* H_0

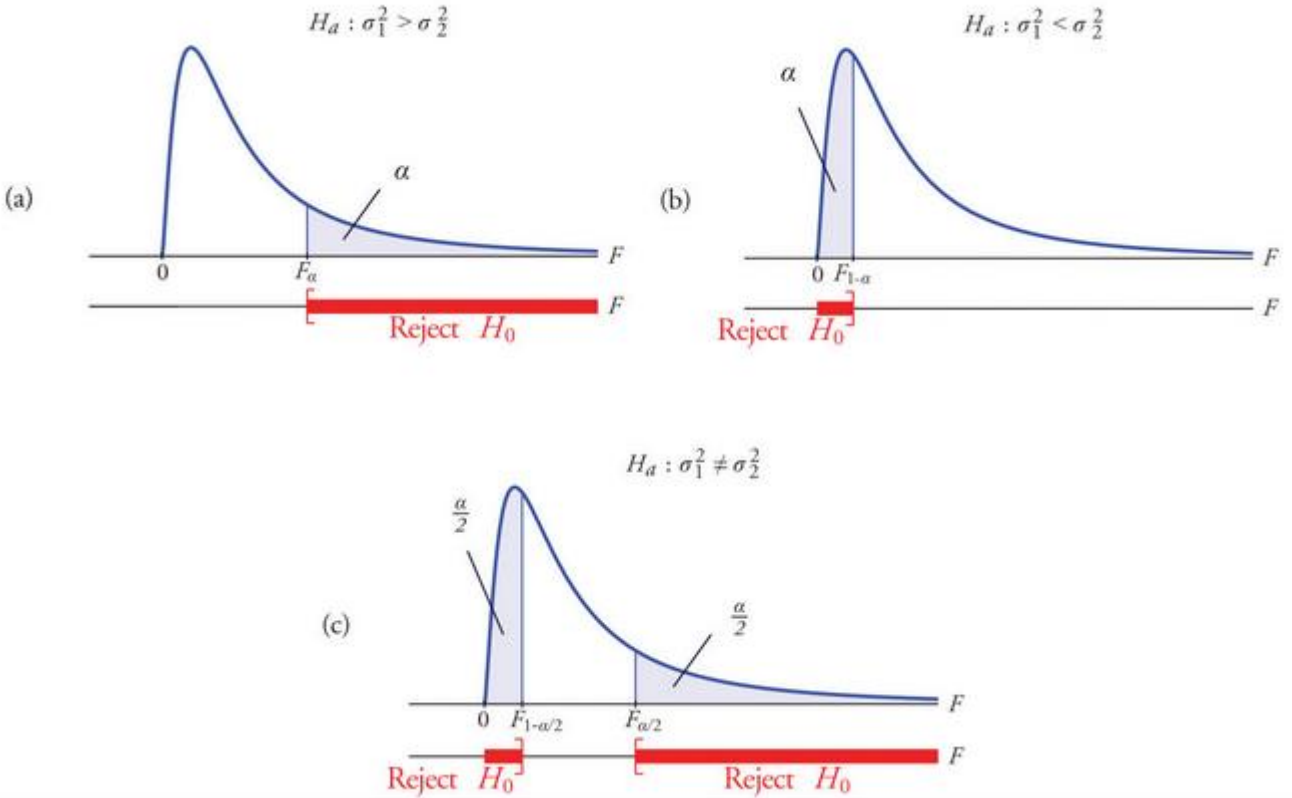
  If $F^* > F(1-\alpha;\ df_R - df_F, df_F)$, conclude H_1

Can you ever accept a null hypothesis?

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 $F \geq F_{\alpha/2}$

Tests in GWAS using the regression framework

- **Example 1:**

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

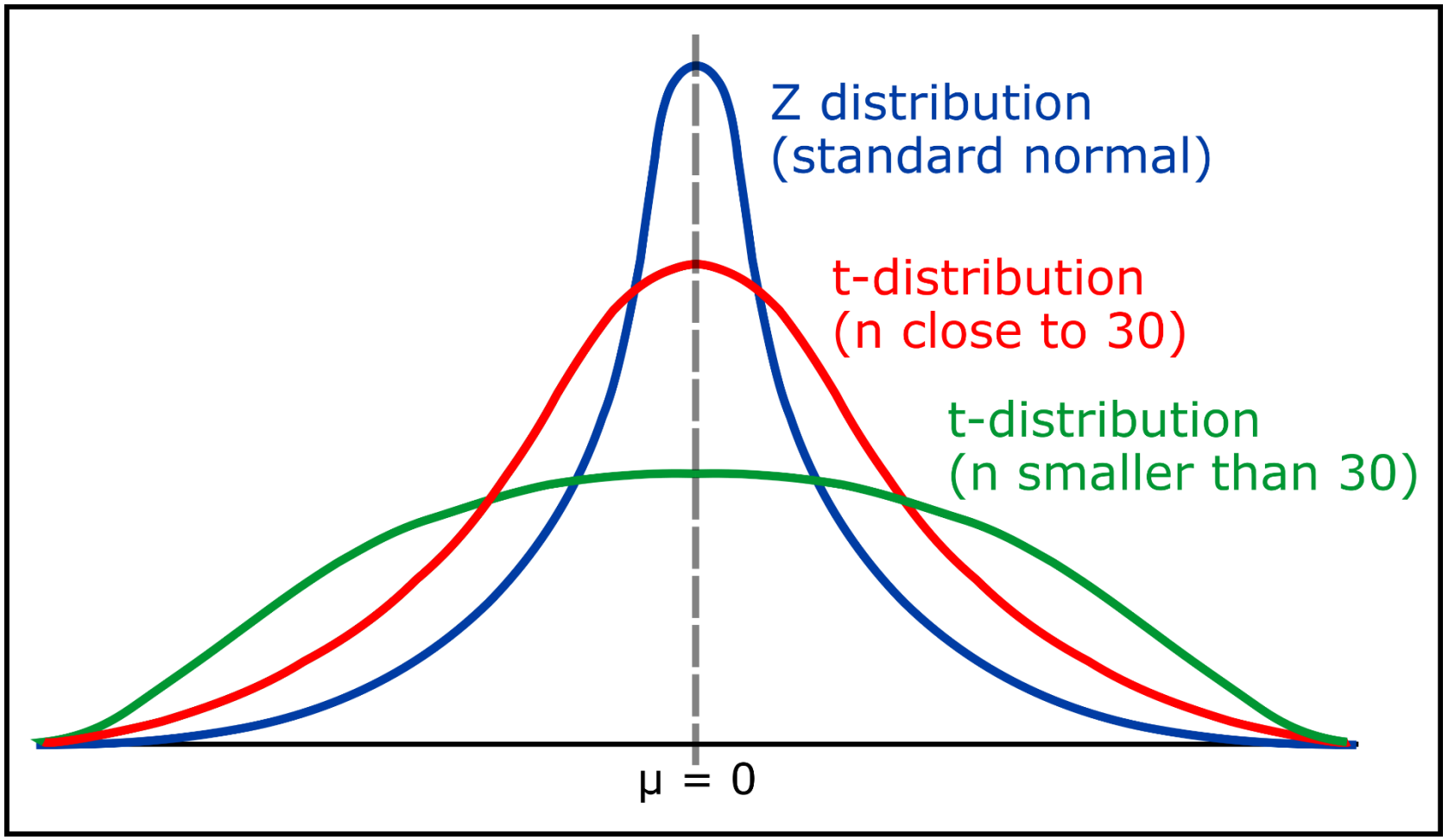
- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$
- $\mathrm{df_F} = \mathrm{n} - 2$ (this links to df in variance estimation)
- $\mathrm{df_R} = \mathrm{n} - 1$ (this links to df in variance estimation)

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

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

Why is the t-test more flexible?

What are different shapes of a t distribution (parameter n)?



