B I O I N F O R M A T I C S

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 3: GENOME-WIDE ASSOCIATION STUDIES

1 Setting the pace

1.a "The Human Genome Project"

1.b The concept of a genetic marker

1.c The rise of Genome-Wide Association studies (GWAs)

2 Components of a GWAs

3 Study Design

3.a Marker level

3.b Subject level

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

4 Pre-analysis Steps

3.a Quality-Control: The Travemünde criteria

Hardy-Weinberg equilibrium

3.b Linkage disequilibrium

3.c Confounding: population stratification

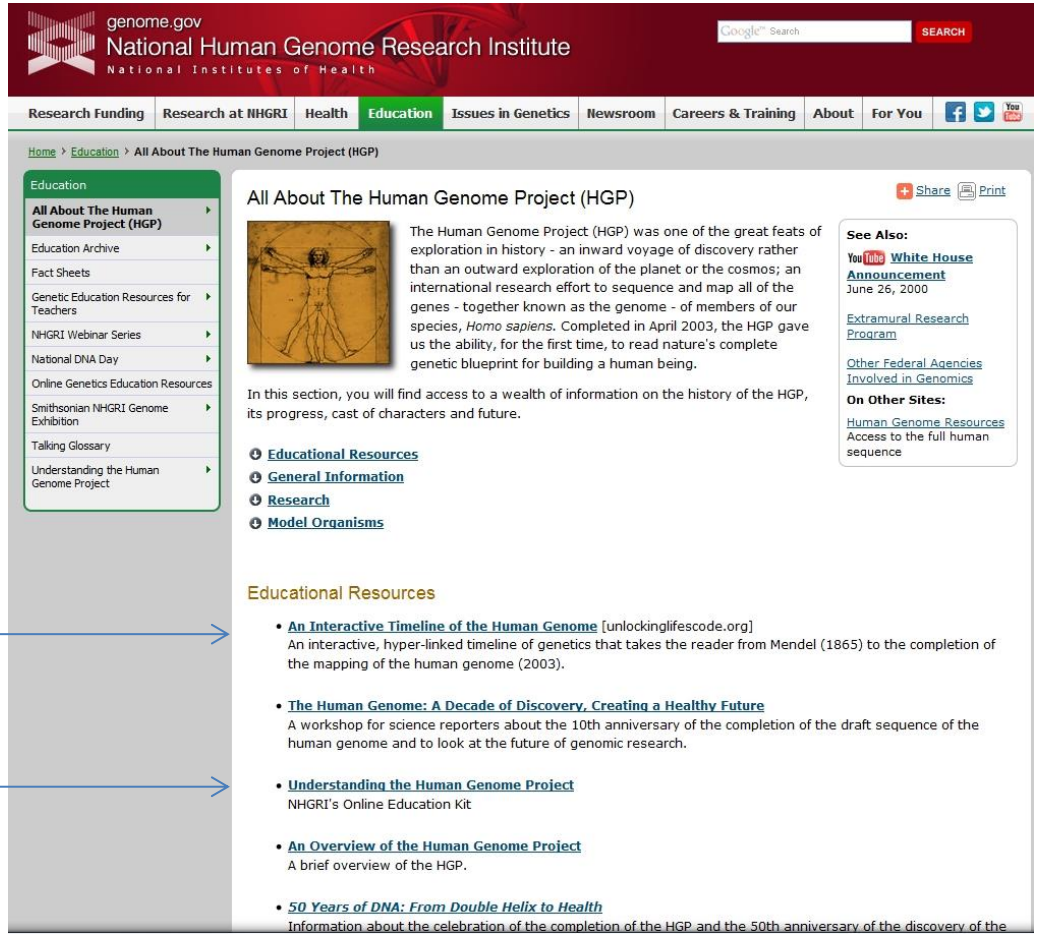
5 Testing for Associations

6 Replication and Validation

7 GWA Interpretation and Follow-Up

1 Setting the pace

1.a “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

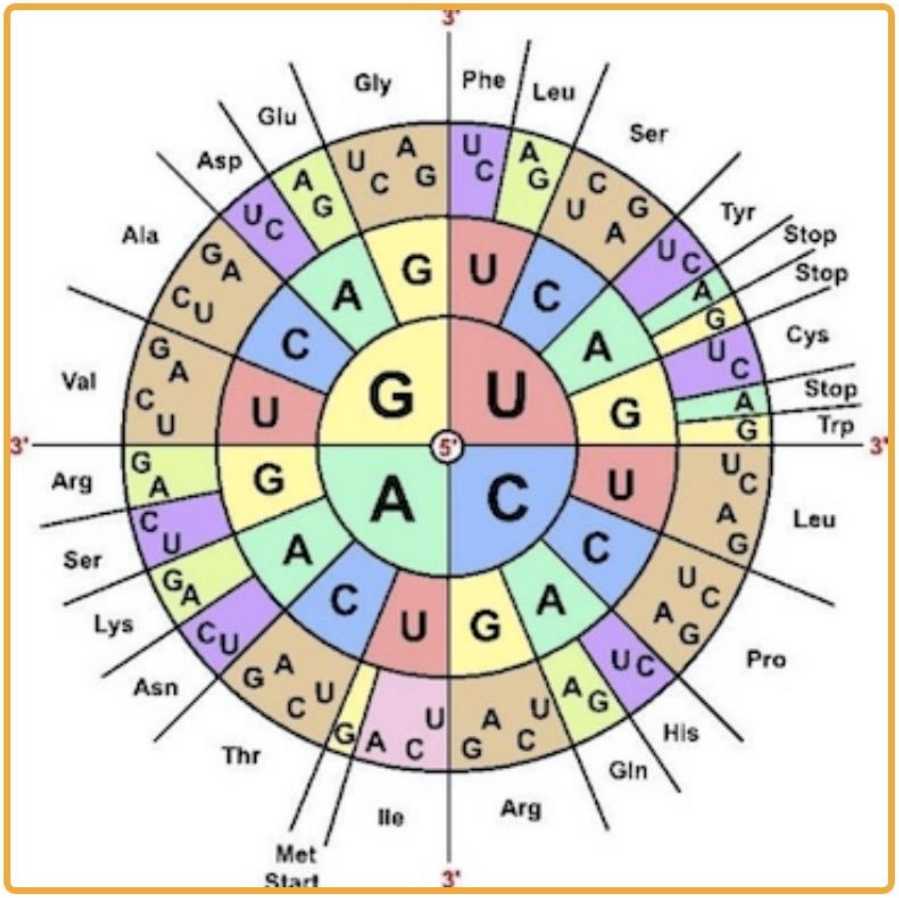
Historical overview

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

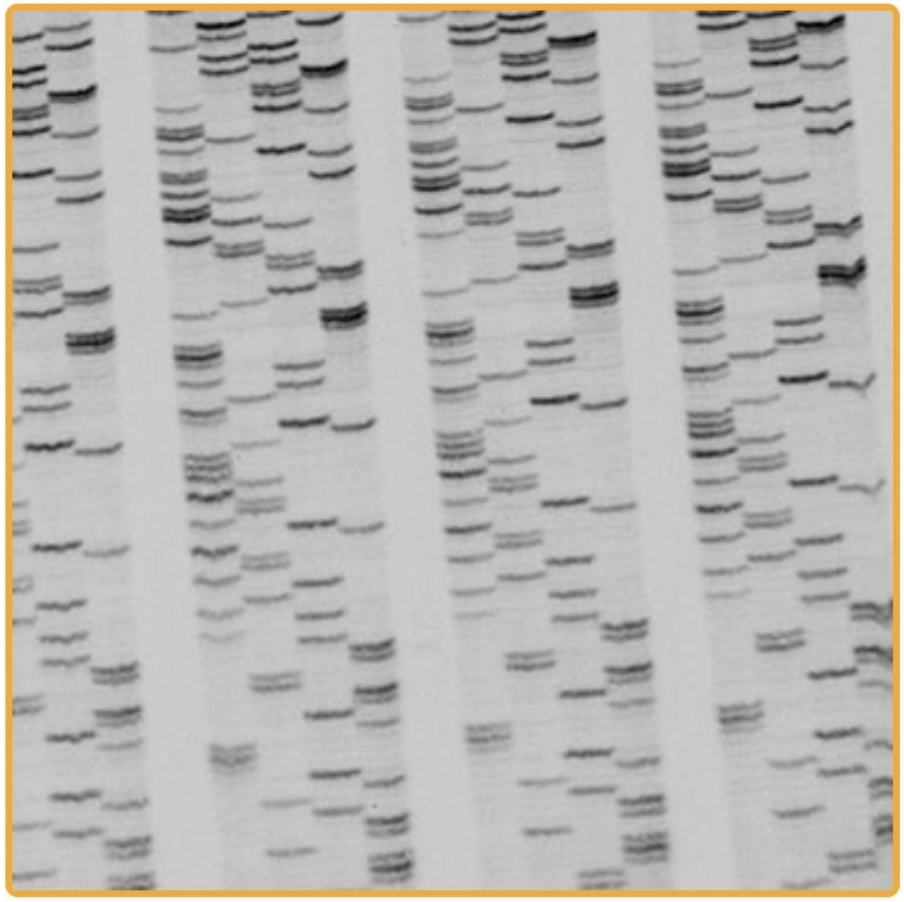


Marshall Nirenberg cracks the genetic code for protein synthesis

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

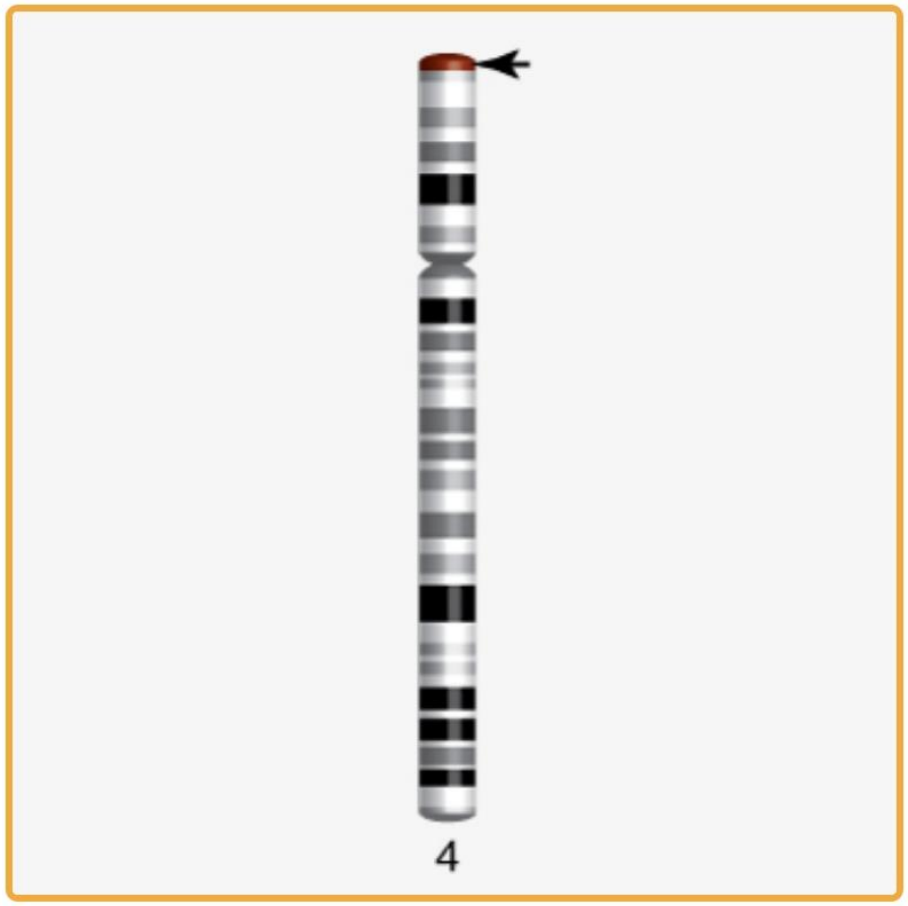


Frederick Sanger develops rapid DNA sequencing technique

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

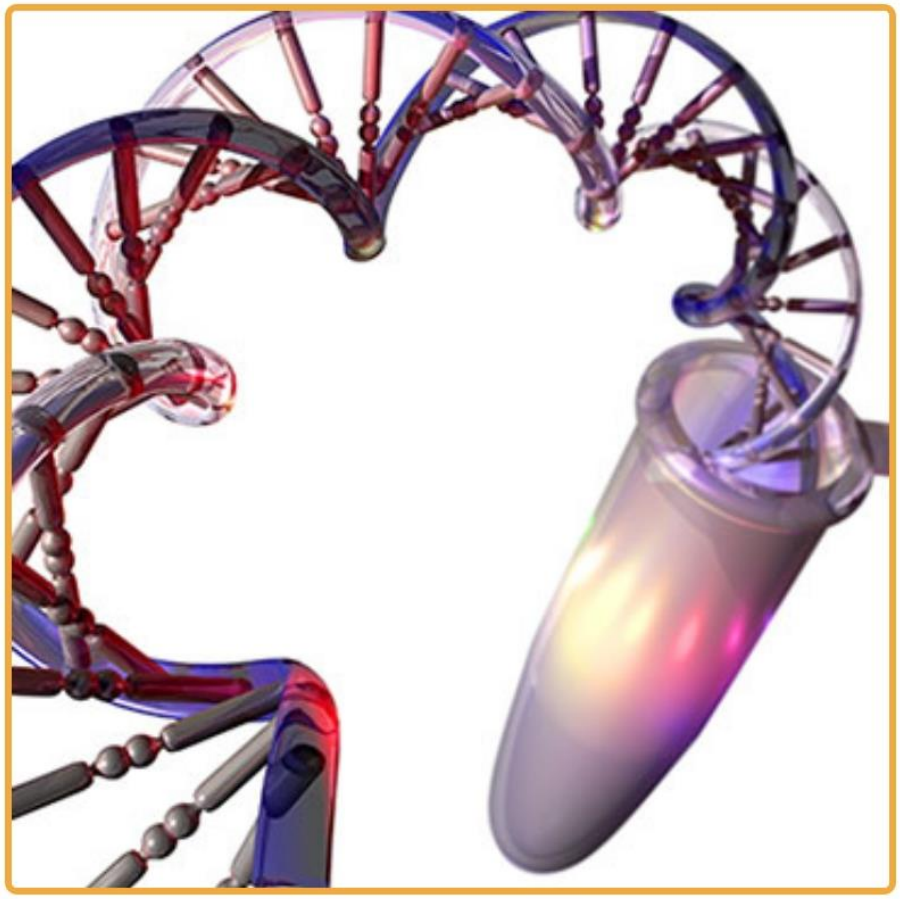


First genetic disease mapped, Huntington's Disease

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

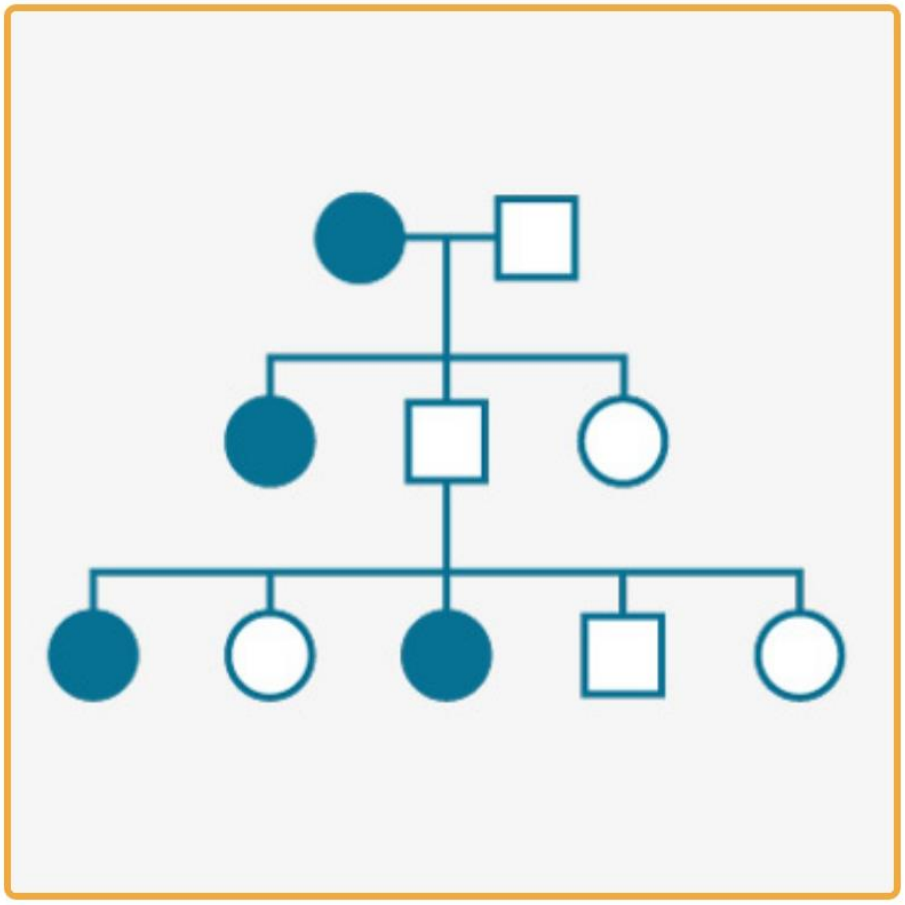


Invention of polymerase chain reaction (PCR) technology for amplifying DNA

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] Keriallen 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 Wan,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

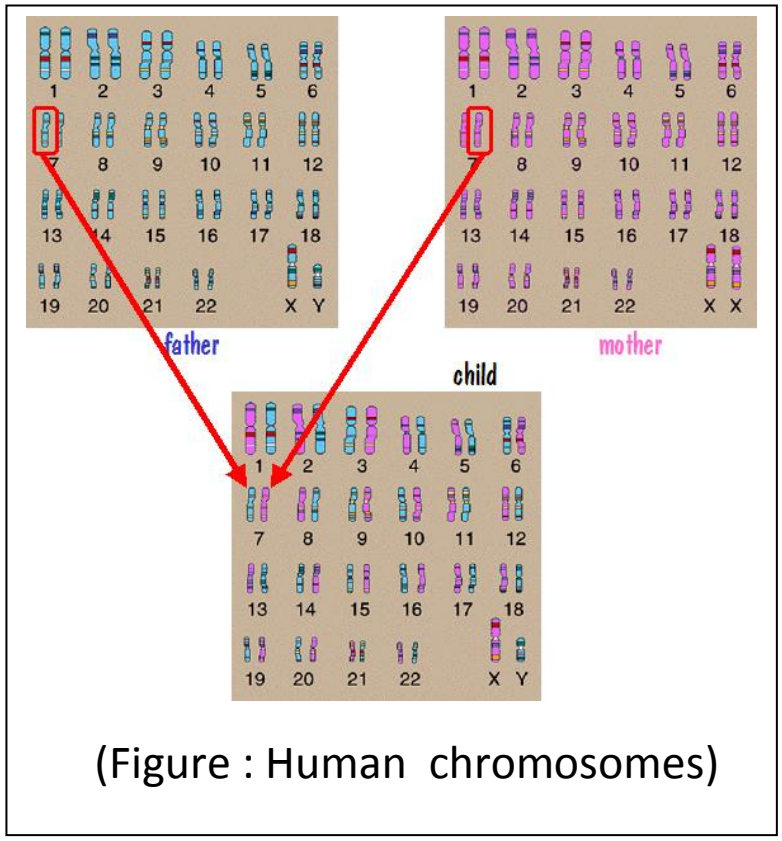
16 FEBRUARY 2001 VOL 291 SCIENCE www.sciencemag.org

- 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**;
  - 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.