Tests in GWAS using the regression framework

- **Example 2:**

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

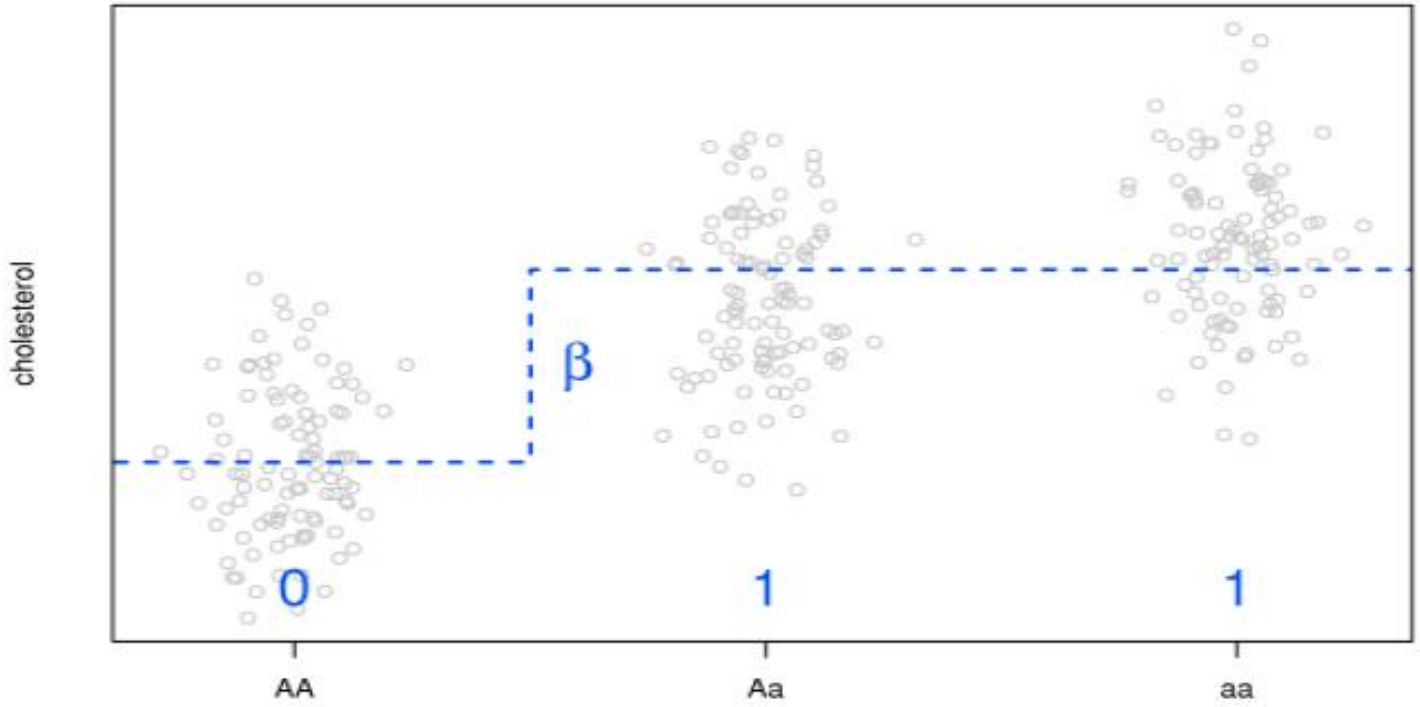
  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 \neq 0$
  - $\mathrm{df_F} = \mathrm{n} - 4$
  - $\mathrm{df_R} = \mathrm{n} - 3$

How many dfs would the corresponding F-test have?

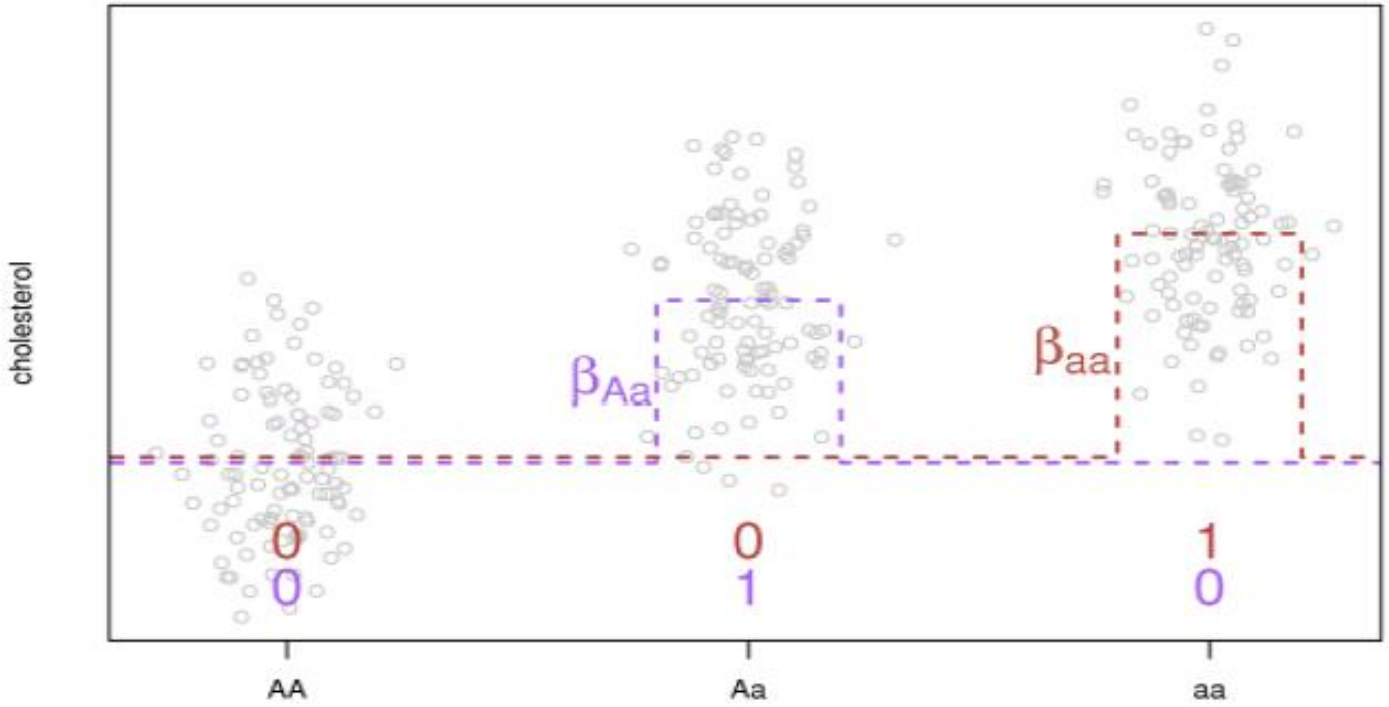
The impact of different encoding schemes for SNPs

	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?

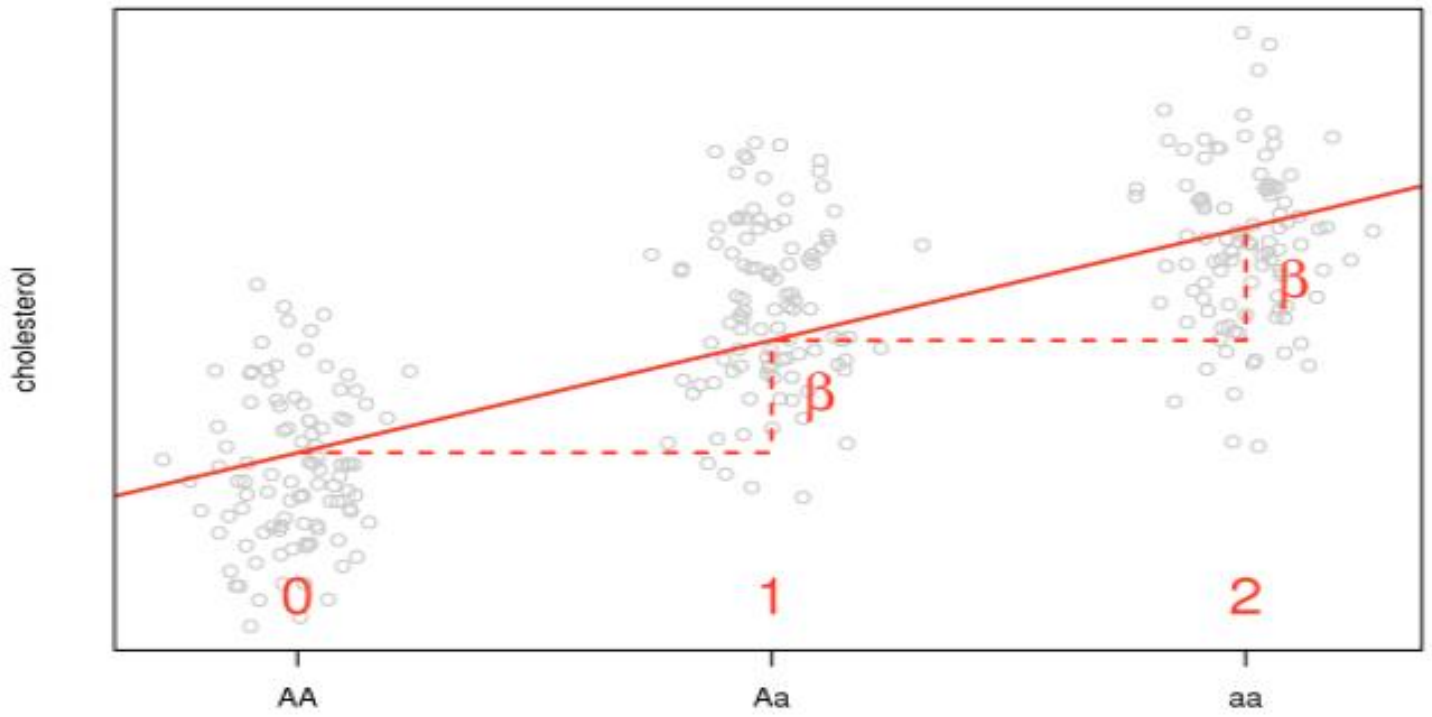


Which encoding scheme provides a good fit to the data?