Terminology: 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)

Terminology: 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”

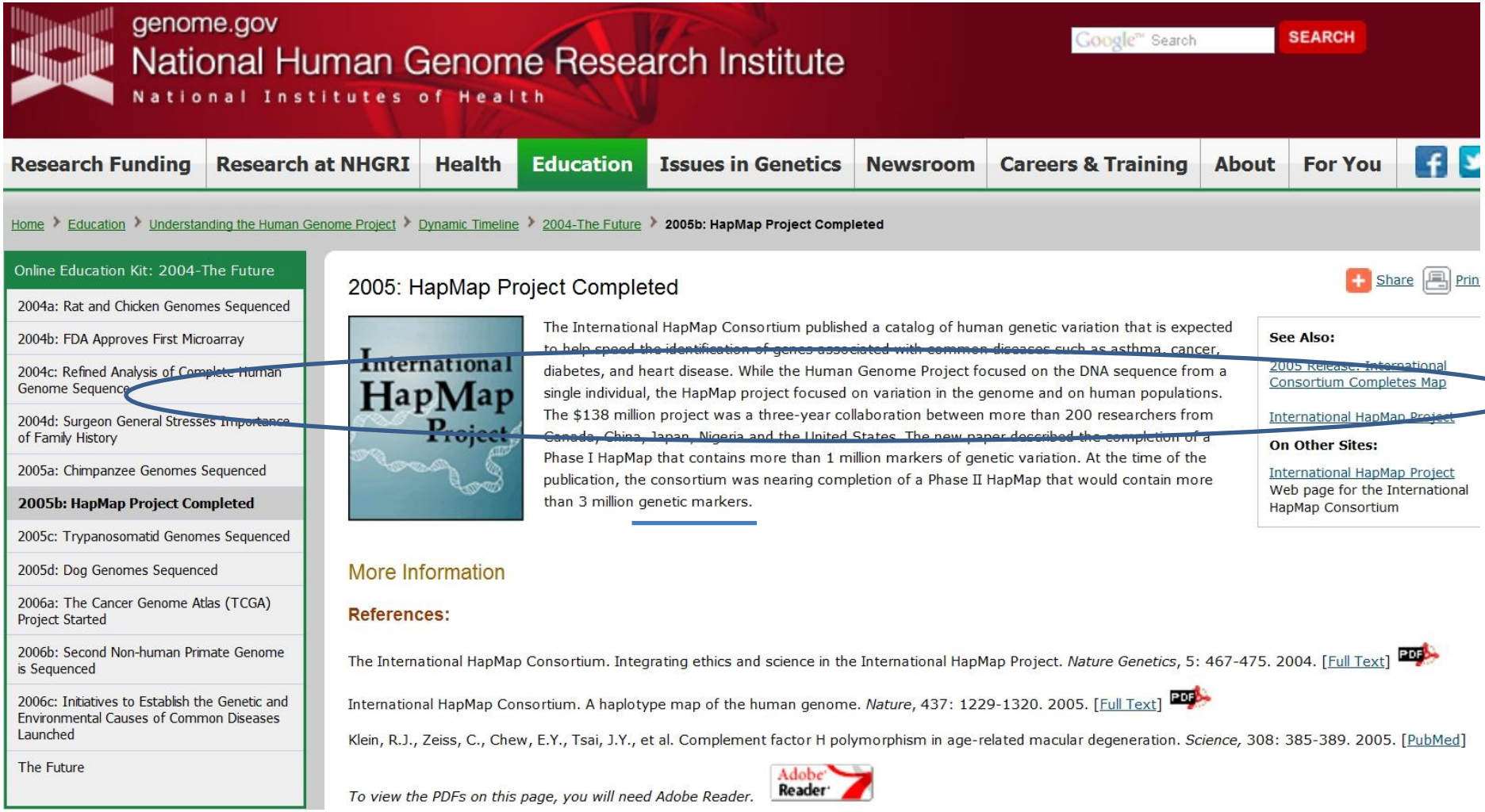
Terminology: Alleles

- 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

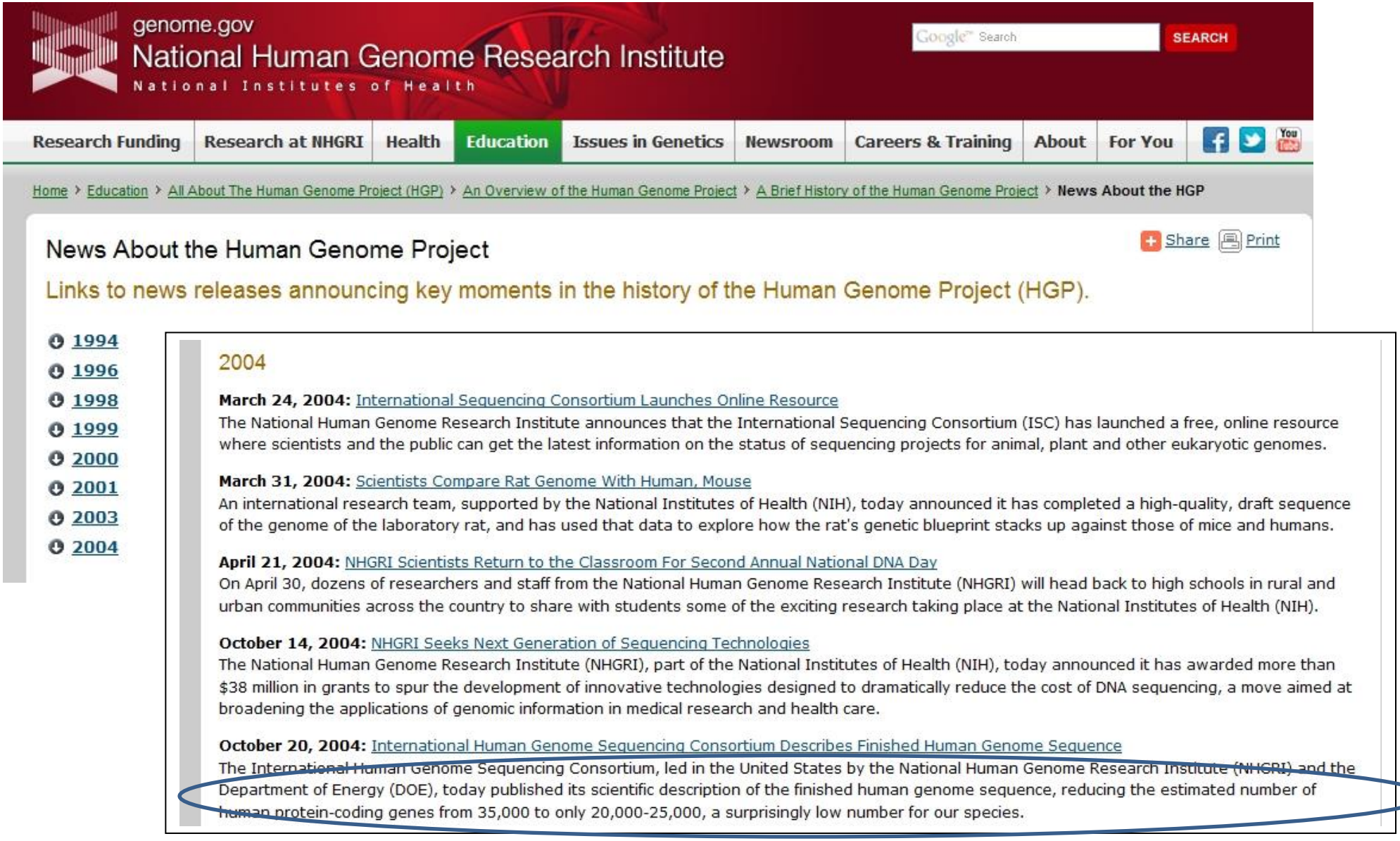
- Genetic marker: polymorphic DNA sequence at single locus

[Mutations ~polymorphisms (see later)]

Historical overview



News about the HGP



Historical overview

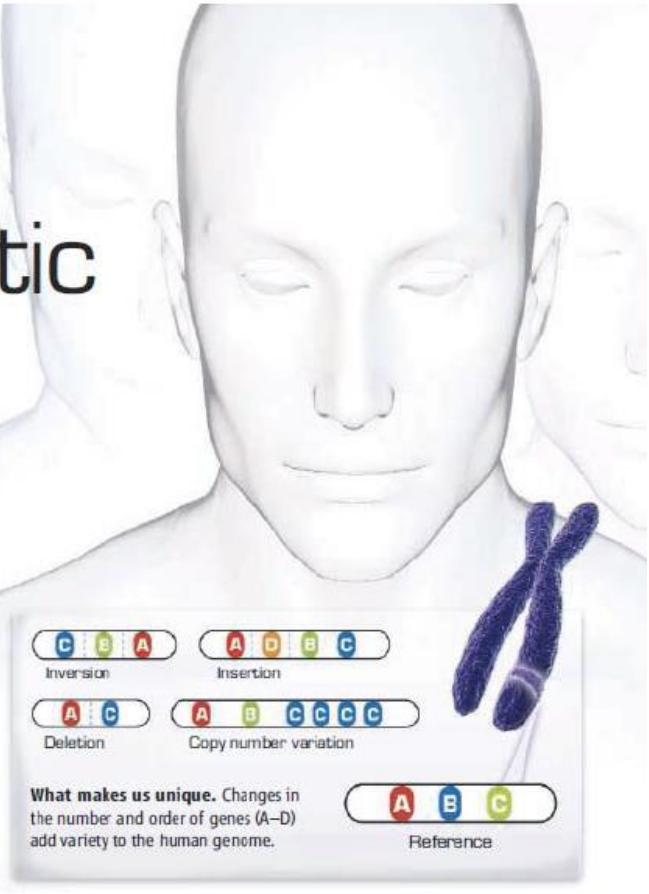
BREAKTHROUGH OF THE YEAR

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

Historical overview: associating genetic variation to disease outcomes

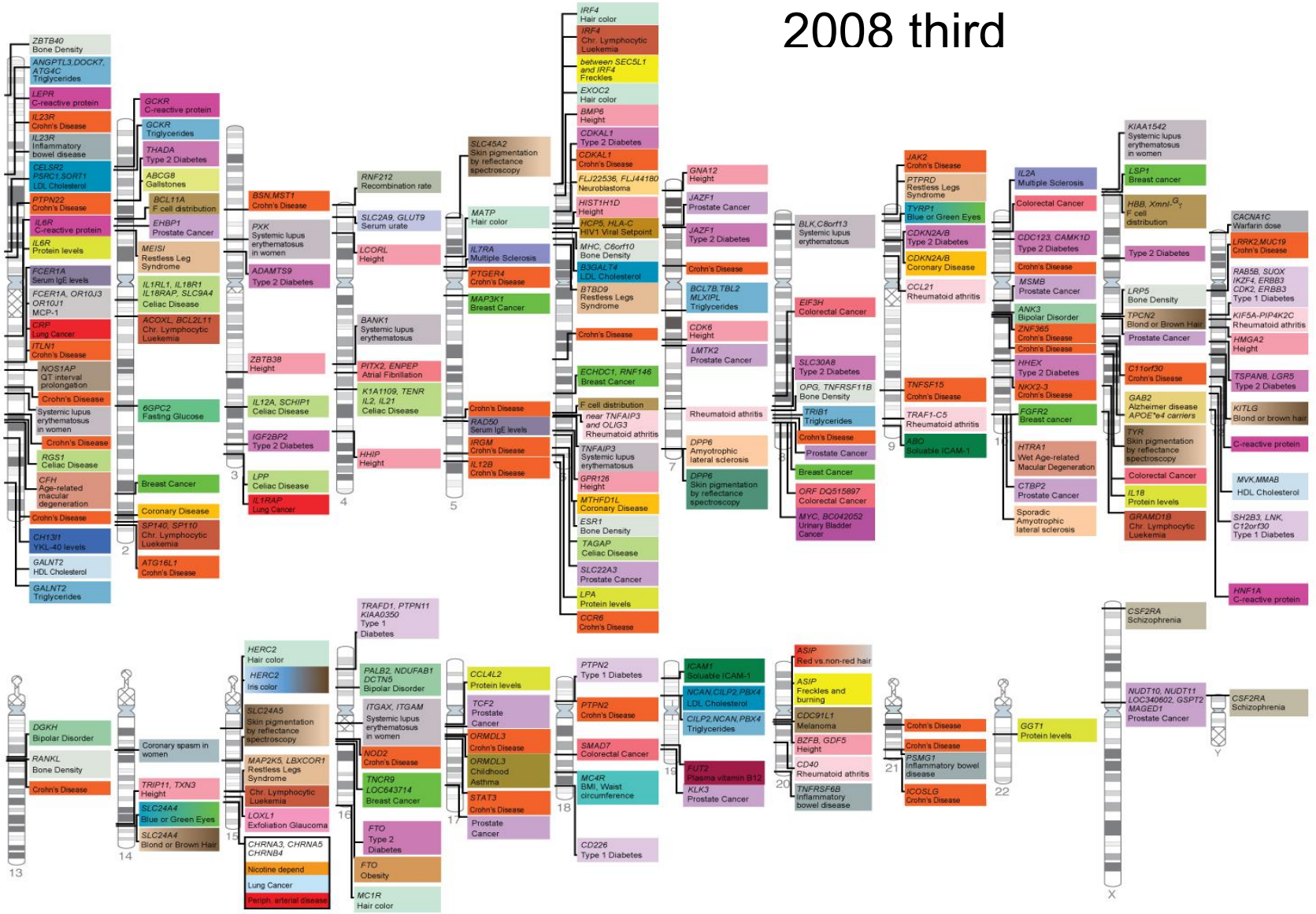
BREAKTHROUGH OF THE YEAR: The Runners-Up

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

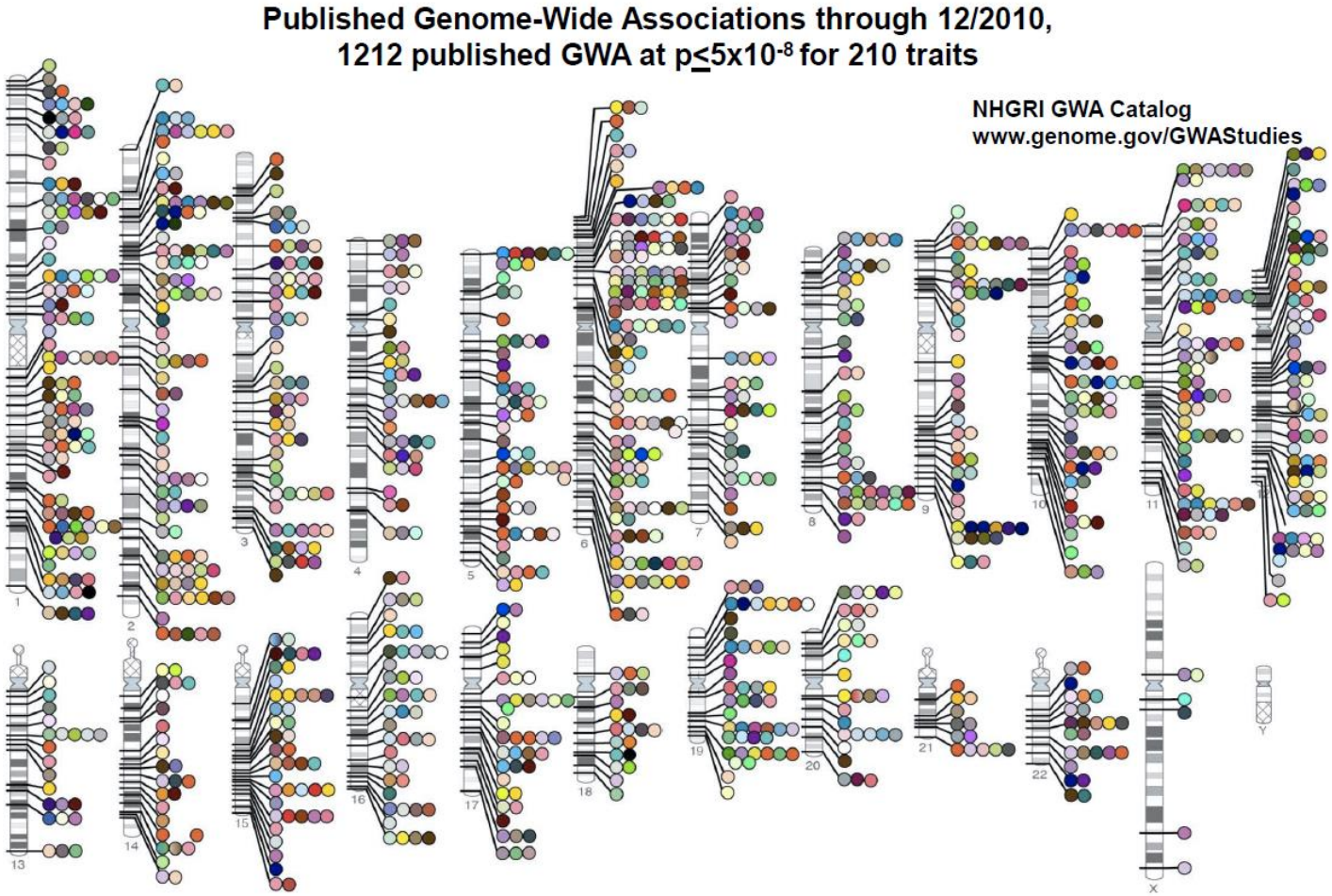
Areas to Watch in 2007

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?

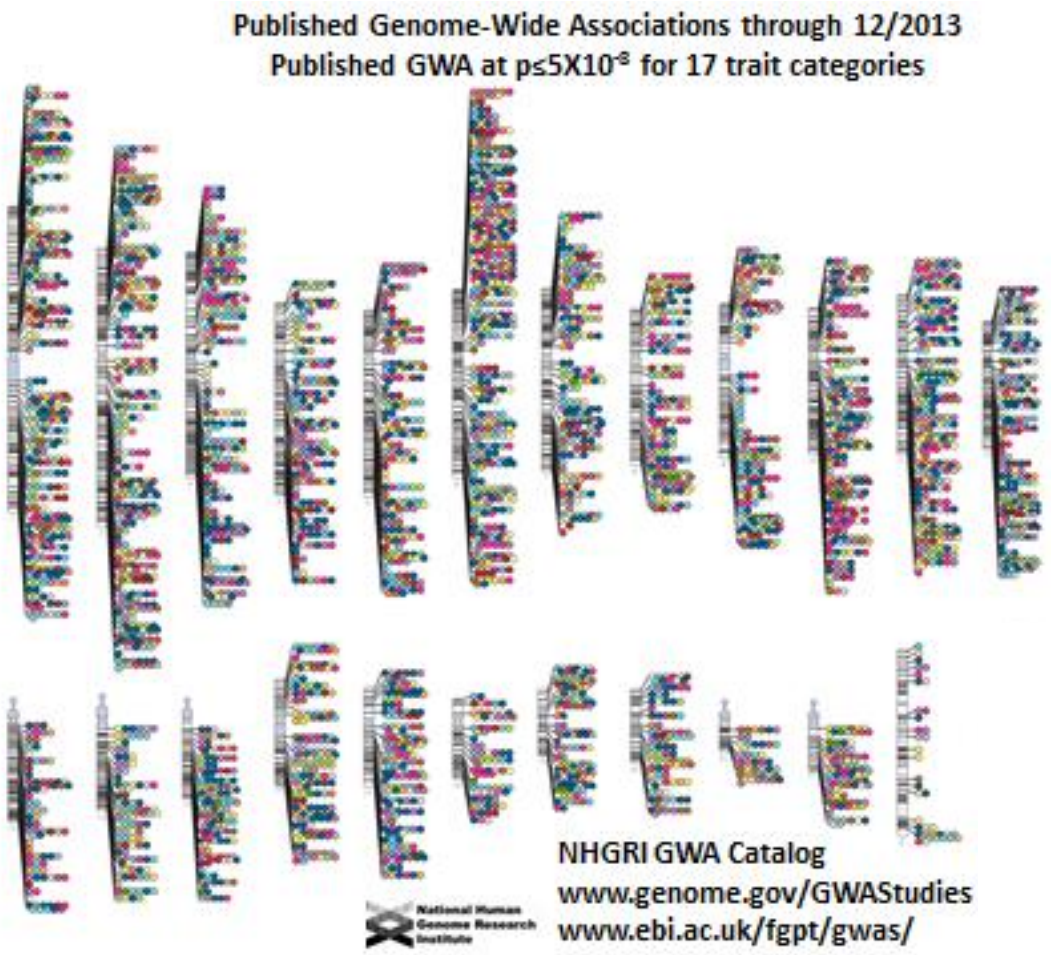
Historical overview: GWAs as a tool to “map” diseases



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

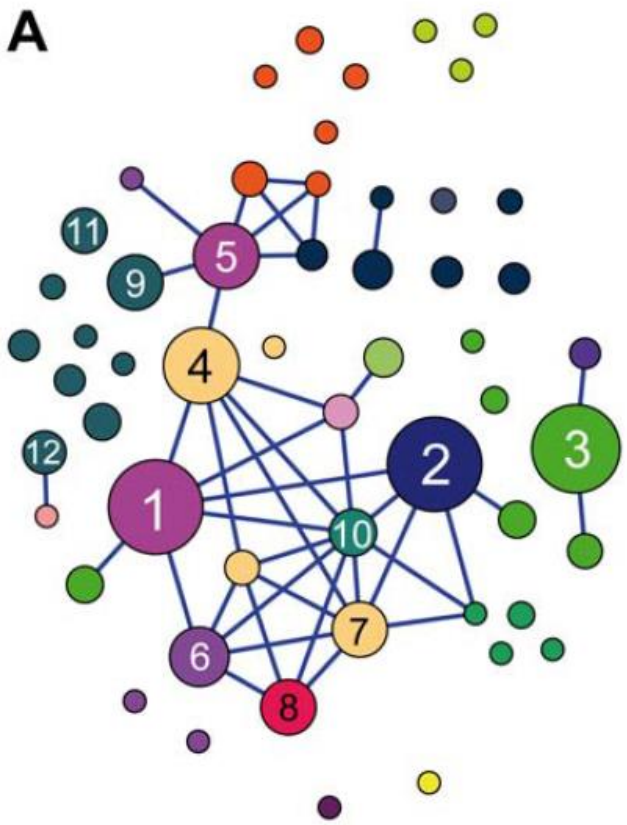


Historical overview: trait categories



- 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)

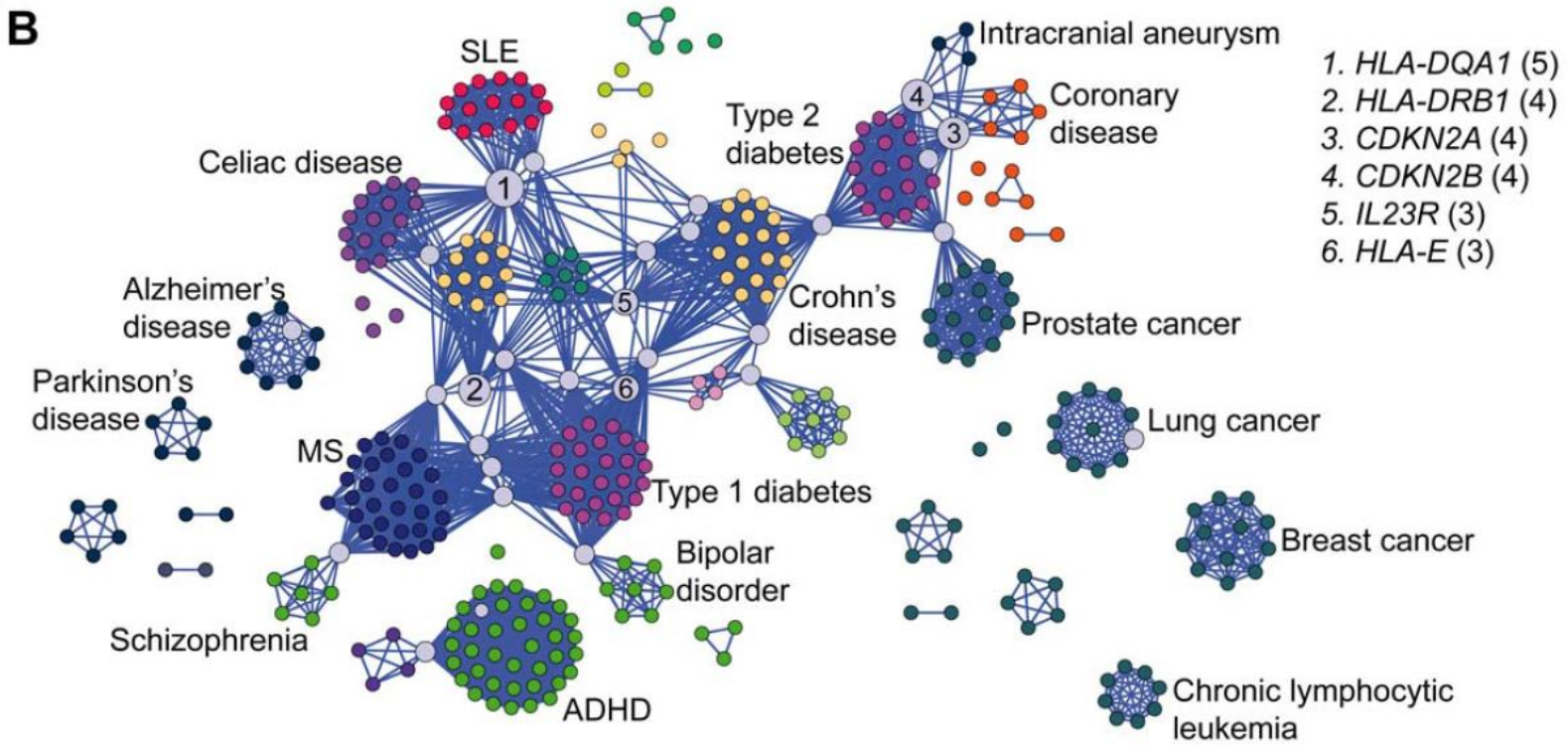


1. Type 1 diabetes (36)
2. Multiple sclerosis (36)
3. ADHD and conduct disorder (33)
4. Crohn's disease (27)
5. Type 2 diabetes (22)
6. Celiac disease (19)
7. Ulcerative colitis(17)
8. Systemic lupus erythematosus (17)
9. Prostate cancer (17)
10. Rheumatoid arthritis (13)
11. Breast cancer (12)
12. Lung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

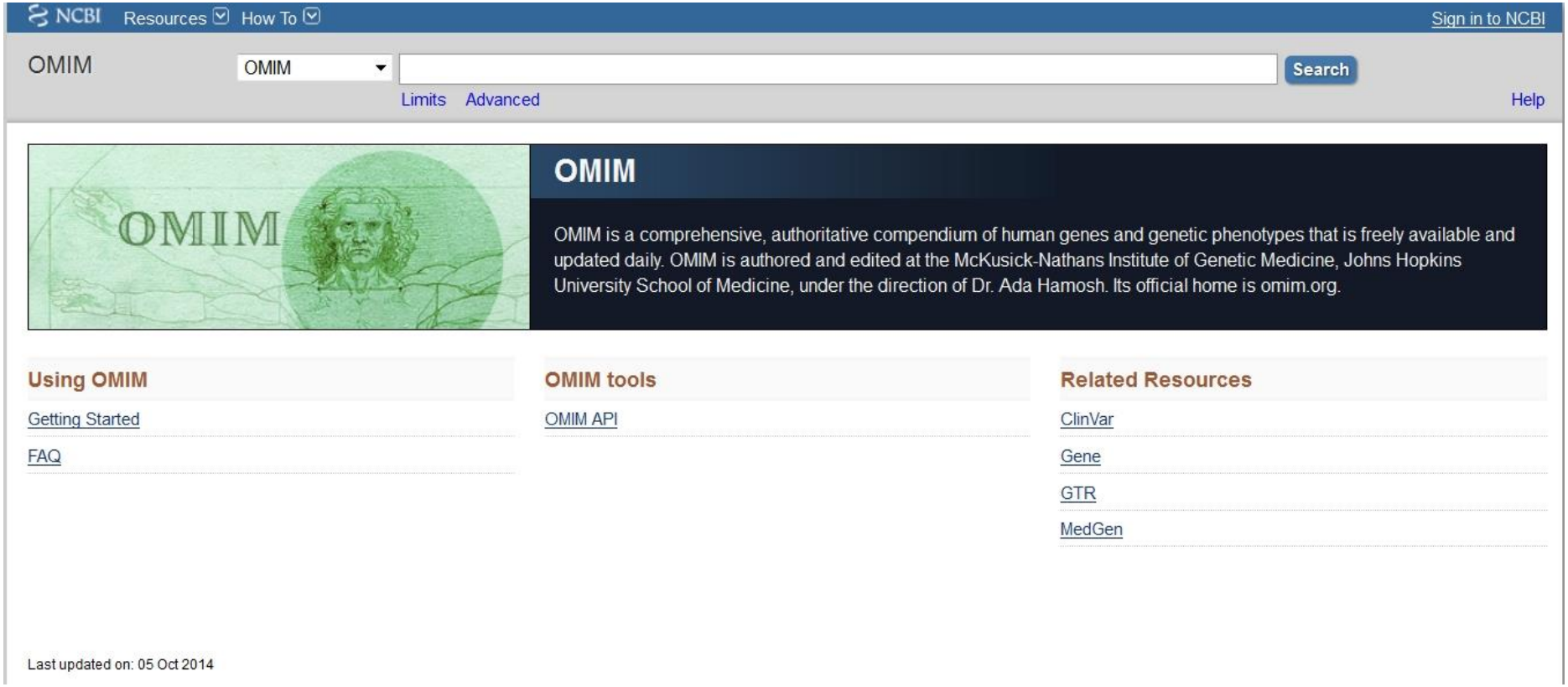
(Barrenas et al 2009: complex disease network – nodes are diseases)

Historical overview: inter-relationships (networks)



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

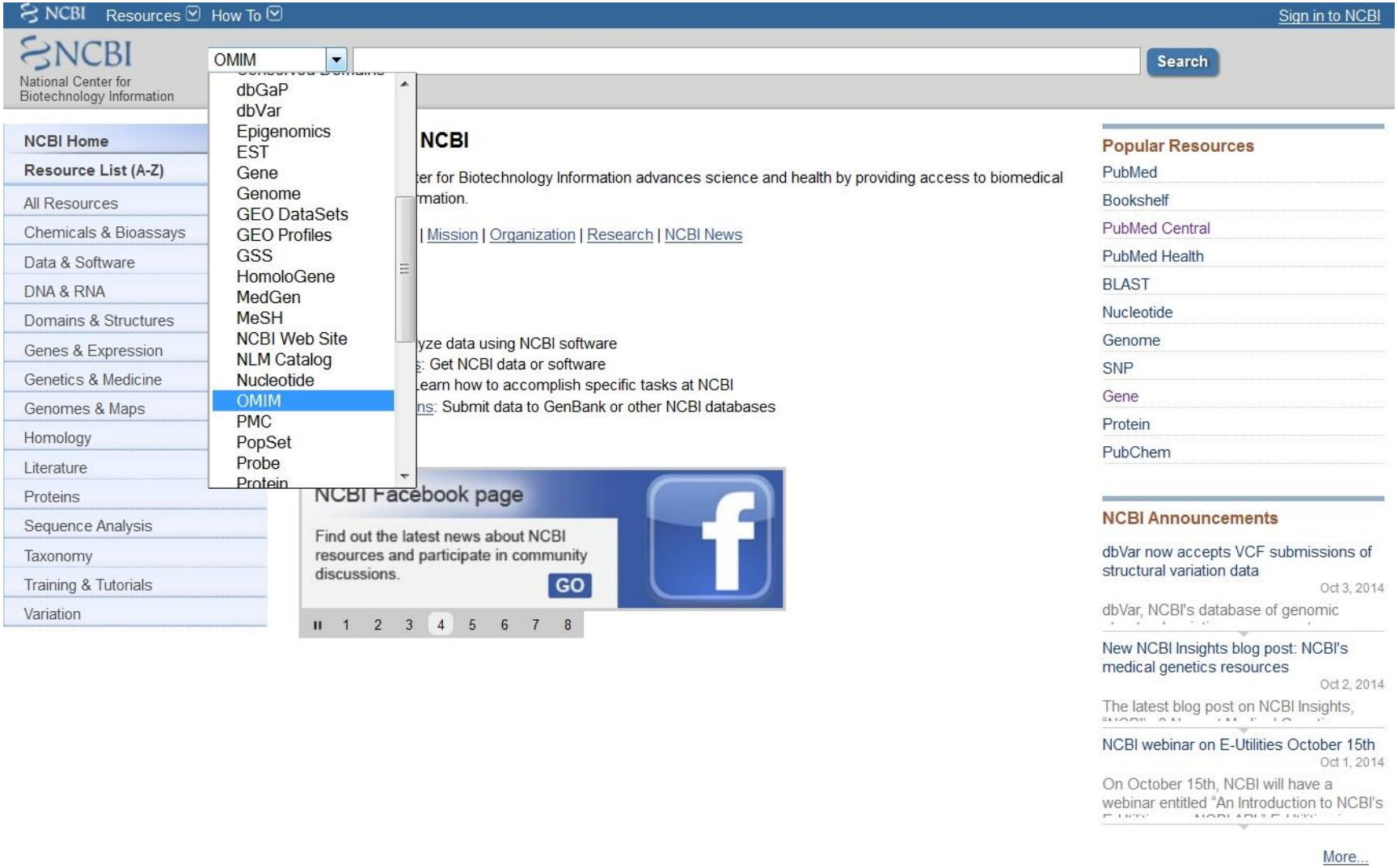
Historical overview: monitoring the progress



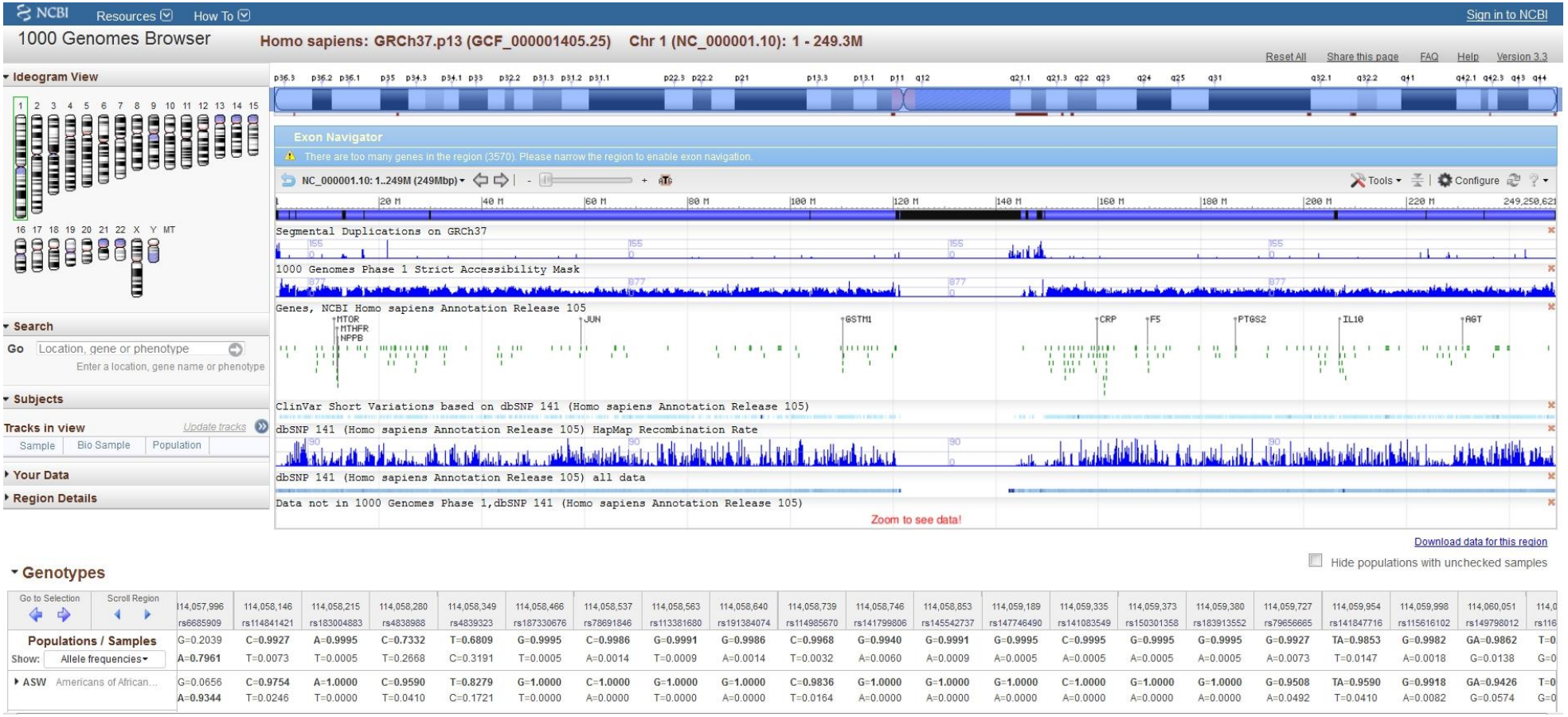
OMIM: molecular dissection of human disease

- 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



1.b The concept of a genetic marker

The evolution 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."