Robust vs overkill ?

Which encoding scheme provides a good fit to the data?



Most commonly used

Regression analysis in R

- Main functions
  - The basic syntax for doing regression in R is **lm()** to fit linear models
  - The R function **glm()** can be used to fit generalized linear models (i.e., when the response is not normally distributed)
- General syntax rules in R model fitting are given on the next slide.

Regression analysis in R

Syntax	Model	Comments
$Y \sim A$	$Y = \beta_o + \beta_1 A$	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)
$Y \sim A + I(A\text{^}2)$	$Y = \beta_o + \beta_1 A + \beta_2 A^2$	Polynomial model; note that the identity function **I( )** allows terms in the model to include normal mathematical symbols.
$Y \sim A + B$	$Y = \beta_o + \beta_1 A + \beta_2 B$	A first-order model in A and B without interaction terms.
$Y \sim A:B$	$Y = \beta_o + \beta_1 AB$	A model containing only first-order interactions between A and B.
$Y \sim A*B$	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 AB$	A full first-order model with a term; an equivalent code is $Y \sim A + B + A:B$.
$Y \sim (A + B + C)\text{^}2$	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 AC$	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 $Y \sim A*B*C - A:B:C$.

Model diagnostics are model-dependent …

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

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

QQ plots for model diagnostics – Q for Quantile

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

  What is the 0.5 quantile for normally distributed data?

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

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

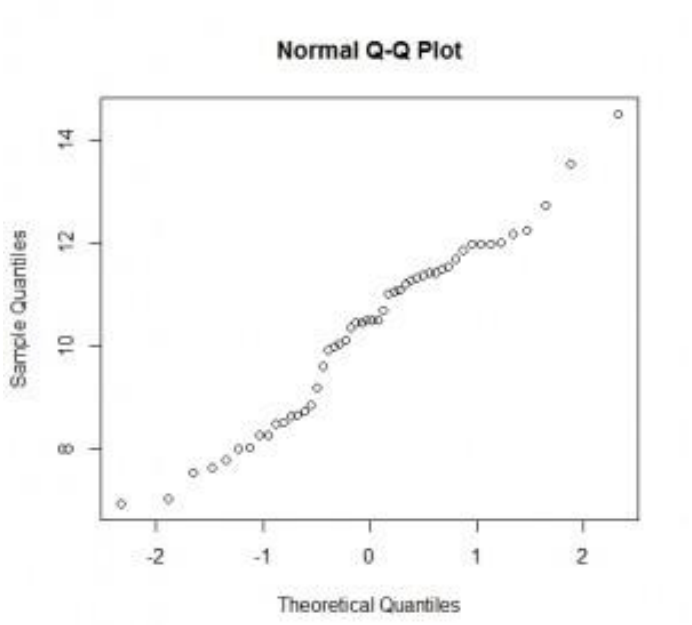
- Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution.
  The number of quantiles is selected to match the size of your sample data.

  The quantile function in R offers 9 different quantile algorithms!

  See help(quantile)

QQ plots for model diagnostics – Q for Quantile

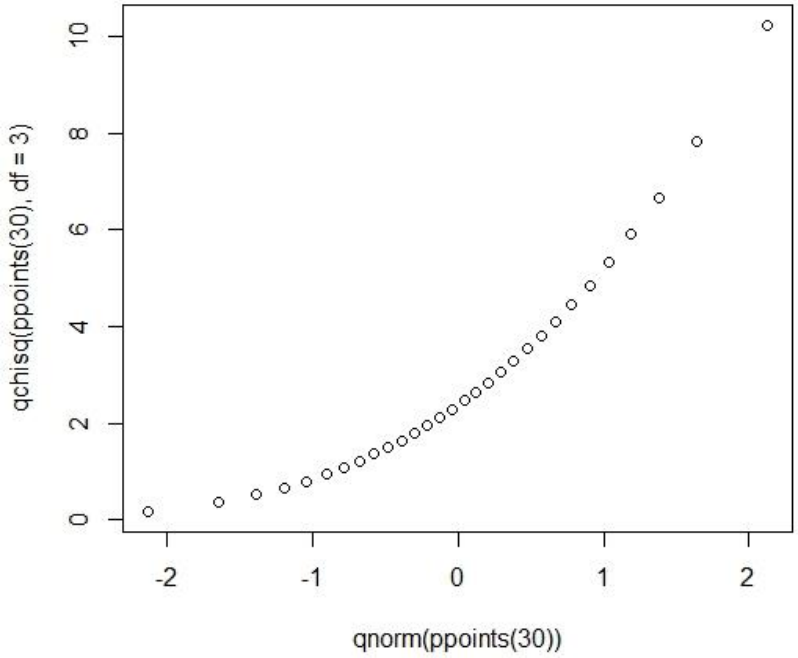
- A Q-Q plot is a scatterplot created by plotting **two sets of quantiles** against one another.
- If both sets of quantiles come from the same distribution, we should see the points forming a line that's roughly straight.
- Here's an example of a Normal Q-Q plot when both sets of quantiles truly come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



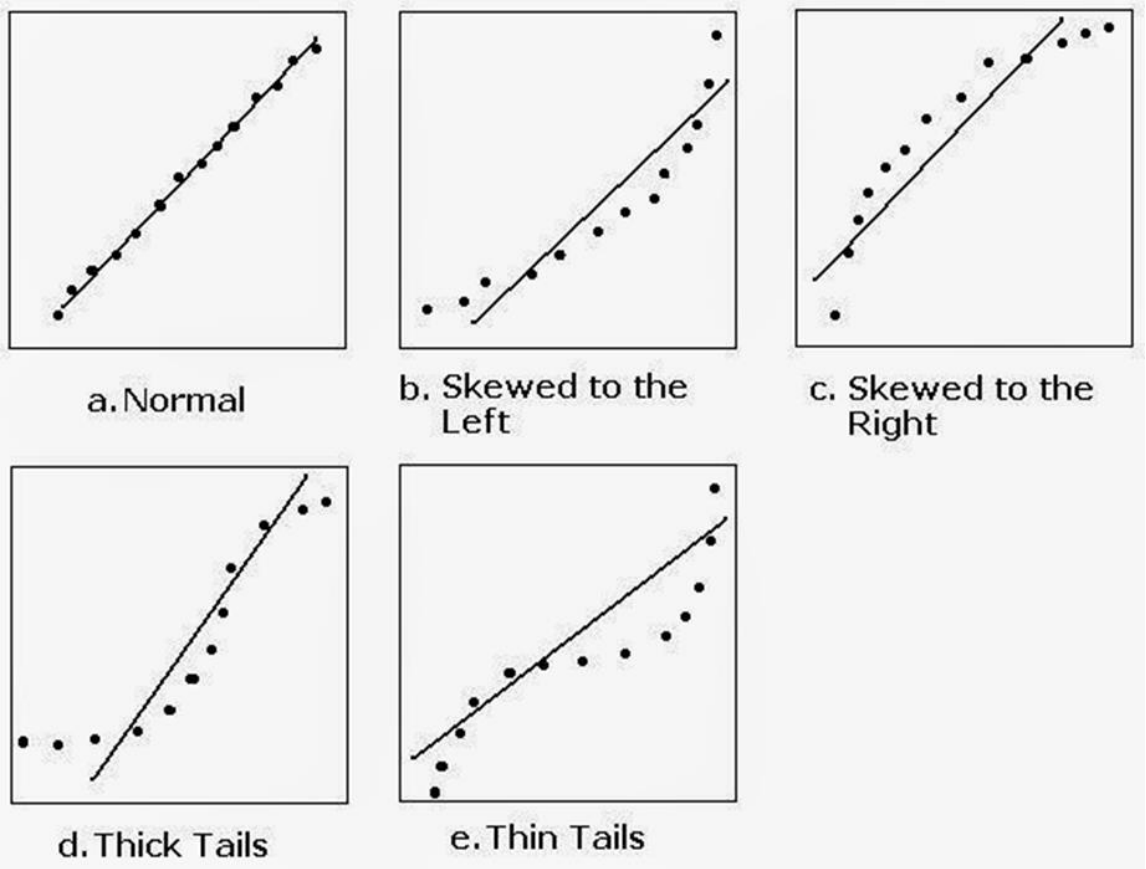
Examples of QQ plots: no straight line

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

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



Examples of QQ plots: some frequent scenarios



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)

Testing for association between case/control status and a SNP

- The **genotype test involves a 2df test** (note that two variables X1 and X2 were needed for genotype coding).
- It has been shown that usually, the additive coding gives adequate power, even when the true underlying mode of inheritance is NOT additive (note that the **additive coding can be achieved by only using 1 variable** (X1)).
- For large sample sizes, a “test for trend” (risk for disease, or average trait increases/decreases with increasing number of “a” copies) theoretically follows a chi-squared distribution with **1df**.