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

Be critical

(date of publication = 2004)

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

The evolution of molecular markers

- Nowadays, **genetic markers represent sequences of DNA** which have been traced to specific locations on the chromosomes and associated with particular traits.
- They demonstrate **polymorphism**, which means that the genetic markers in different organisms of the same species are different.
- A classic example of a genetic marker:
  the area of the DNA which codes for blood type in humans –
  all humans have and need blood, but the blood of individual humans can be very different as a result of polymorphism in the area of the genome which codes for blood.

Types of genetic markers – example 1: microsatellites

- Synonymous: short tandem repeat, STR
- Number of repeats varies between individuals
  - Mononucleotide, dinucleotide, trinucleotide, tetranucleotide, non-integer STRs
- Determine allele length (e.g., 133, 136, 139, 142, …)
- Occurrence in non-coding regions
- High mutation frequency $\approx 10^{-2} – 10^{-4}$ events per locus per generation
- Not easy to score automatically
- Frequent but not dense enough for some applications

(Ziegler and Van Steen, Brazil 2010)

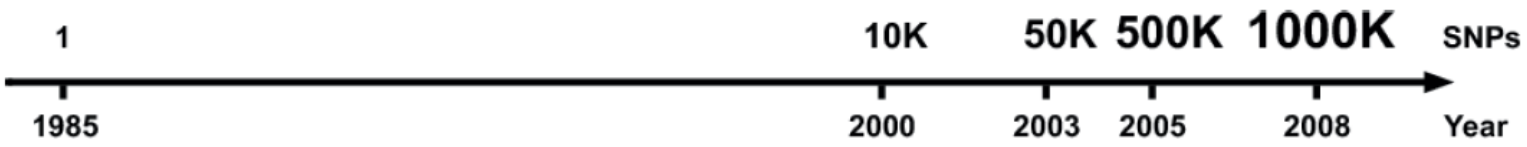
Types of genetic markers – example 2: single nucleotide polymorphisms

- 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)

Why are SNPs preferred over STRs?

- SNPs very frequent ➔ dense marker map
- Some SNPs functionally relevant ➔ candidate variations for disease
- SNPs more stable, i.e., lower mutation rate
- Genotyping in highly automated fashion



(Ziegler and Van Steen, Brazil 2010)

1.c The rise of Genome-Wide Association studies (GWAs)

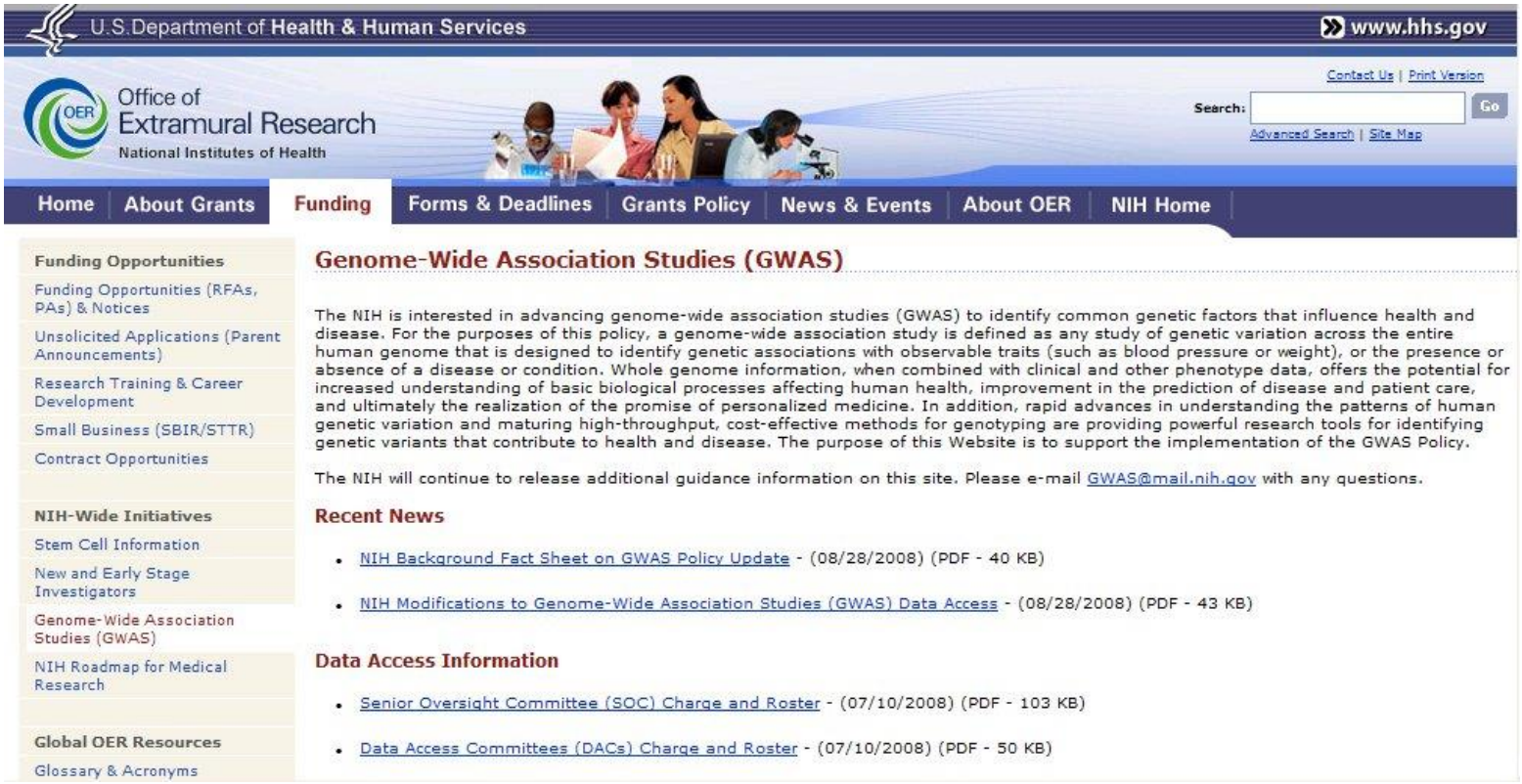
Definition

- 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 for improved public health

- 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		

What do we need to carry out a genome-wide association study?

- The tools include

  - computerized databases that contain the reference human genome sequence,

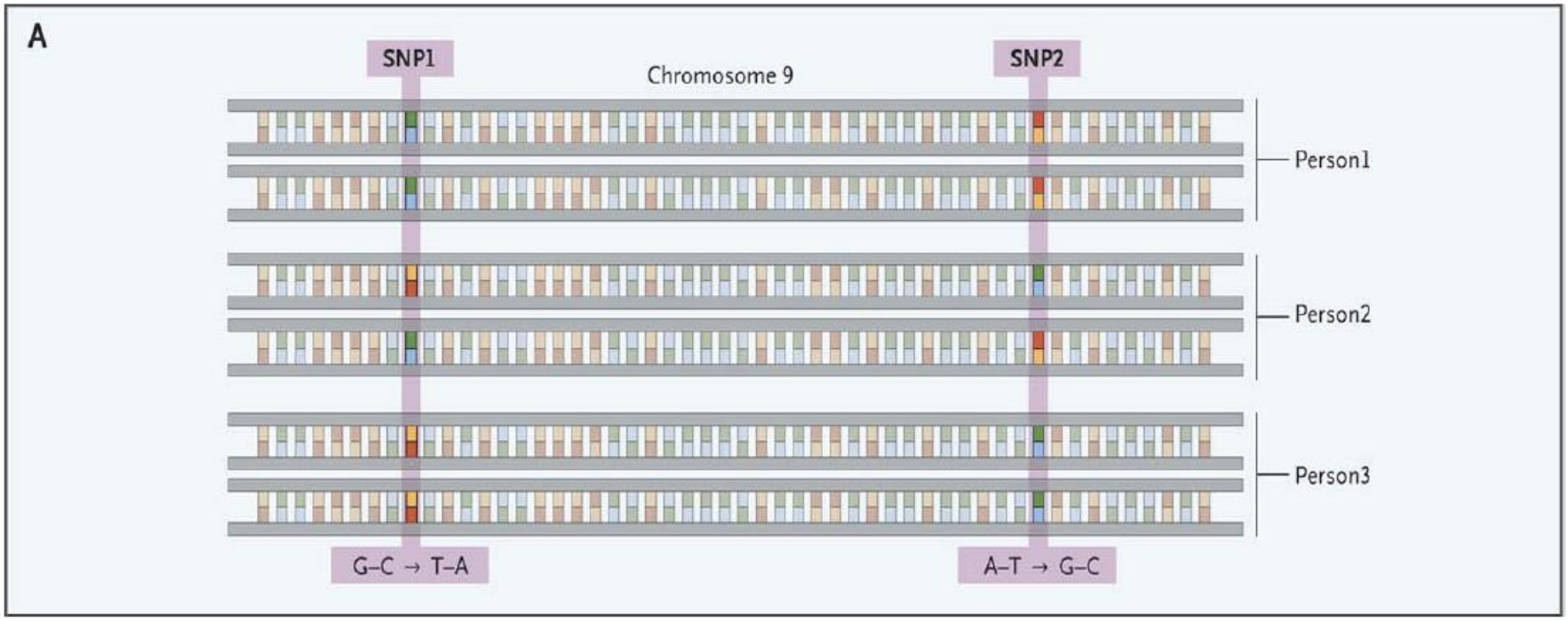
  - a map of human genetic variation and

  - a set of new technologies that can quickly and accurately analyze (whole-genome) samples for genetic variations that contribute to the onset of a disease.

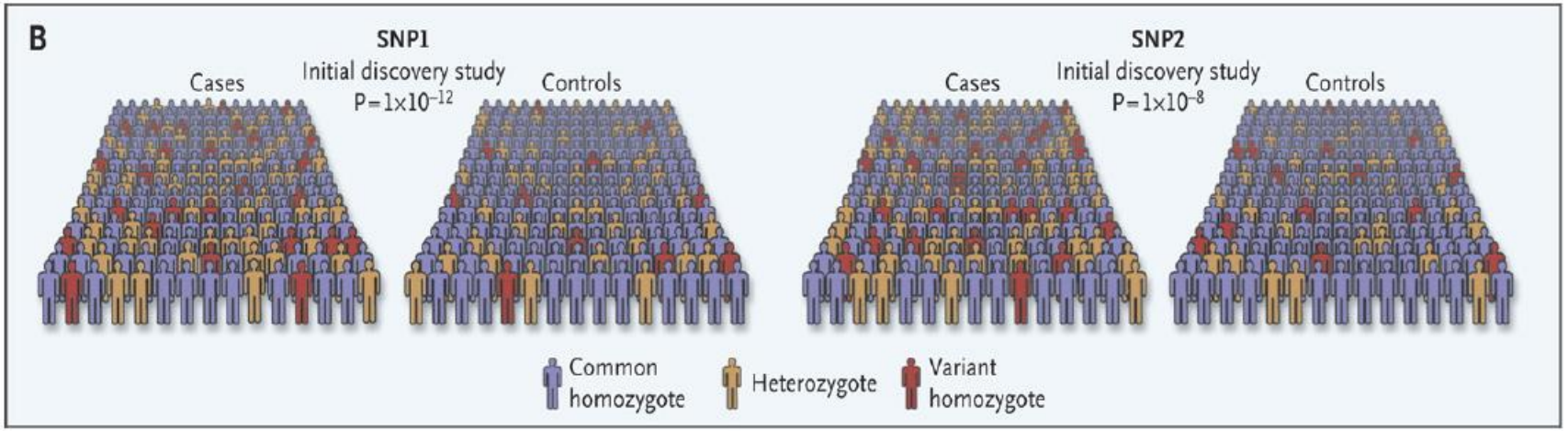
(http://www.genome.gov/pfv.cfm?pageID=20019523)

Genome-wide association studies in practice

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)



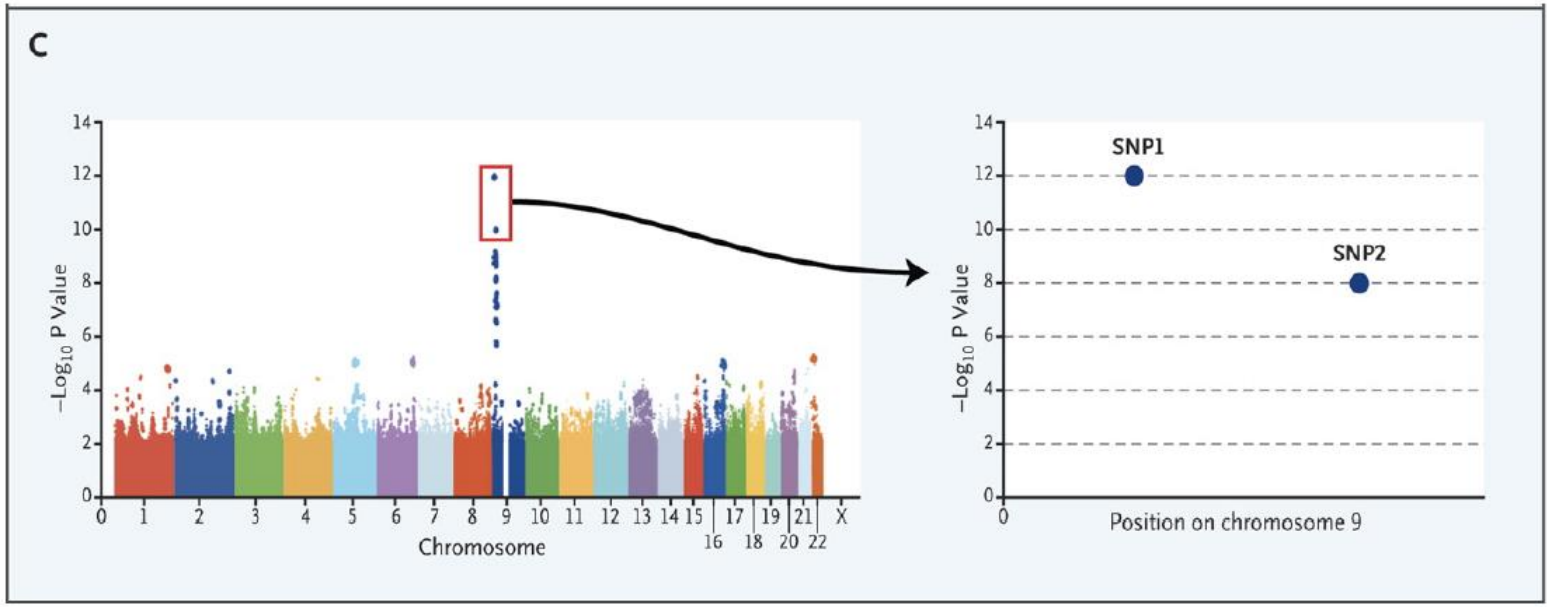
Genome-wide association studies in practice



- 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)

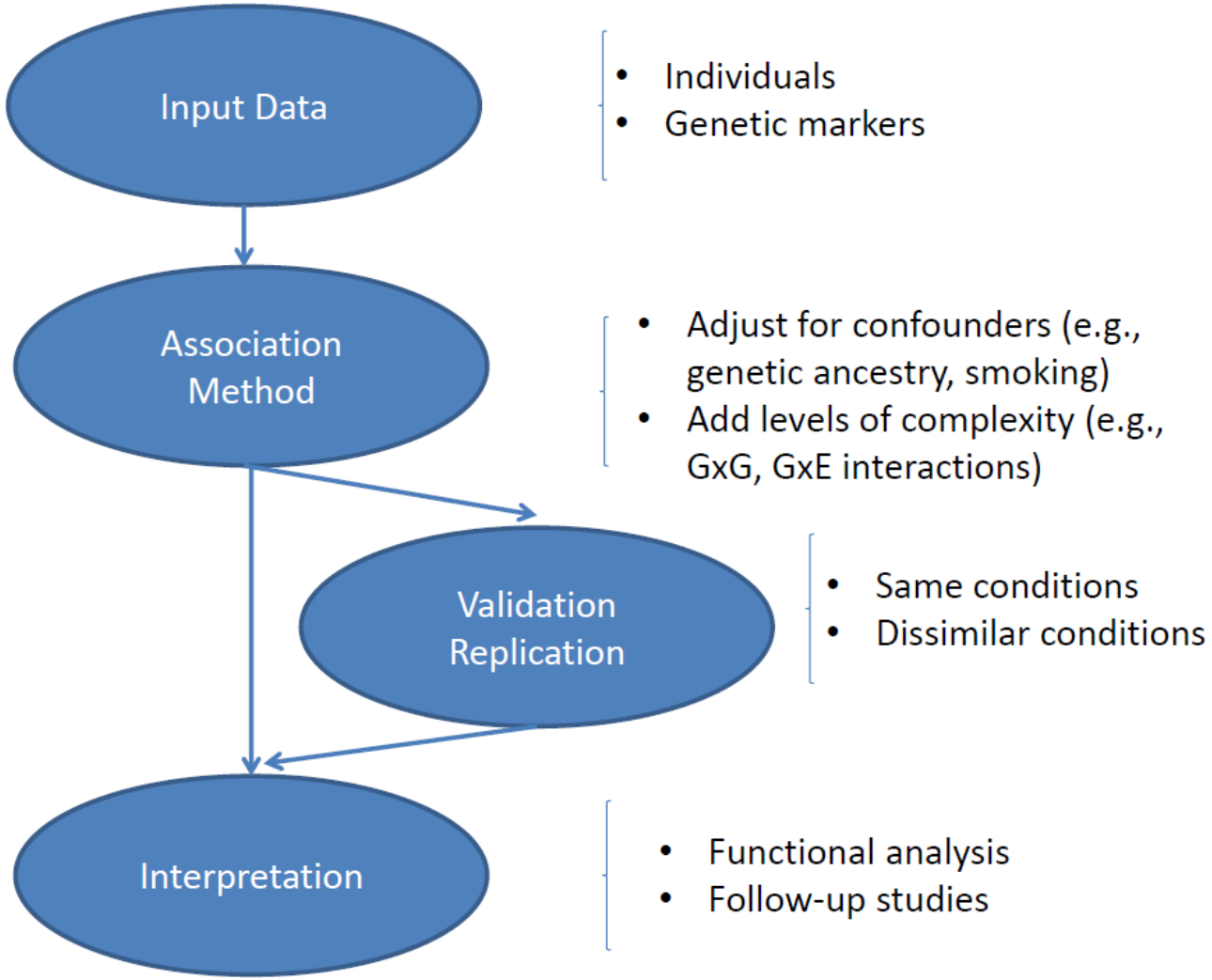
Genome-wide association studies in practice



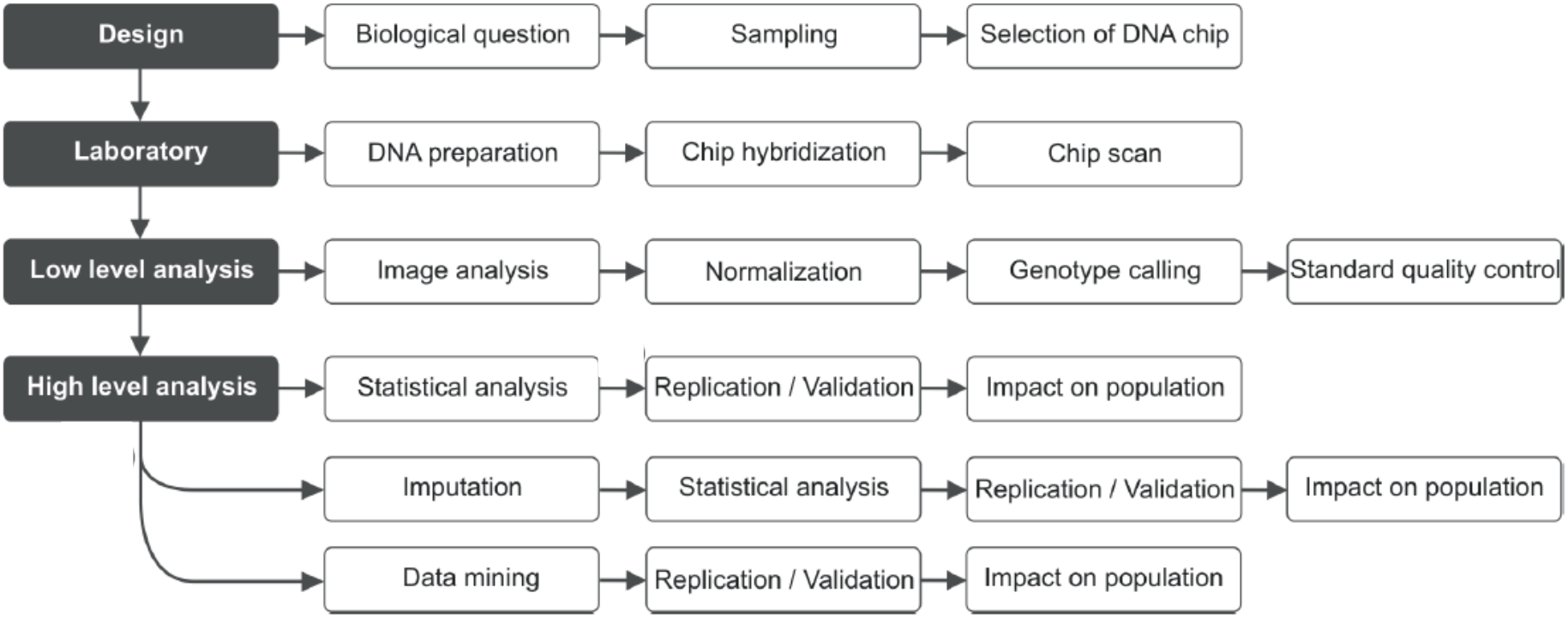
- The plot in Panel C shows the P values for all genotyped SNPs that have survived a quality-control screen (each chromosome, a different color).
- The results implicate a locus on chromosome 9, marked by SNPs 1 and 2, which are adjacent to each other (graph at right), and other neighboring SNPs.

(Manolio 2010)

2 Components of a GWAs



Detailed flow of a genome-wide association study



(Ziegler 2009)

GWAs belong to Bioinformatics research (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

GWAs belong to Bioinformatics research (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

GWAs belong to Bioinformatics research (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

GWAs belong to Bioinformatics research (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

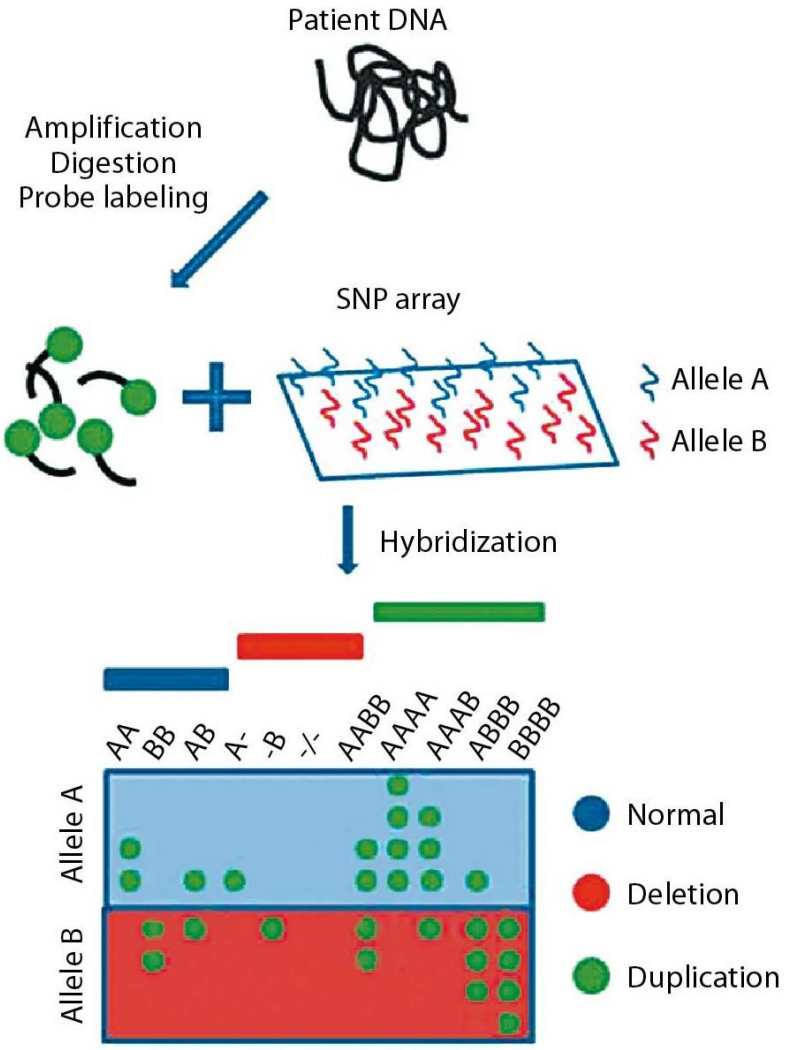
3 Study Design

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., exones 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).

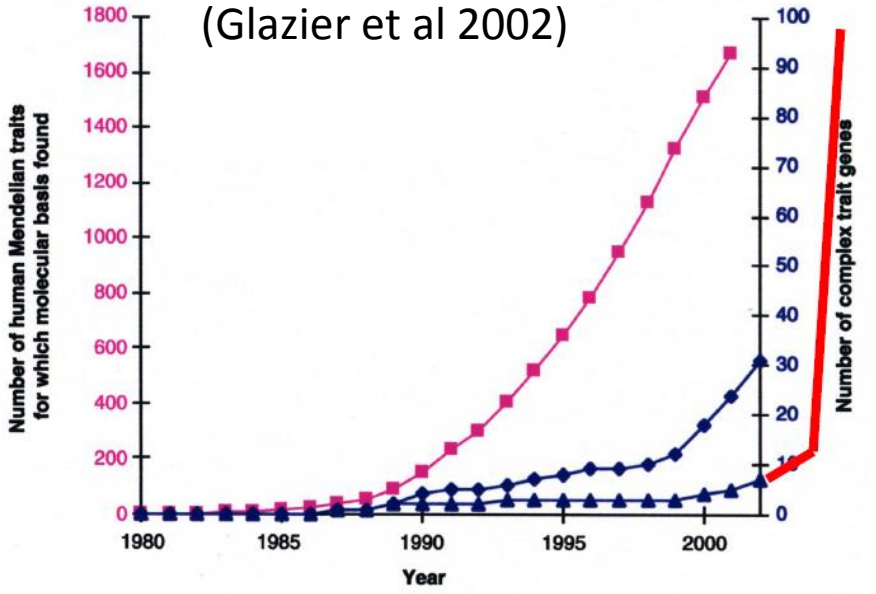


Hypothesis: Common Disease – Common Variant (CDCV)

- This hypothesis predicts that the genetic risk for **common diseases will often be due to disease-predisposing alleles with relatively high frequencies**; there will be one or a few predominating disease alleles at each of the major underlying disease loci (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?
- Whether the CDCV hypothesis is true for most diseases is yet unknown but there are a few prototypical examples: the APOE e4 allele in Alzheimer disease
- An alternative hypothesis = Commond Disease Rare Variant (CDRV) hypothesis

Number of traits for which molecular basis exists

(Glazier et al 2002)



PINK : Human Mendelian traits

BLUE middle line : All complex traits

BLUE bottom line + red extension: Human complex traits

Complex disease (definition):

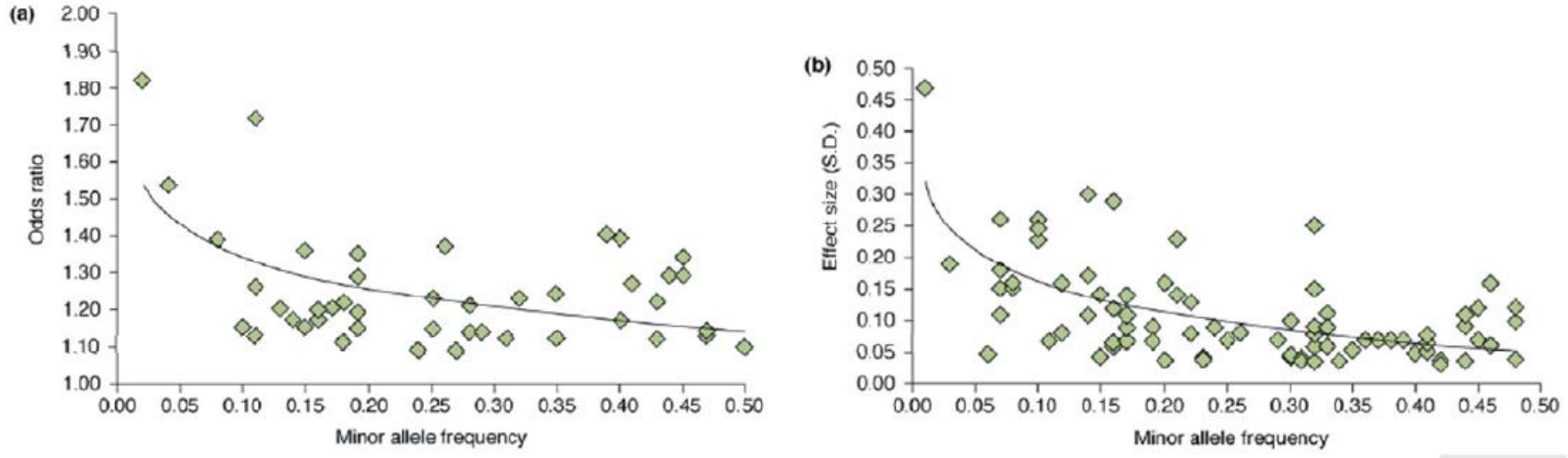
The term complex trait/disease refers to any phenotype that

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

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

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 ….

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. Cases usually mild cases, etc
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. Lower power when compared with case-control studies
3. Sensitive to genotyping errors

4 Pre-analysis steps

3.a Quality control – The Travemünde Criteria

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

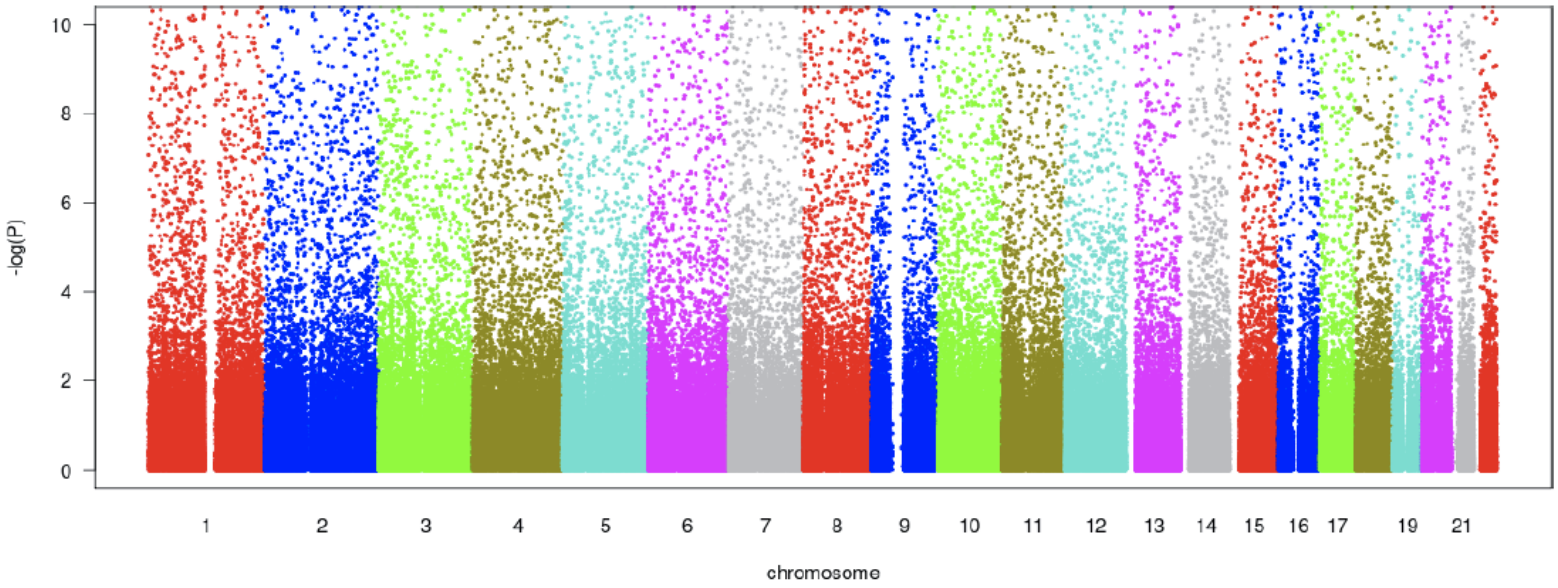
tfam file: First 6 columns of standard ped file

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

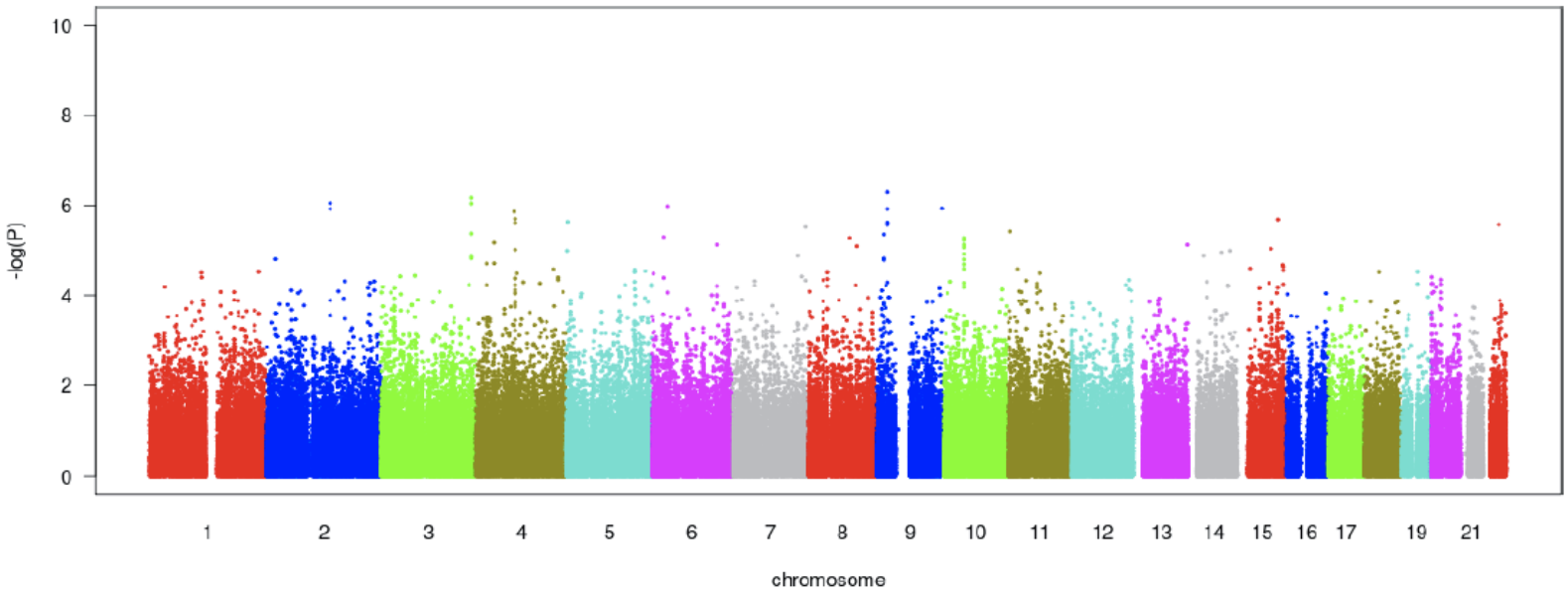


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(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)