Instead of

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

and modelling

$$E[Y|SNP] = \beta_0 + \beta_1 SNP$$ (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 g() is called a **link function**

and thus

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

For a binary trait Y:

$$E[Y|SNP] = Prob(Y = 1|SNP)$$

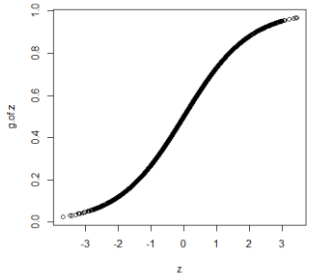
$$= \frac{\exp(\boldsymbol{\eta})}{(1+\exp(\boldsymbol{\eta}))} = \frac{1}{(1+\exp(-\boldsymbol{\eta}))} = \boldsymbol{g\_inv}(\boldsymbol{\eta})$$

where

g_inv is the **logistic function (sigmoid function)**

(squashing the linear predictor to an acceptable range)

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



Since

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

we have

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

and thus

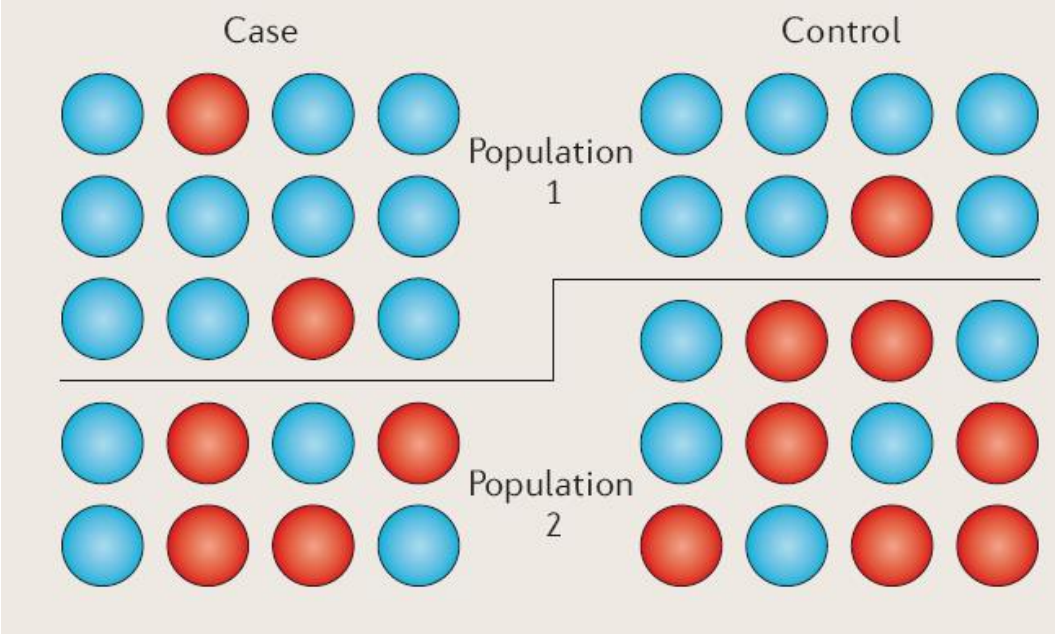
$$\boldsymbol{g}(E[Y|SNP]) = \beta_0 + \beta_1 SNP = \log\left(\frac{Prob(Y = 1|SNP)}{1 - Prob(Y = 1|SNP)}\right) = \boldsymbol{\eta}$$

(g is called **the logit function**)

APPLICATION OF REGRESSION FRAMEWORK

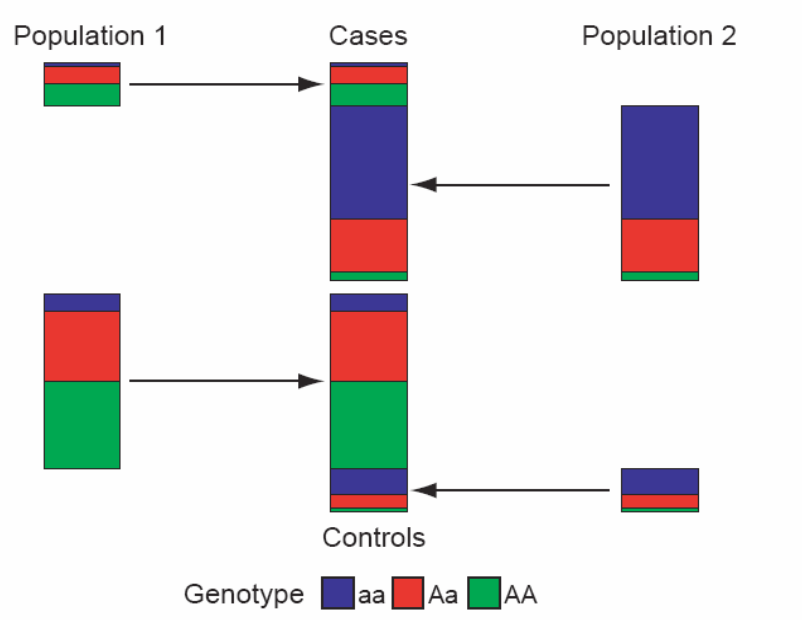
What is spurious association?

- **Spurious association** refers to false positive association results due to not having accounted for population substructure as a confounding factor in the analysis



What is spurious association?

- Typically, there are two characteristics present:
  - A difference in proportion of individual from two (or more) subpopulation in case and controls
  - Subpopulations have different allele frequencies at the locus.



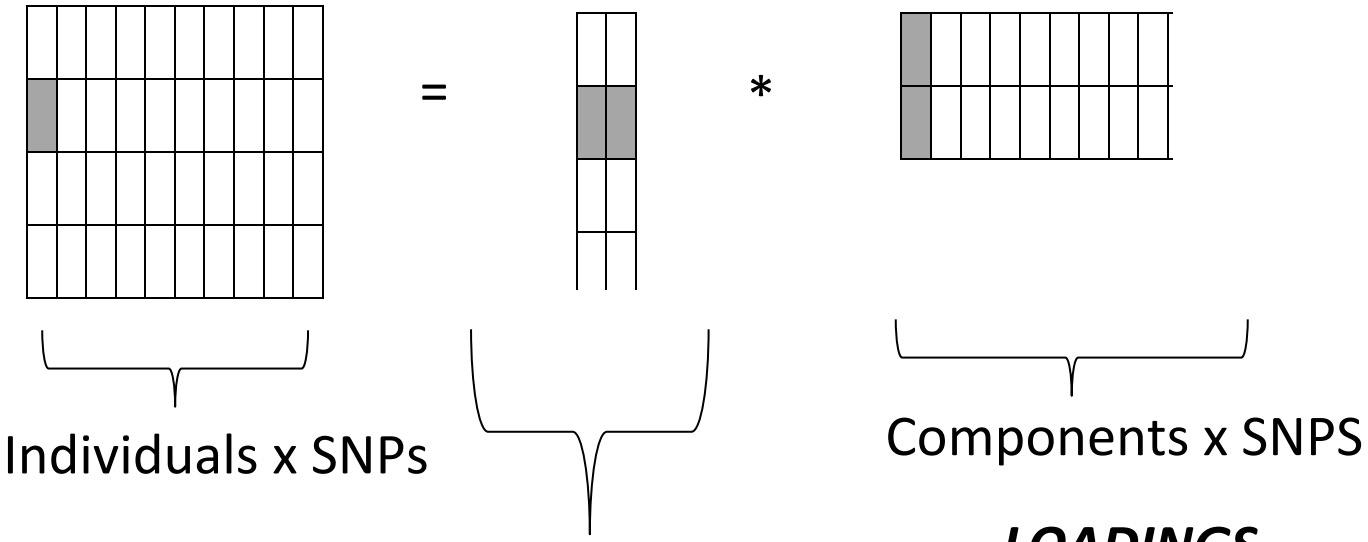
What are typical methods to deal with population stratification?