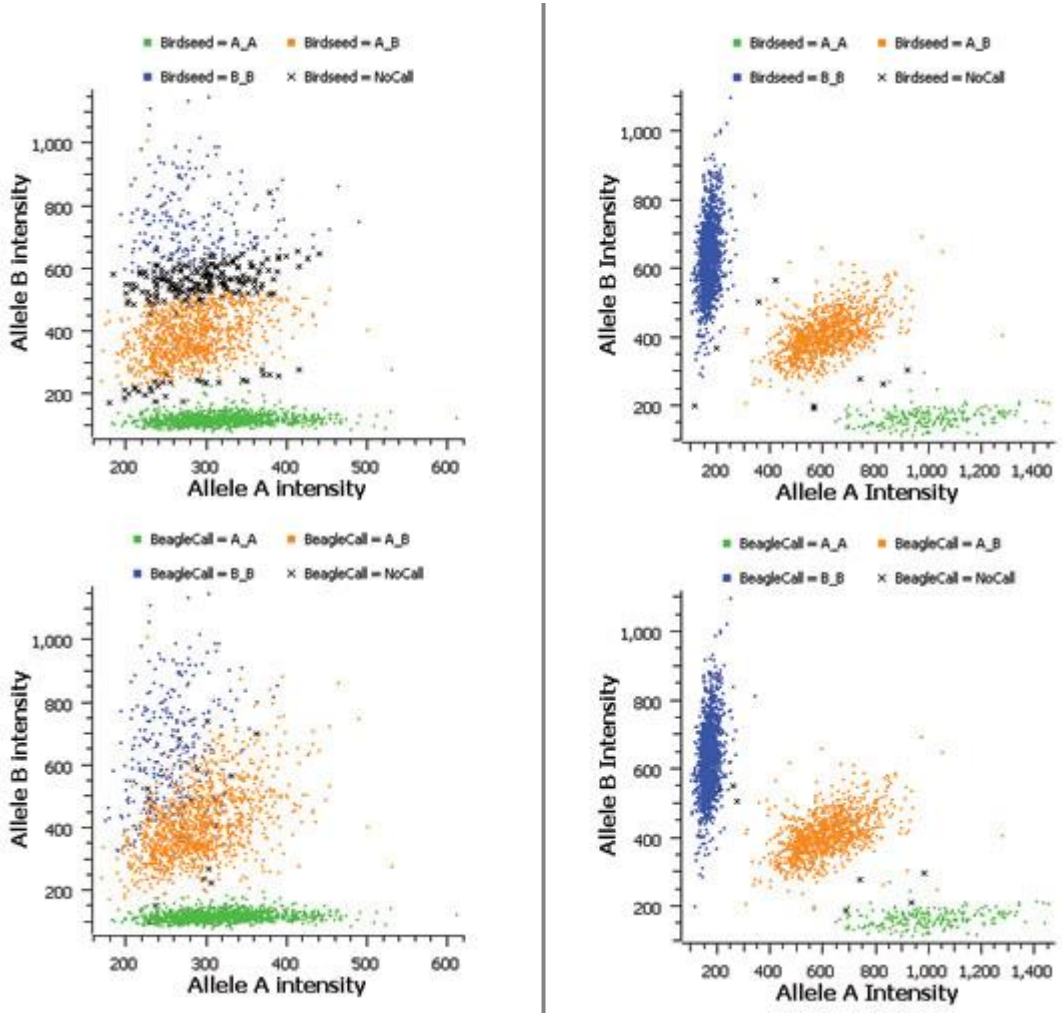
What is the standard quality control?

- Quality control can be performed on different levels:
  - Subject or sample level
  - Marker level (here: SNP level)
  - X-chromosomal SNP level
- Consensus on how to best QC data has led to the so-called “Travemünde criteria” (obtained in the town Travemünde)

Example of QC 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)

Reading genotype calling cluster plots



Allele signal intensity 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 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

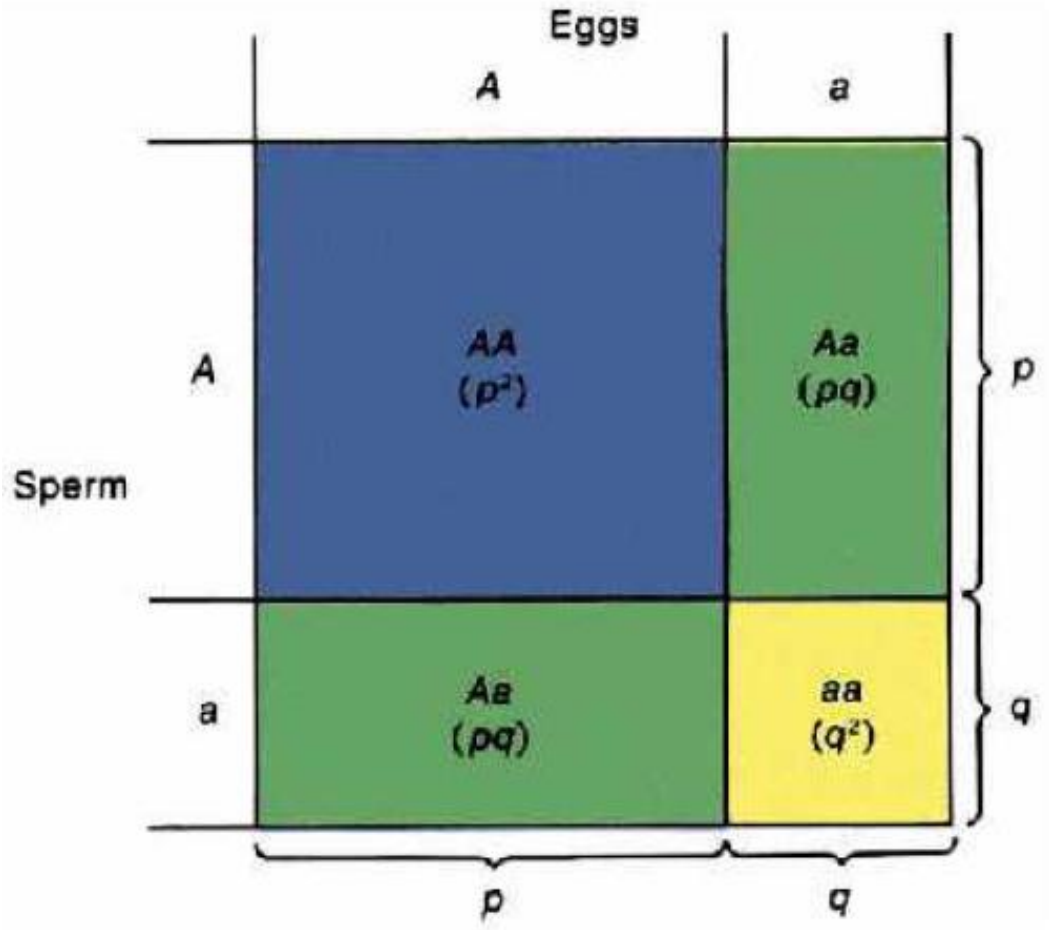
- $P(A_1A_1) = p_{11} = p^2$
- $P(A_1A_2) = p_{12} = 2pq$
- $P(A_2A_2) = p_{22} = q^2$

the population is said to be in HWE at the SNP

(Ziegler and Van Steen 2010)

What is Hardy-Weinberg Equilibrium (HWE)?

Consider diallelic SNP with alleles A and a



Distorting factors to HWE causing evolution to occur

1. **Non-random mating**

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

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

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

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

The Travemünde criteria

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

(Ziegler 2009)

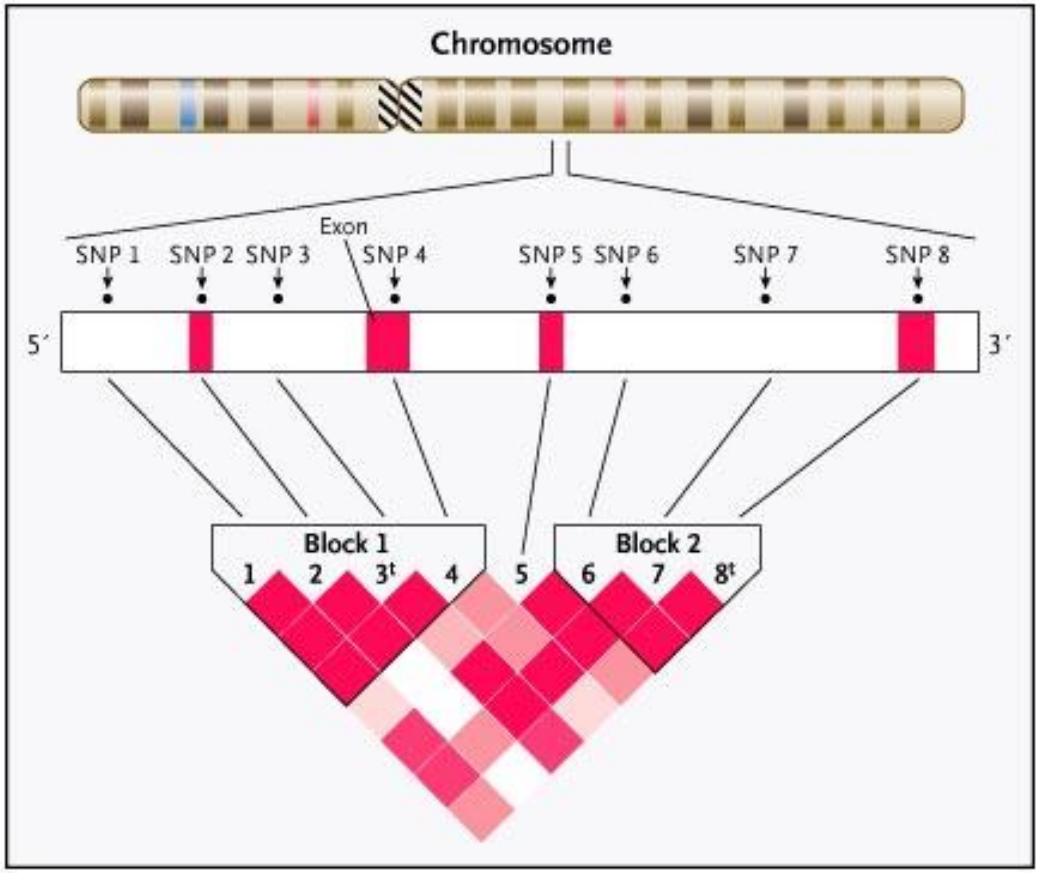
The Travemünde criteria

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

(Ziegler 2009)

4.b Linkage disequilibrium

Mapping the "relationships" between SNPs (Christensen and Murray 2007)



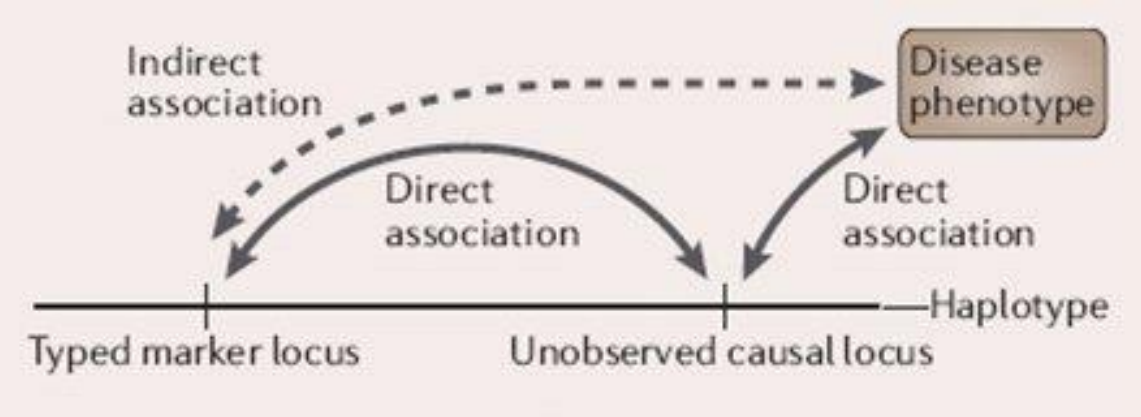
(HaploView software)

Linkage disequilibrium (LD) between genetic markers

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



Distances among SNPs

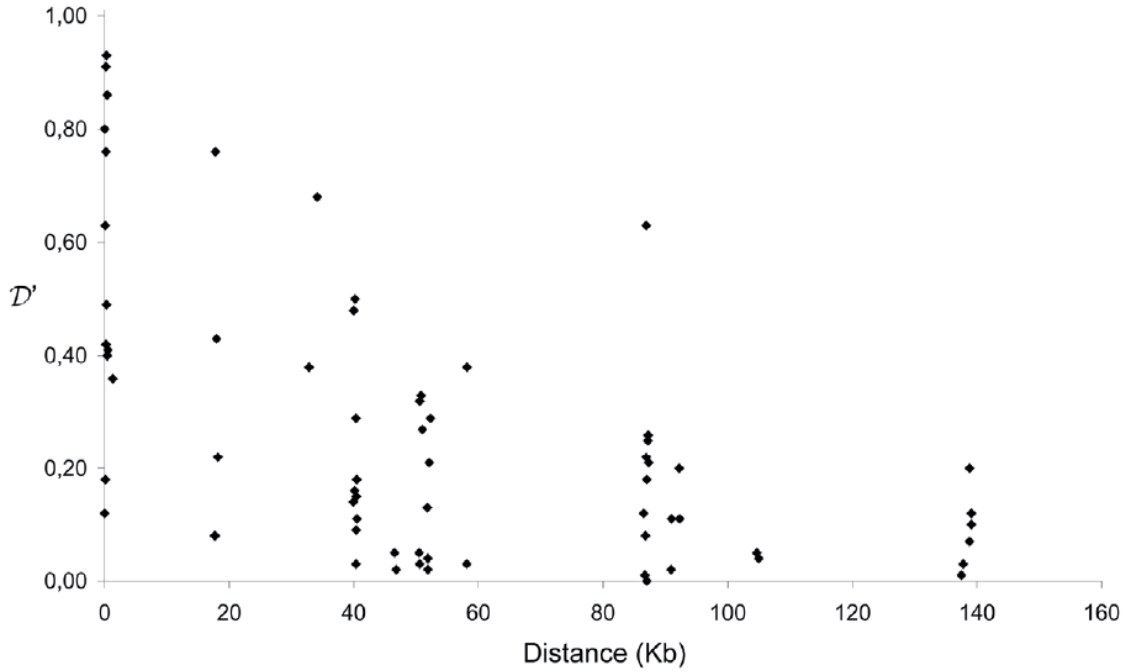
- The measure D is defined as the difference between the observed and expected (under the null hypothesis of independence) proportion of "haplotypes" bearing specific alleles at two loci: $p_{AB} - p_A p_B$

	A	a
B	p_{AB}	p_{aB}
b	p_{Ab}	p_{ab}

- A **haplotype** is a linear arrangement of alleles on the same chromosome that have been inherited as a unit.

- Because of its interpretation, the **measure r^2 (coefficient of determination**) is most often used for GWAs

How far does linkage disequilibrium extend?



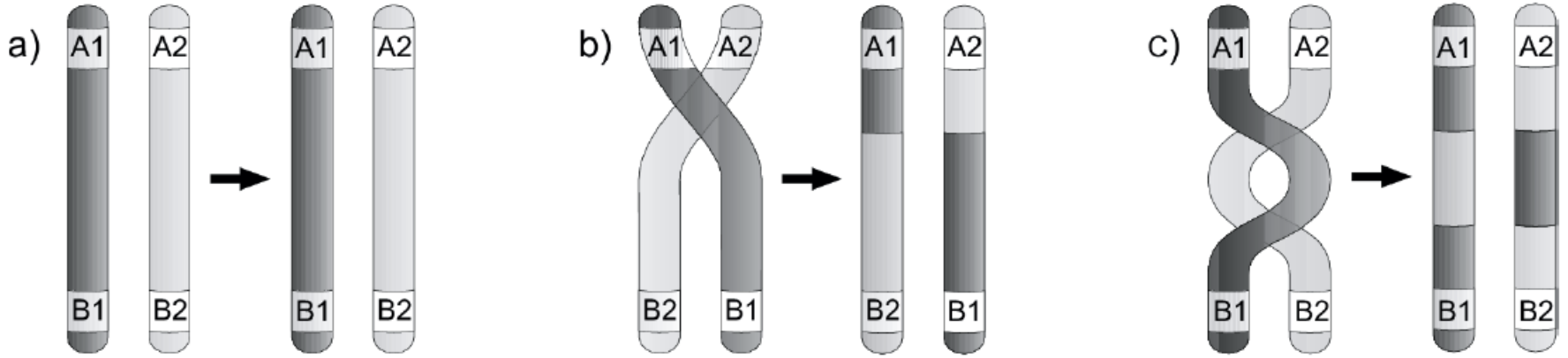
D' (Lewontin's *D* prime) is the absolute ratio of D compared with its maximum value.

D' =1 : complete LD

(Hecker et al 2003)

- LD is usually a function of distance between the two loci. This is mainly because recombination acts to break down LD in successive generations (Hill, 1966).

Recombination



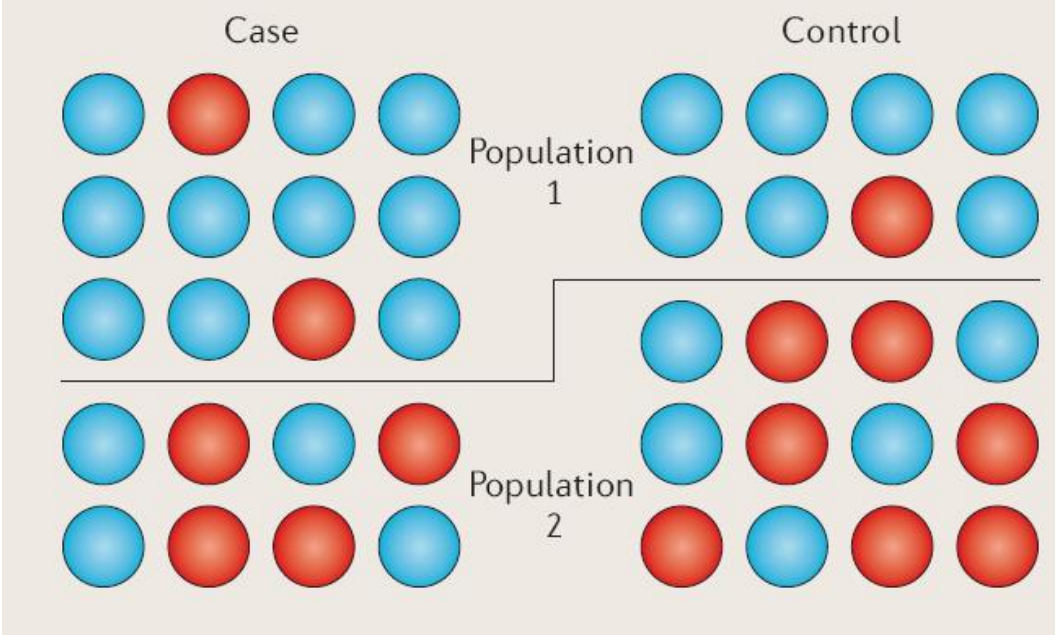
- Relevant measure: recombination fraction (probability of odd number of crossovers) between two chromosomal positions
- Strong correlation between recombination fraction and distance in base pairs

(Ziegler and Van Steen, Brazil 2010)

4.c Confounding: population stratification

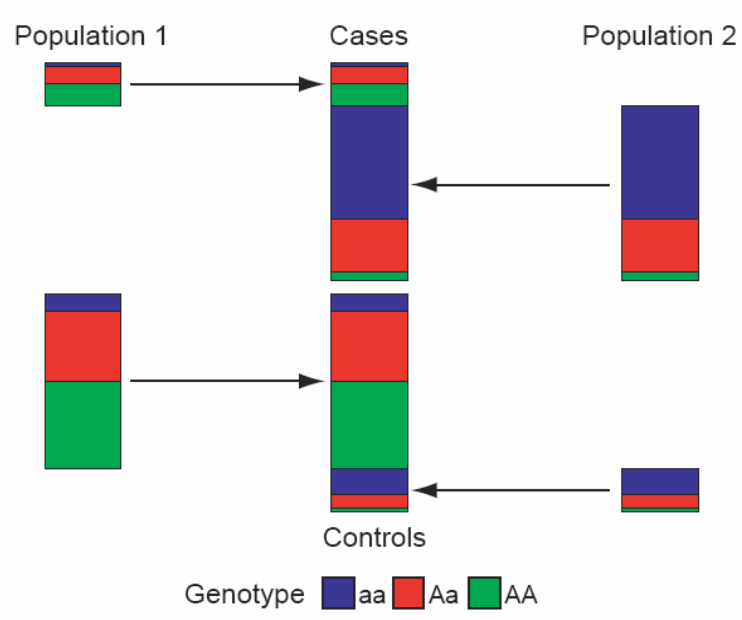
What is spurious association?

- **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.



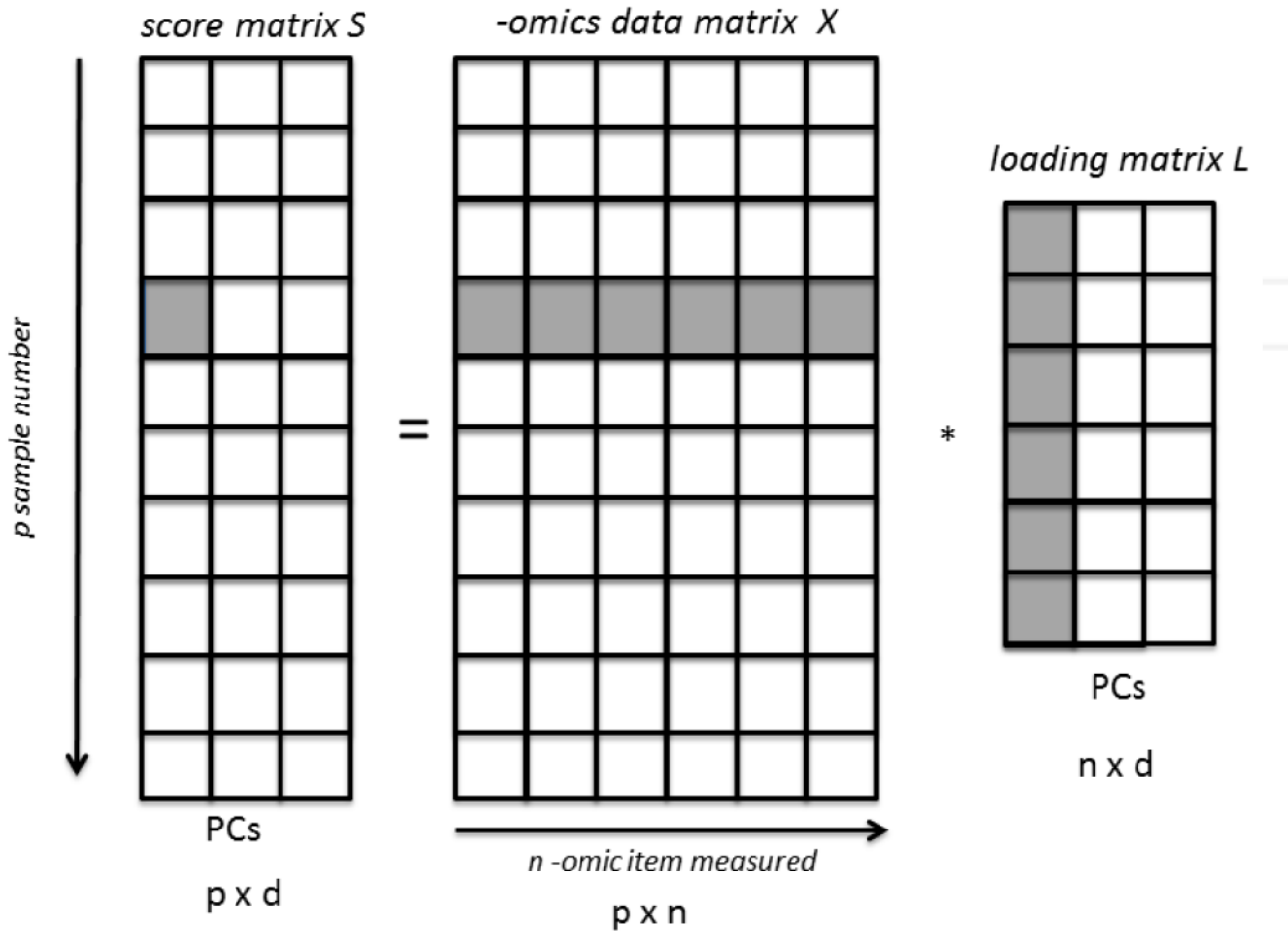
What are typical methods to deal with population stratification?

- 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 (and omics: http://cdn.intechopen.com/pdfs-wm/30002.pdf)



$$s_{4,1} = x_{4,1}.l_{1,1} + x_{4,2}.l_{2,1} + x_{4,3}.l_{3,1} + x_{4,4}.l_{4,1} + x_{4,5}.l_{5,1} + x_{4,6}.l_{6,1}$$

Principal components

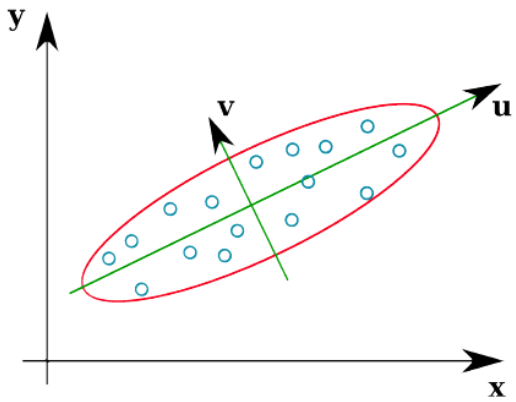


Figure 1: PCA for Data Representation

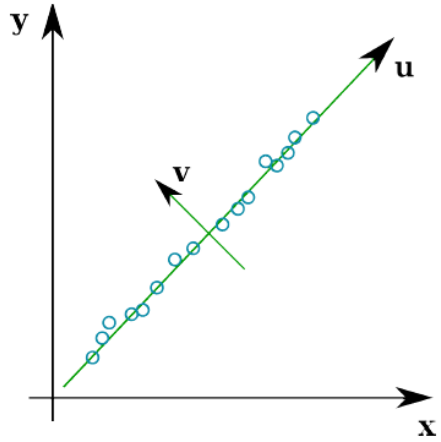


Figure 2: PCA for Dimension Reduction

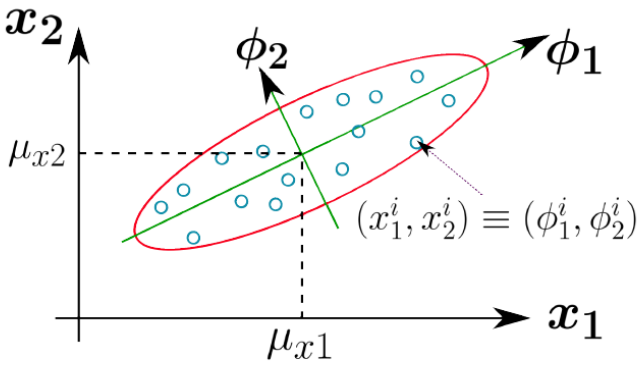
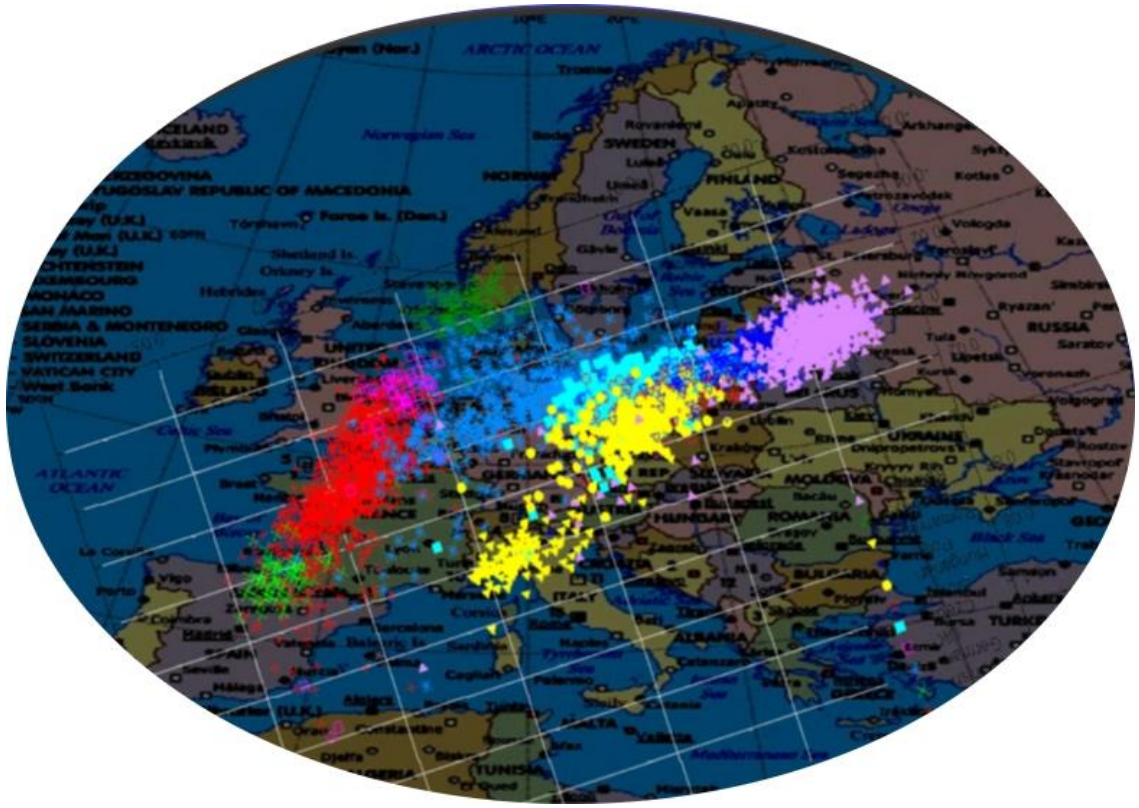


Figure 3: The PCA Transformation

- Find eigenvectors of the covariance matrix for standardized (x1, x2, …) [→SNPs]
- These will give you the direction vectors indicated in Fig3 by phi_1 and phi_2
- These determine the axes of maximal variation

Principal components

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



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

5 Testing for Associations

One SNP at a time

	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

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 colums → **genotype test**

	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 colums → **allelic test (ONLY to be used under HWE)**

	A	**a**
Cases		
Controls		

Sum of entries is 2 x (cases + controls)

Testing for association between case/control status and a SNP

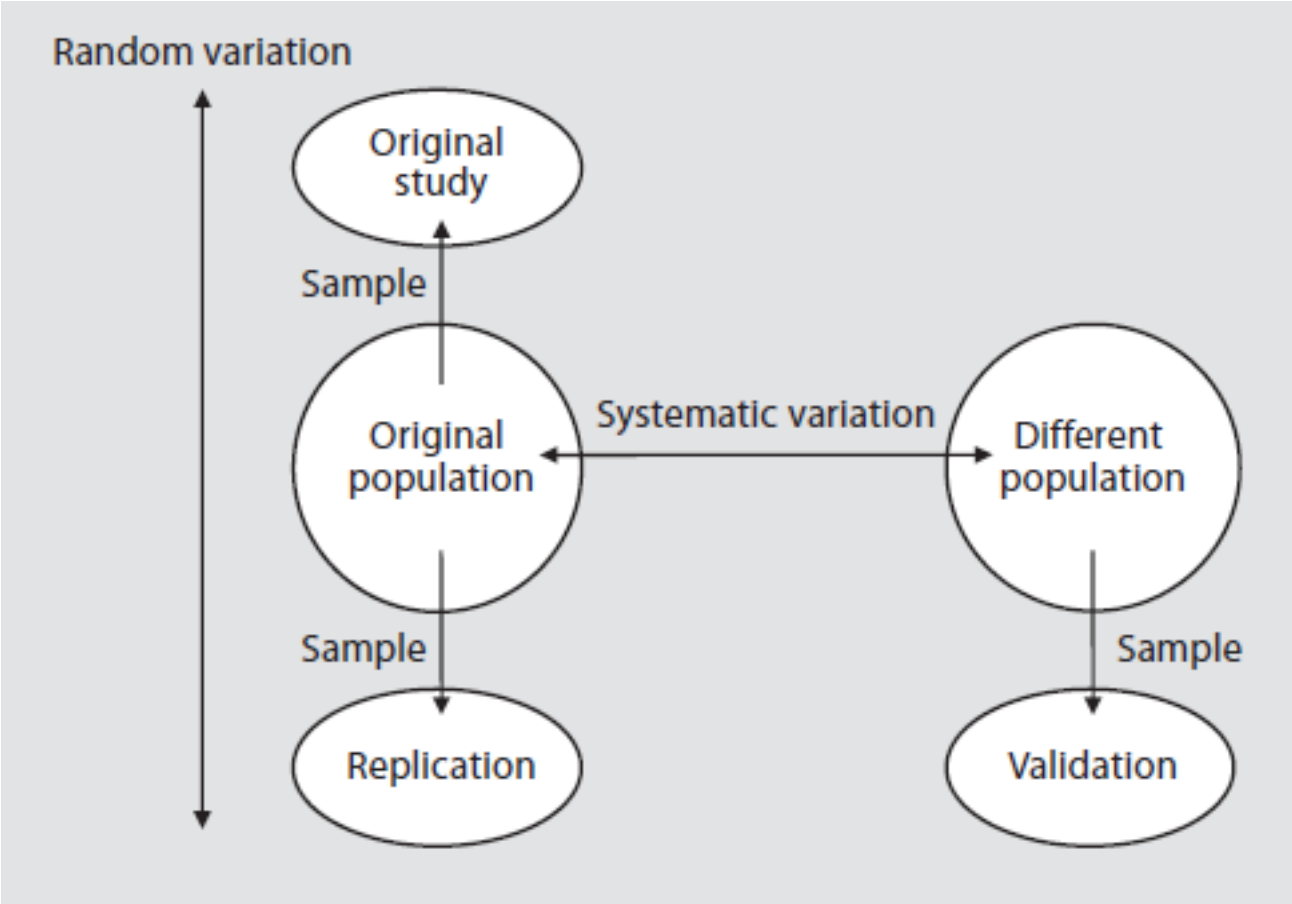
- The genotype test involves a 2df test (note that two variables X1 and X2 were needed for genotype coding).
- It has been shown that usually, the additive coding gives adequate power, even when the true underlying mode of inheritance is NOT additive (note that the additive coding can be achieved by 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.
- How can we adjust such a test for population substructure?