- Methods to deal with spurious associations generated by population structure generally require a number (at least >100) of widely spaced null SNPs that have been genotyped in cases and controls in addition to the candidate SNPs.

- These methods large group into:

  - **Principal components**
  - Structured association methods: “First look for structure (population clusters) and **second** perform an association **analysis** conditional on the cluster allocation”
  - **Genomic control methods**: “**First analyze** and second downplay association test results for over optimism” → see later

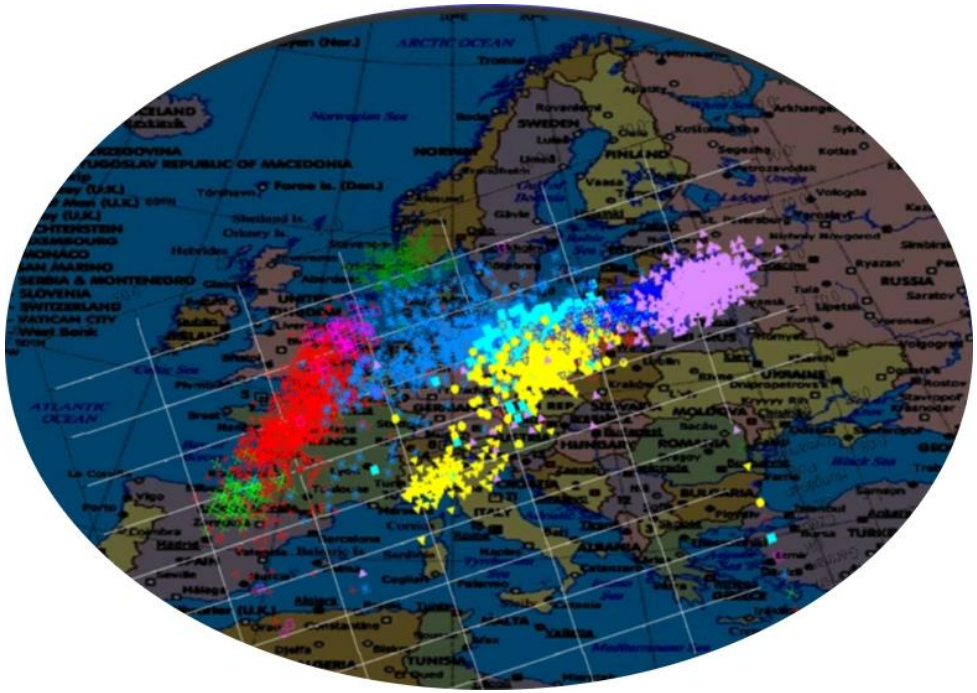
Principal components



- Principal components analysis (PCA) is one of a family of techniques for taking high-dimensional data, and using the dependencies between the variables to represent it in a more tractable, lower-dimensional form, without losing too much information.

Principal components in population genetics

- In European data, the first 2 principal components “nicely” reflect continuous axes of variation due to shared ancestry



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

Principal components in population genetics

- **Example 2:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 PC_1 + \beta_3 PC_2 + \varepsilon$$

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

Genomic control

- In Genomic Control (GC), a 1-df association test statistic is computed at each of the null SNPs, and a parameter λ is calculated as the empirical median divided by its expectation under the chi-squared 1-df distribution.
- Then the association test is applied at the candidate SNPs, and if λ > 1 the test statistics are divided by λ.

  - Under H_0 of no association p-values uniformly distributed
  - In case of population stratification: inflation of test statistics
  - $\hat{\lambda} = \frac{\text{median}(\chi_1^2, \chi_2^2, \ldots, \chi_L^2)}{\text{median}(\mathcal{L}(\chi_1^2))} = \frac{\text{median}(\chi_1^2, \chi_2^2, \ldots, \chi_L^2)}{0.456}$
  - $\chi_{GC}^2 = \chi^2 / \hat{\lambda}$

```
> median(rchisq(10,1))
[1] 0.9641272
> median(rchisq(100,1))
[1] 0.5001173
> median(rchisq(1000,1))
[1] 0.4206546
> median(rchisq(10000,1))
[1] 0.4686072
> median(rchisq(100000,1))
[1] 0.455271
> median(rchisq(1000000,1))
[1] 0.4548966
```

A note about MULTIPLE TESTING – Bonferroni

- Recall we are testing ~1 Million markers, more or less
- Several strategies to adjust the p-values for doing so many tests
  - Bonferroni
    - 0.05/{# tests, i.e., # markers, M}
    - most widely used in practice
    - Pr(Reject any test | null hypothesis true) = 0.05
  - False Discovery Rate (FDR)
  - Permutation

A note about multiple testing – FDR

- False Discovery Rate (FDR) limits the expected number of false positives
- Less stringent control than Bonferroni, e.g.
- *"Another way to look at the difference is that a p-value of 0.05 implies that 5% of all tests will result in false positives. An FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives. The latter is clearly a far smaller quantity."*

  (http://www.nonlinear.com/support/progenesis/samespots/faq/pq-values.aspx)

Table 10.1 Number of errors committed when testing M null hypotheses

	Declared non-significant	Declared significant	Total Total
true null hypotheses	U	V	M_0
non-true null hypotheses	T	S	$M - M_0$
	$M - R$	R	M

Then the false discovery rate is given by

$$E\left(\frac{V}{V+S}\right).$$

A note about multiple testing – permutations

- Many of the tested genotype markers are correlated with each other (in LD), and so the tests are correlated
- Bonferroni adjusts as if they were completely independent
- Permutation will be more powerful, but… ?

A final note about multiple testing

- Nan Laird comments: *"Given the many false positive findings in the history of genetic association studies, one rather errs on being too conservative."*
- Initial GWAS had a lot of false positives (recall, replication, replication, replication...)

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: randomized controlled trials

- Randomized trials are studies in which human volunteers are randomly assigned to receive either the agent being studied or an inactive placebo, usually under double-blind conditions (where neither the participants nor the investigators know which substance each individual is receiving), and their health is then monitored for a period of time.
- This type of study can provide strong evidence for a causal effect, especially if its findings are replicated by other studies.

(https://www.acsh.org)

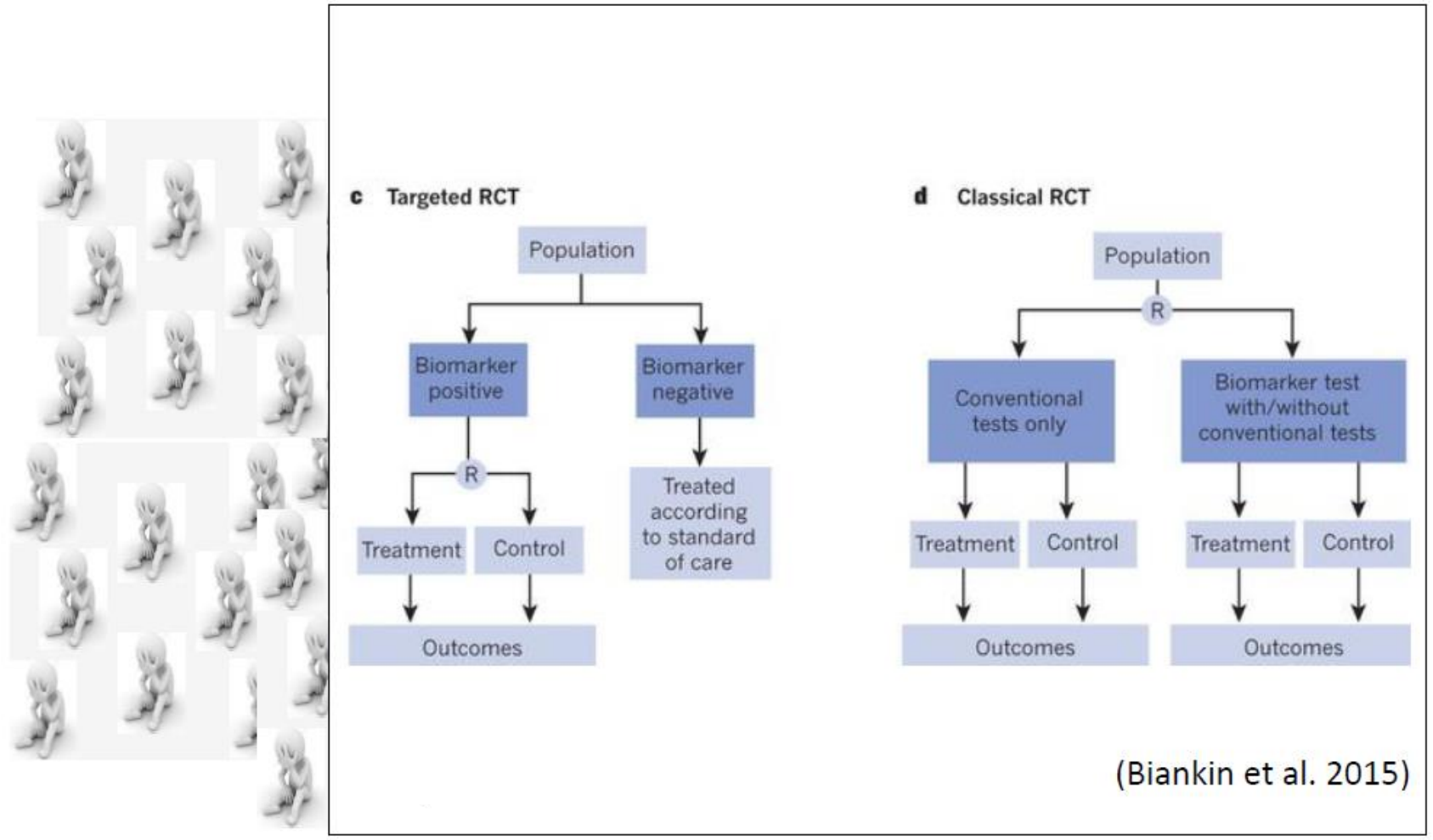
Philos Stud (2010) 147:59–70
DOI 10.1007/s11098-009-9450-2

What are randomised controlled trials good for?