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
  - $\hat{\lambda} = \frac{\text{median}(\chi_1^2, \chi_2^2, \ldots, \chi_L^2)}{\text{median}(\mathcal{L}(\chi_1^2))} = \frac{\text{median}(\chi_1^2, \chi_2^2, \ldots, \chi_L^2)}{0.456}$
  - $\chi_{GC}^2 = \chi^2 / \hat{\lambda}$

6 Replication and Validation

The difference



(Igl et al. 2009)

Guidelines for replication studies

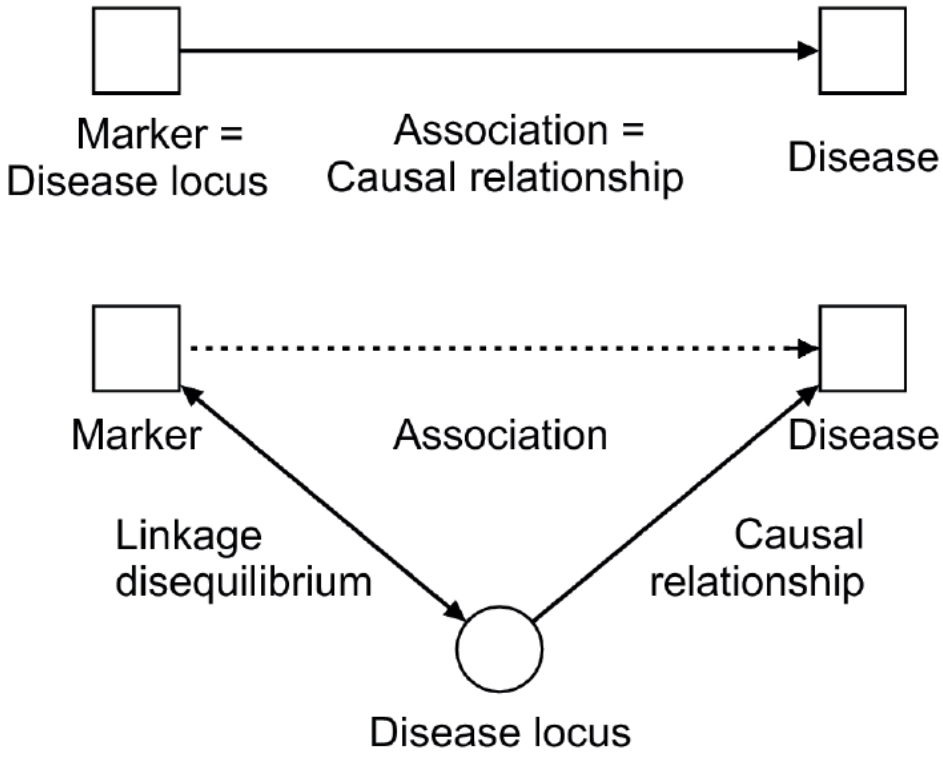
- Replication studies should be of sufficient size to demonstrate the effect
- Replication studies should conducted in independent datasets
- Replication should involve the same phenotype
- Replication should be conducted in a similar population
- The same SNP should be tested
- The replicated signal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

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)

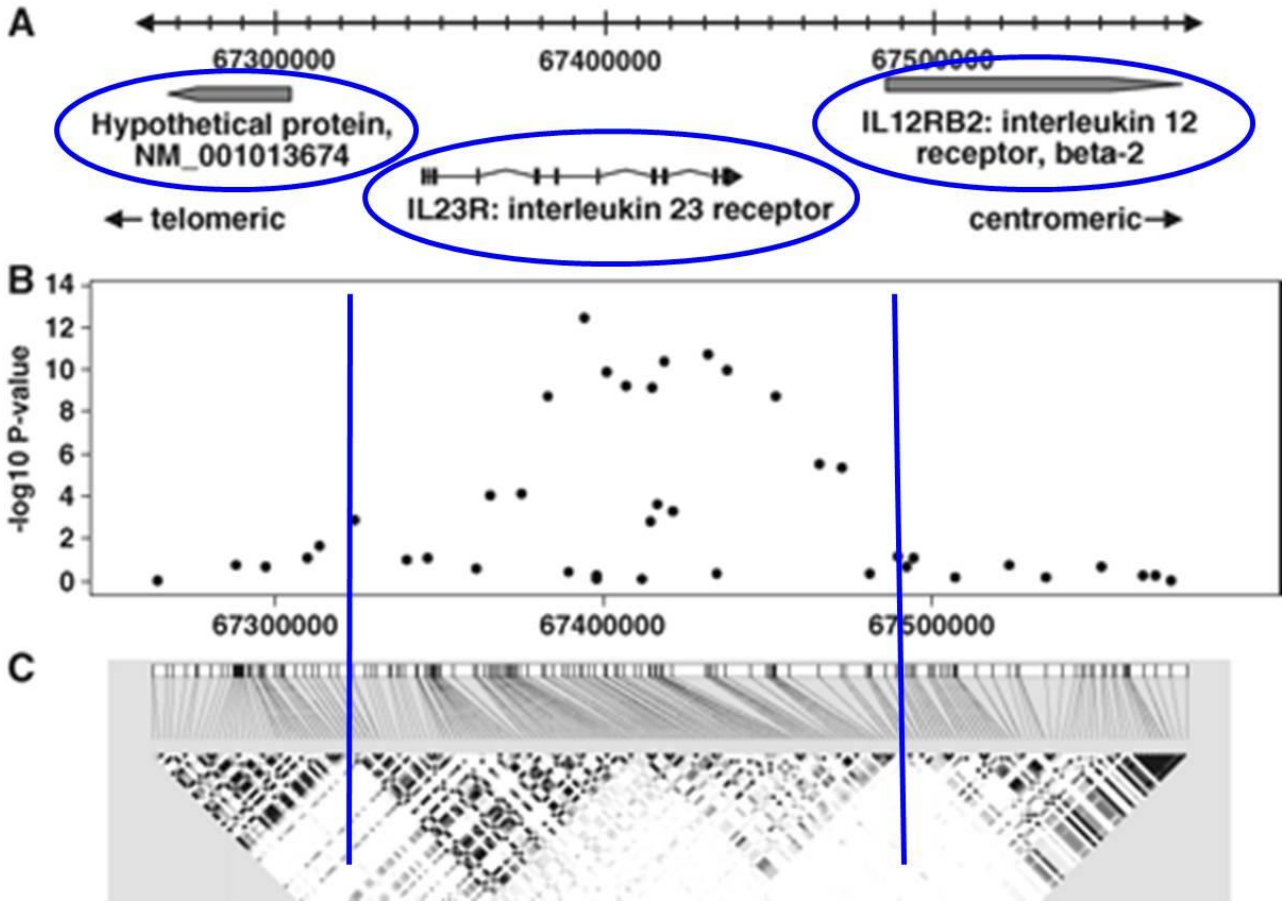
7 GWA Interpretation and Follow-Up

Finding the causal locus



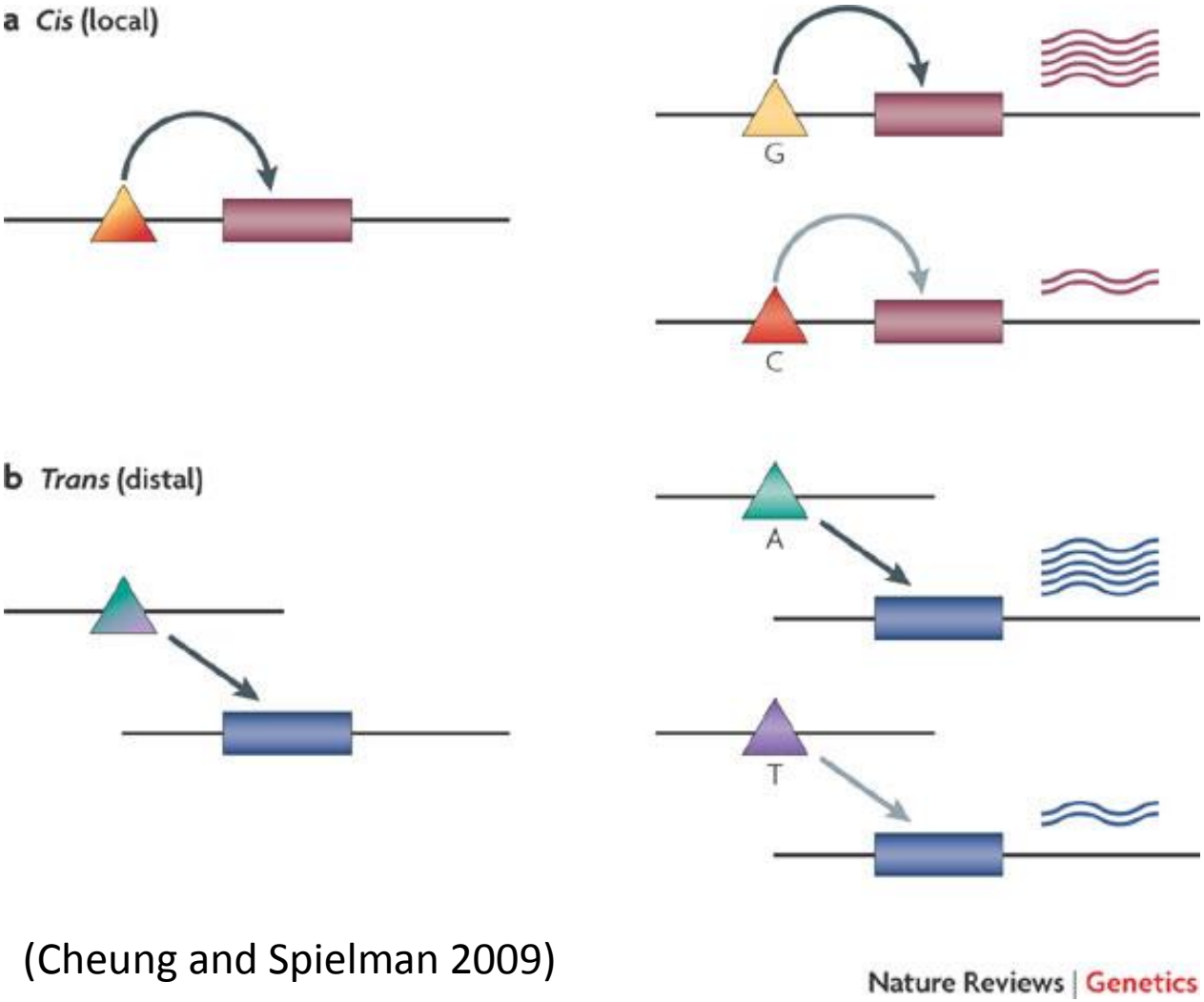
(Ziegler and Van Steen, Brazil 2010)

Finding the causal locus



(Duerr et al 2006)

Finding the causal locus



(Cheung and Spielman 2009)

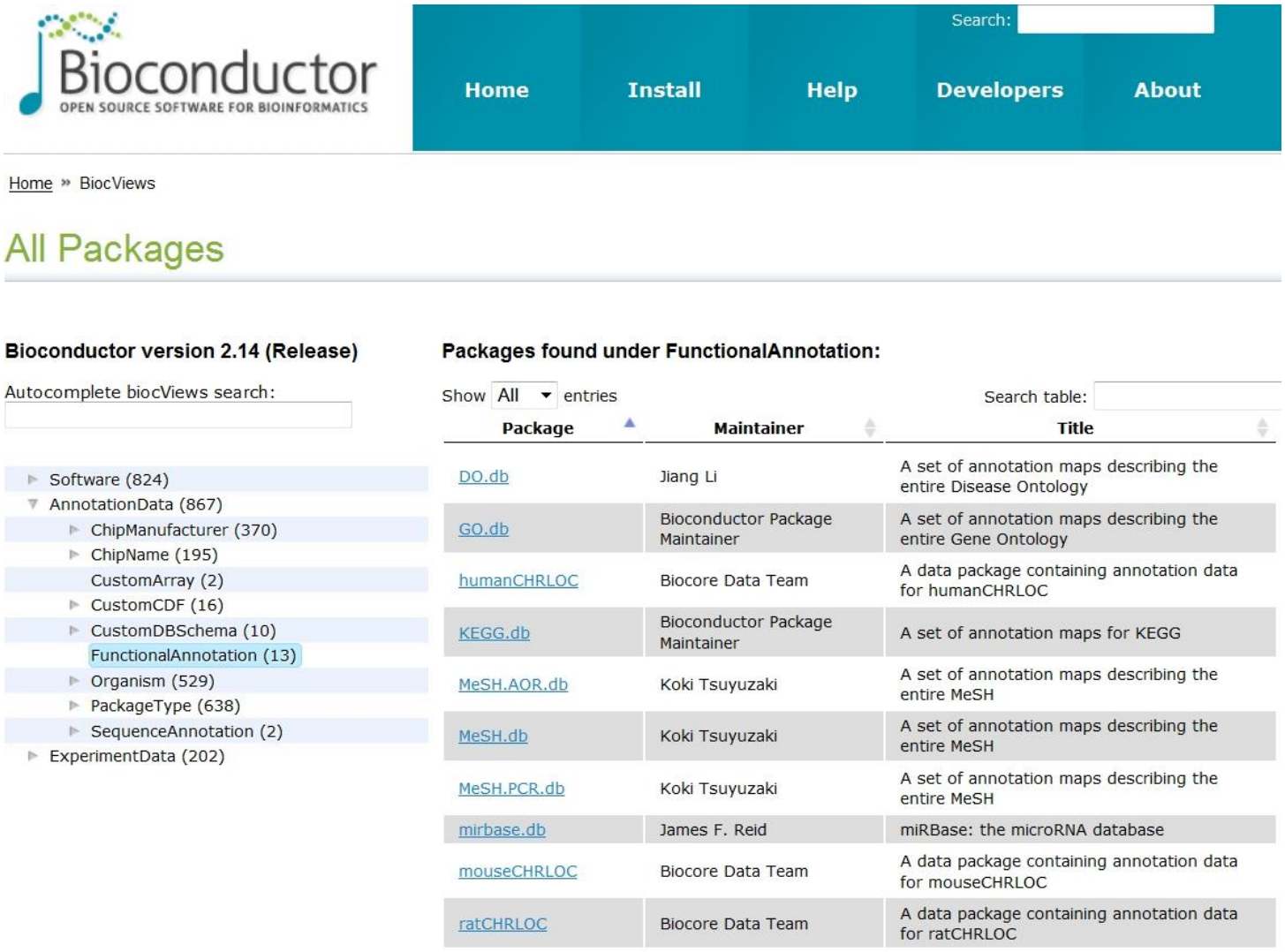
- Cis-acting variants are found close to the target genes and trans-acting variants are located far from the target genes, often on another chromosome.
- Different allelic forms of the cis- and trans-acting variants have different influence on gene expression.

Finding function

- 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.
- Many **functional annotation databases** have been developed in order to classify genes according their various roles in the cell. E.g., the comprehensive Gene Ontology (GO) is one of the most widely used by many functional enrichment tools

(Glass and Girvan 2014)

Functional annotation in R

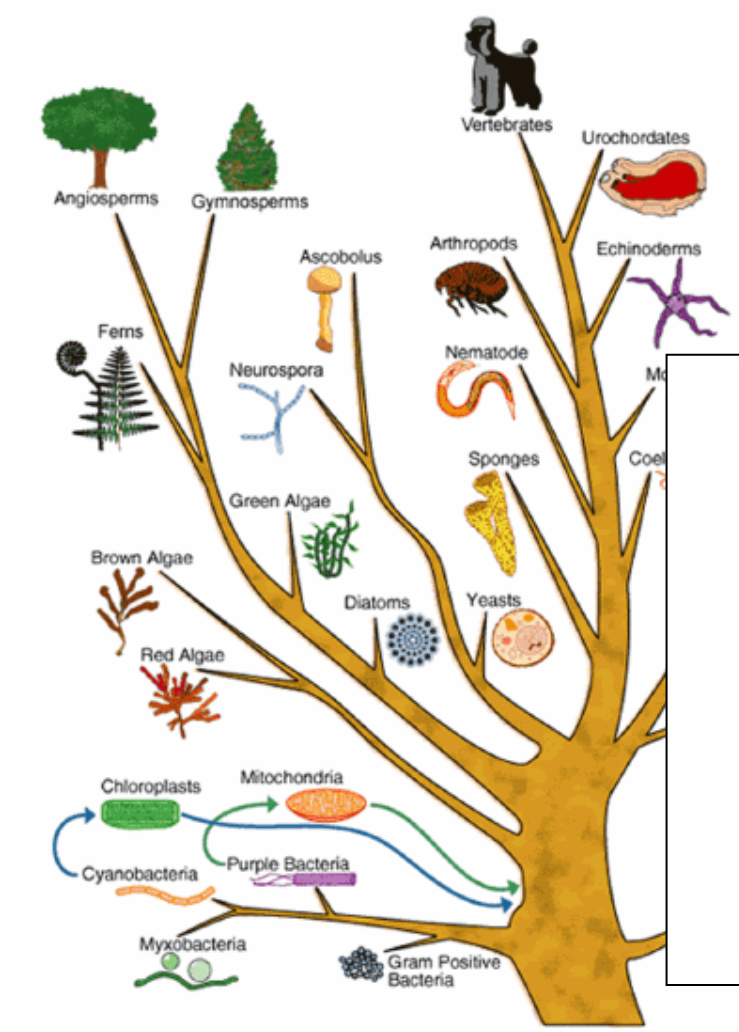


Finding function

- Once a genomic region of interest has been sequenced, bioinformatic analysis can be used to determine if the sequence is similar to that of a known gene.
- This is where sequences from model organisms are helpful.
- For example, let's say 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.
    - It is a good bet that the human sequence also is part of a gene that codes for a membrane protein that regulates salt balance.

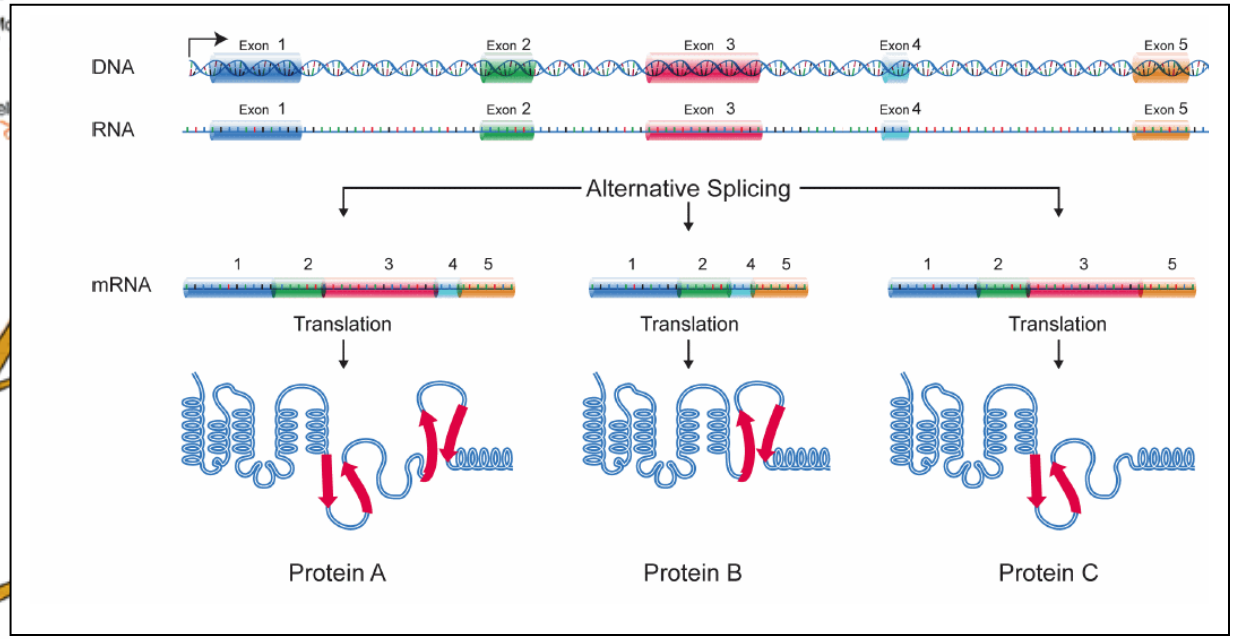
→Links to DNA-seq lectures (sequence comparisons)

Finding function



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:**



"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

Some References:

- Ziegler A and Van Steen K 2010: IBS short course on “Genome-Wide Association Studies”
- Balding D 2006. A tutorial on statistical methods for population association studies. Nature Reviews Genetics, 7, 781-791.
- Kruglyak L 2008. The road to genomewide association studies. Nature Reviews Genetics 9: 314-
- Wang et al 2005. Genome-wide association studies: theoretical and practical concerns. Nature Reviews Genetics 6: 109-
- Peltonen L and McKusick VA 2001. Dissecting human disease in the postgenomic era. Science 291, 1224-1229
- Li 2007. Three lectures on case-control genetic association analysis. Briefings in bioinformatics 9: 1-13.
- Rebbeck et al 2004. Assessing the function of genetic variants in candidate gene association studies 5: 589-
- Robinson 2010. Common Disease, Multiple Rare (and Distant) Variants. PLoS Biology 8(1): e1000293