Nancy Cartwright

Abstract Randomized controlled trials (RCTs) are widely taken as the gold standard for establishing causal conclusions. Ideally conducted they ensure that the treatment 'causes' the outcome—in the experiment. But where else? This is the venerable question of external validity. I point out that the question comes in two importantly different forms: Is the specific causal conclusion warranted by the experiment true in a target situation? What will be the result of implementing the treatment there? This paper explains how the probabilistic theory of causality implies that RCTs can establish causal conclusions and thereby provides an account of what exactly that causal conclusion is. Clarifying the exact form of the conclusion shows just what is necessary for it to hold in a new setting and also how much more is needed to see what the actual outcome would be there were the treatment implemented.

Designing RCTs for testing precision-medicine strategies is an evolving field!



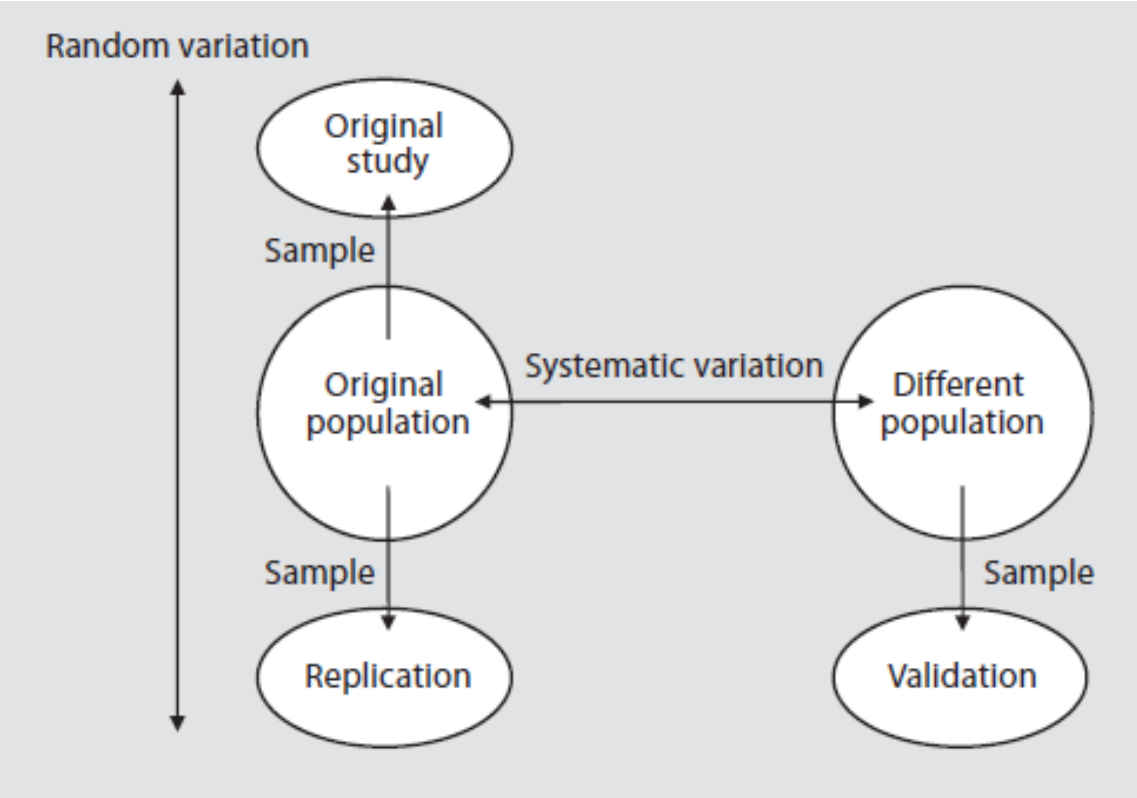
Establishing causation: causal variants for human complex traits

- 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.
  - Also here, corrections for population structure are needed to remove potential spurious associations.

6 Post-analysis Steps

6.a Replication and Validation

The difference



(Igl et al. 2009)

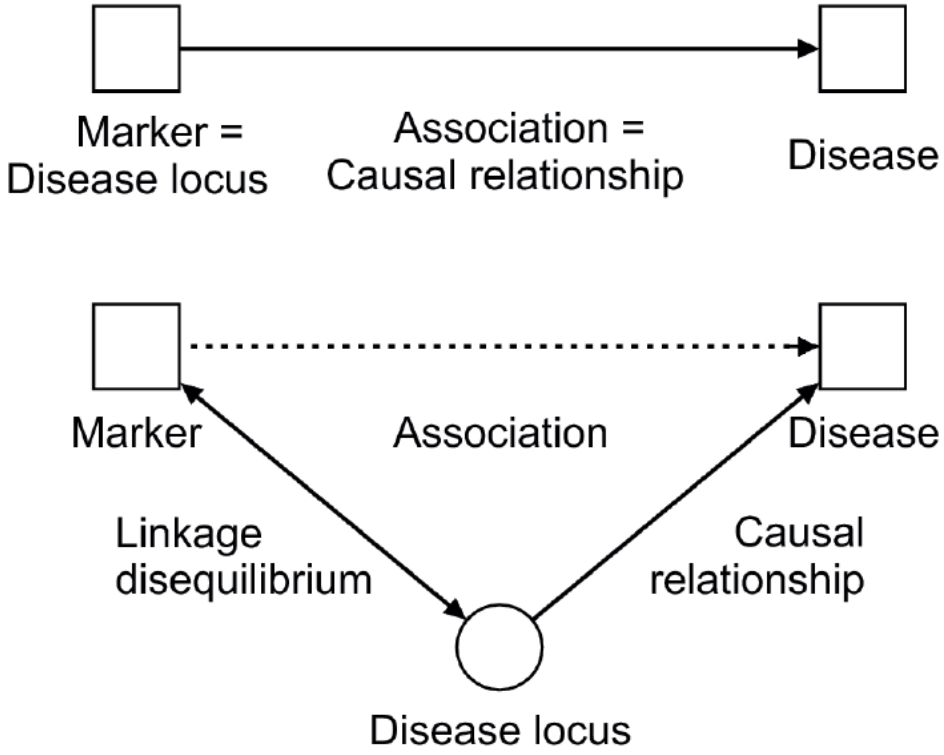
Guidelines for replication studies

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

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

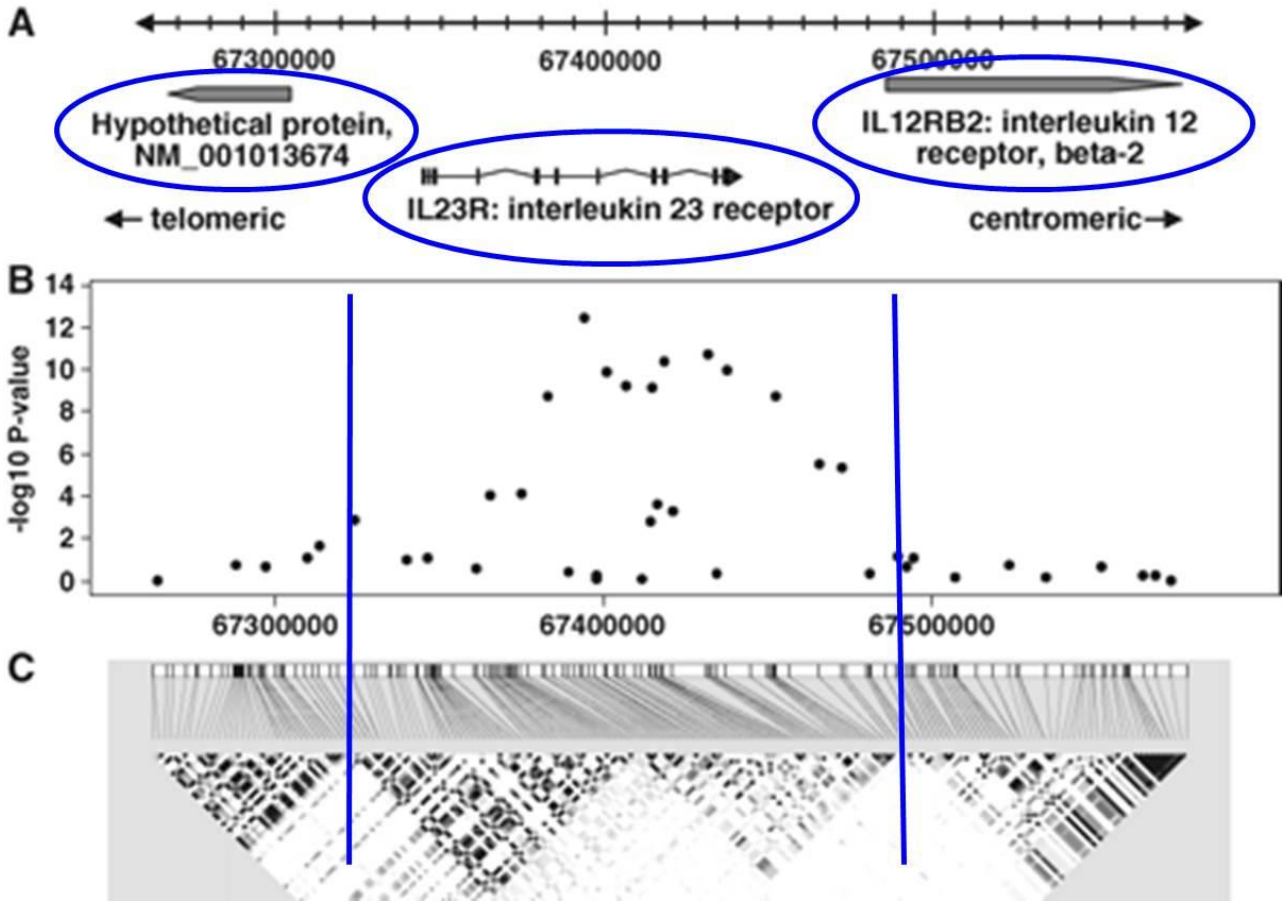
6.b GWA Interpretation and follow-up

Entering the field of functional genomics



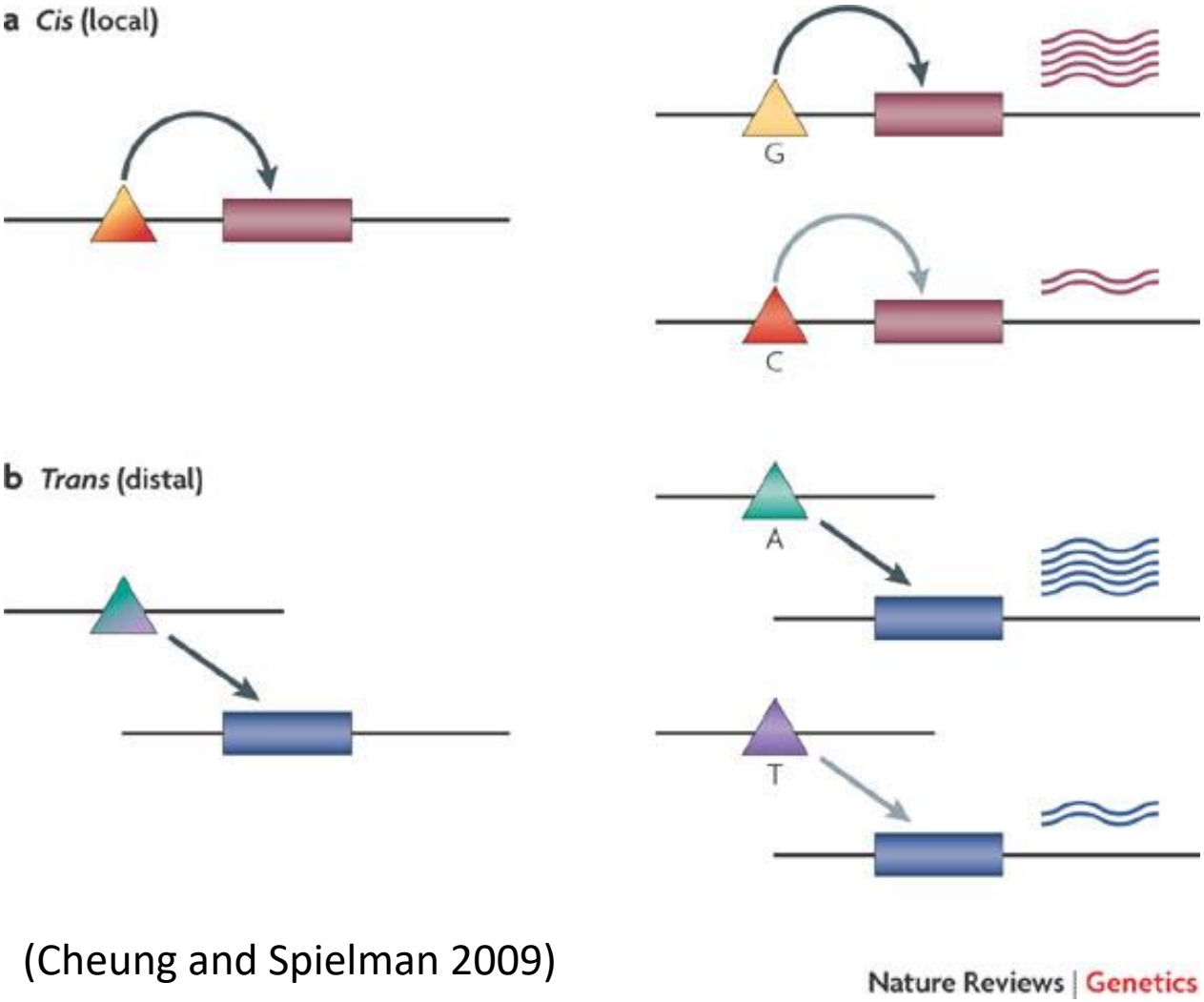
(Ziegler and Van Steen, Brazil 2010)

Finding the “relevant” loci



(Duerr et al 2006)

Finding the “relevant” loci



(Cheung and Spielman 2009)

- Cis-acting variants are found close to the target genes and trans-acting variants are located far from the target genes, often on another chromosome.
- Different allelic forms of the cis- and trans-acting variants have different influence on gene expression.

Finding the “relevant” loci: finding “genes” and “eQTLs” in whole blood

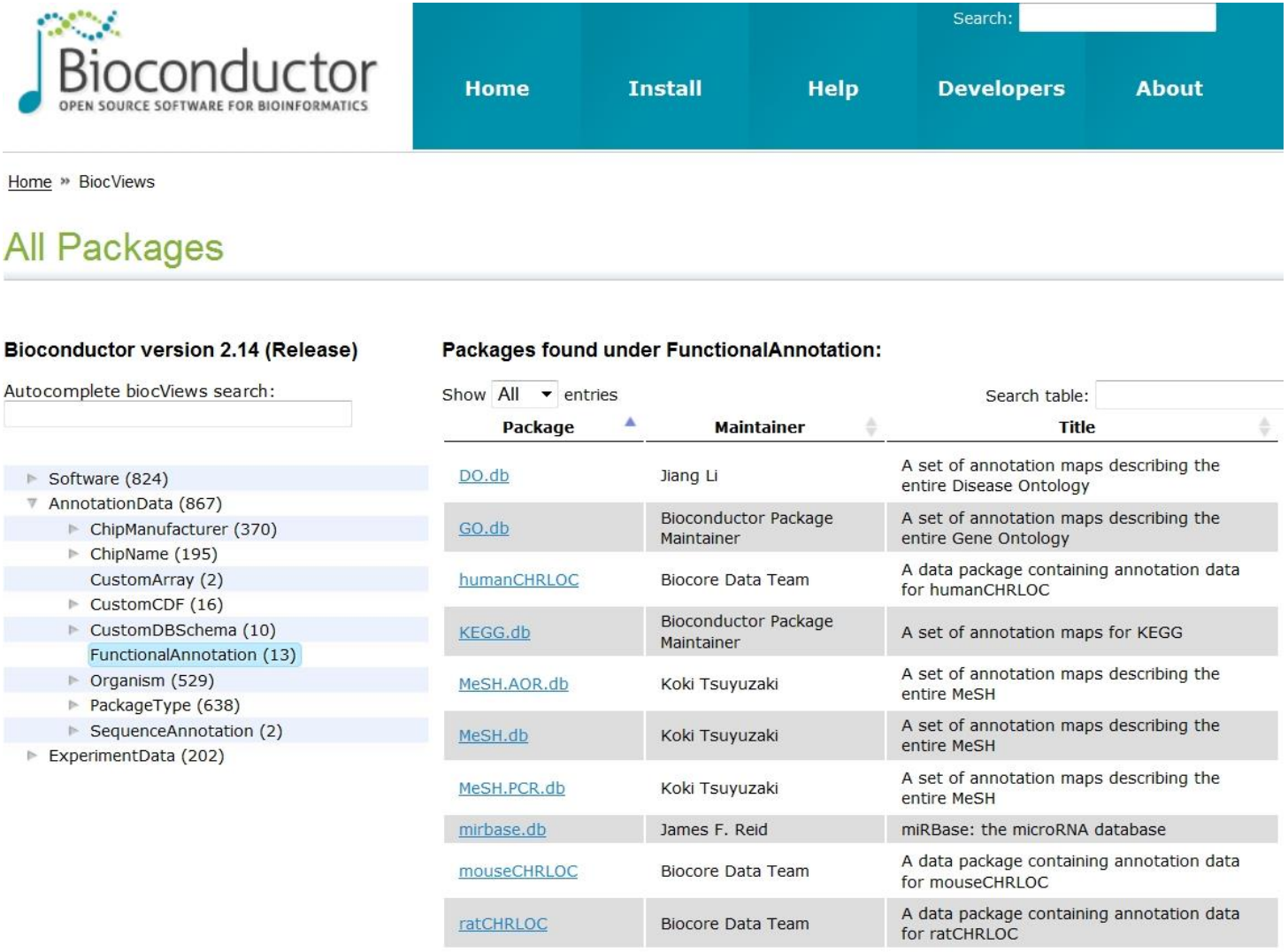
(https://data.broadinstitute.org/mpg/depict/)

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

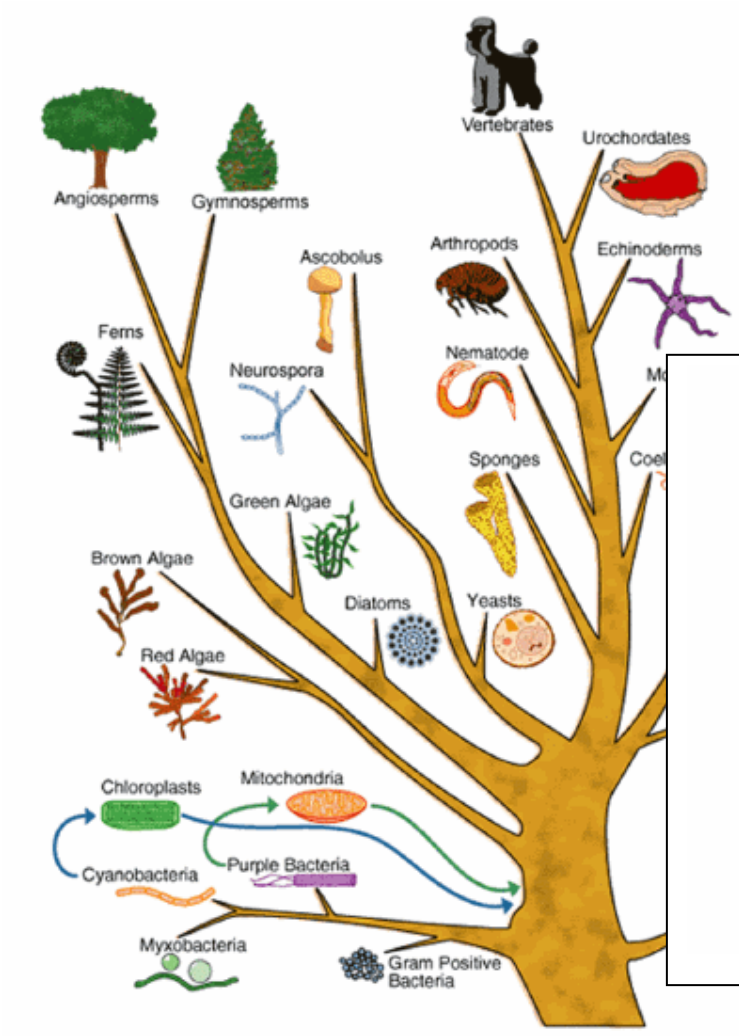
- Evaluating the functional properties of gene sets is commonly used both to verify that the genes implicated in a biological experiment are **functionally relevant** and to **discover unexpected shared functions** between those genes.
- DEPICT can test the null hypotheses:
  - "Genes in associated loci do not enrich for the reconstituted gene set"
  - "Genes in associated are not highly expressed in the given tissue or cell type"
- Many **functional annotation databases** have been developed in order to classify genes according their various roles in the cell (e.g., Gene Ontology (GO) repository, widely consulted by many functional enrichment tools).

(Glass and Girvan 2014)

Functional annotation in R

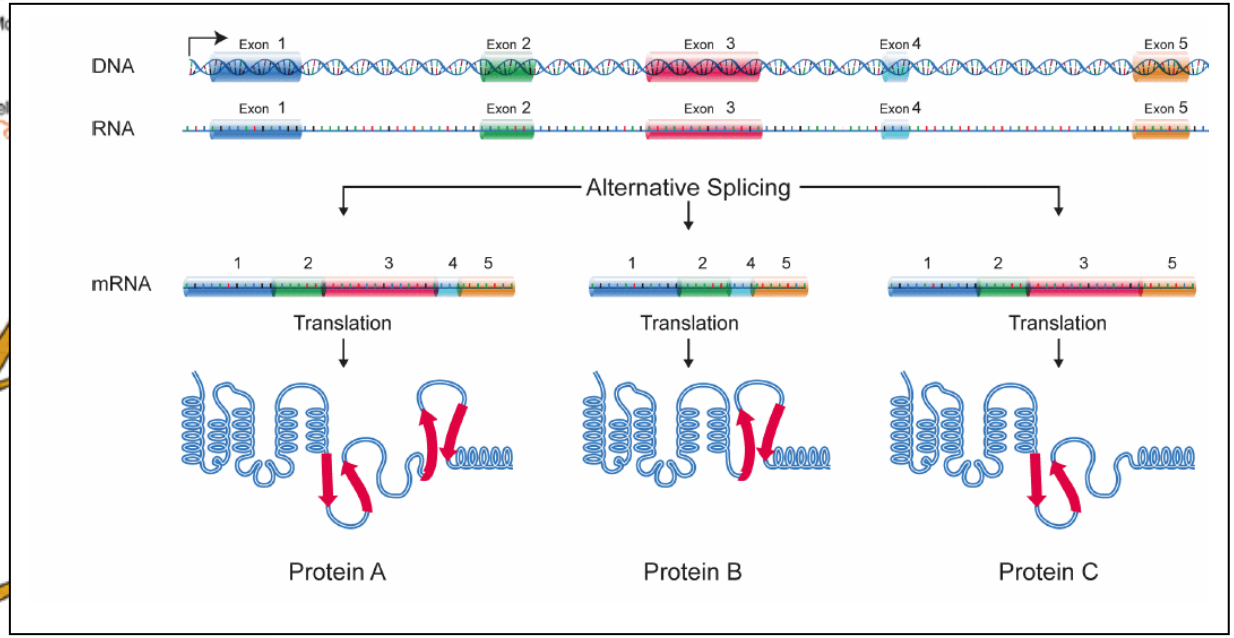


Gene expression as an extra source of information



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:**



More about sequence analyses in the subsequent classes

Model organisms as an extra source of information

- Suppose that we have an unknown human DNA sequence that is associated with the disease cystic fibrosis.
- A bioinformatic analysis finds a similar sequence from mouse that is associated with a gene that codes for a membrane protein that regulates salt balance.
- A good bet may be that the human sequence also is part of a gene that codes for a membrane protein that regulates salt balance.

More about sequence analyses in the subsequent classes

"The more we find, the more we see, the more we come to learn.

The more that we explore, the more we shall return."

Sir Tim Rice, *Aida*, 2000

Main supporting doc to this class (complementing course slides)

√

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Education

Chapter 11: Genome-Wide Association Studies

William S. Bush[1]*, Jason H. Moore[2]

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

Questions?

Homework assignment

(literature based or Q&A reports)

Check out the document

“Critical evaluation of a paper/report”