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 A hype about GWA studies

1.b Speaking the language: relevant questions

1.c Genome-wide association studies

2 GWA Details: Study Design

2.a Marker level

2.b Subject level

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

3 GWA Details: Prior Analyses

3.a Quality Control:

Hardy-Weinberg equilibrium

Missingness

The Travemünde criteria

3.b Linkage disequilibrium and SNP tagging

3.c Confounding: population stratification

4 GWA Details: Testing for Associations

4.a Single SNP (see TA session)

4.b Multiple SNPs

Regression-based tests (see TA session)

Multiple testing correction

4.c Replication and validation

5 GWA Interpretation and Follow-Up

1 Setting the pace

1.a A hype about GWA studies

"May he live in interesting times;

Like it or not we live in interesting times."

Robert Kennedy, June 7, 1966

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."

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 is 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.

Genetic mapping

- 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 - also called linkage mapping - 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)

- 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)

Having a genetic map : now what ?

Science
AAAS

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?

Pennisi 2007 Science 318:1842-3

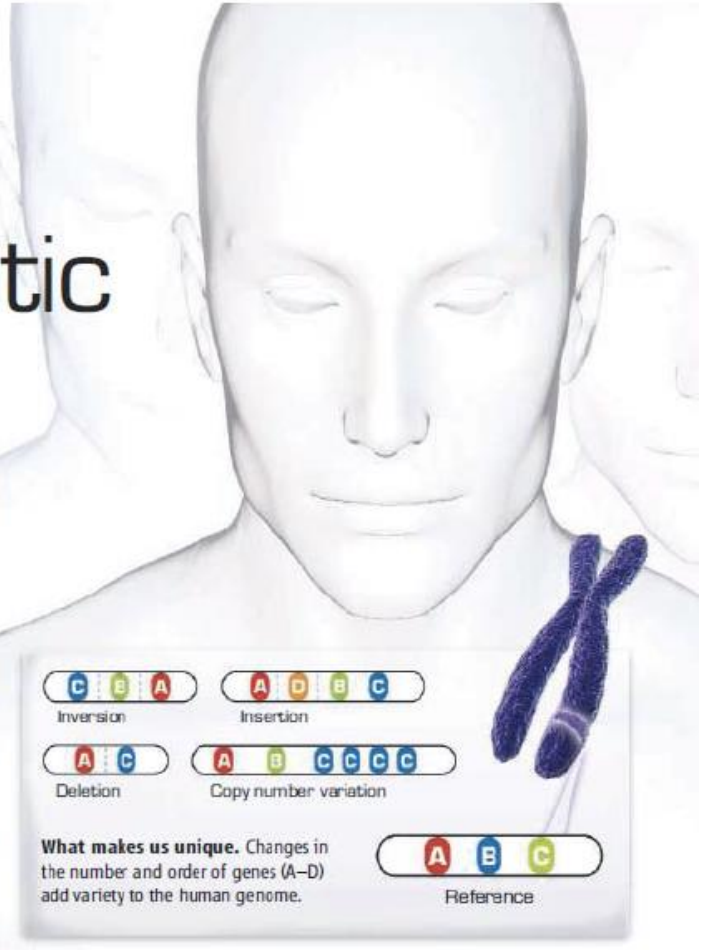
BREAKTHROUGH OF THE YEAR

Human Genetic Variation

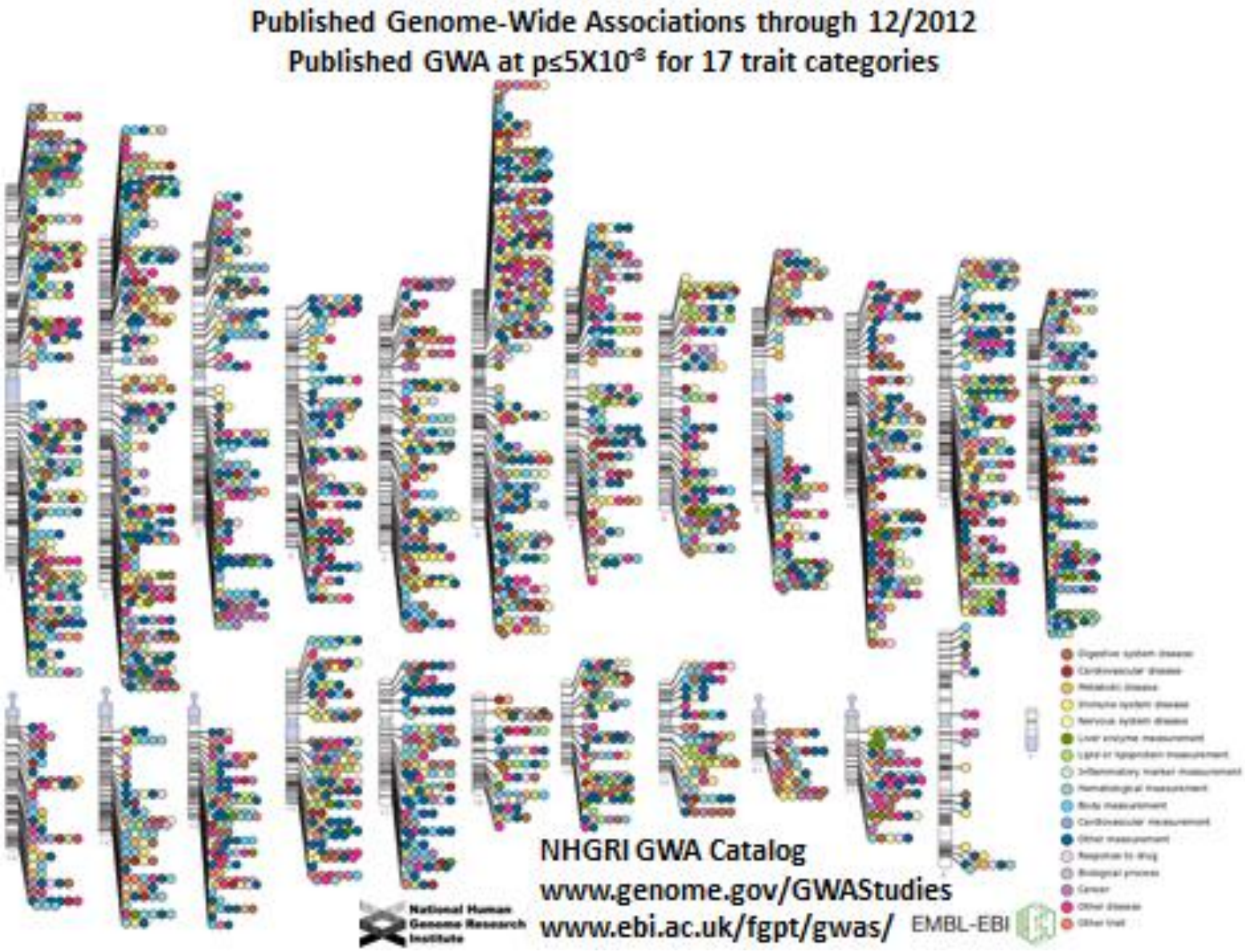
Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another

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.

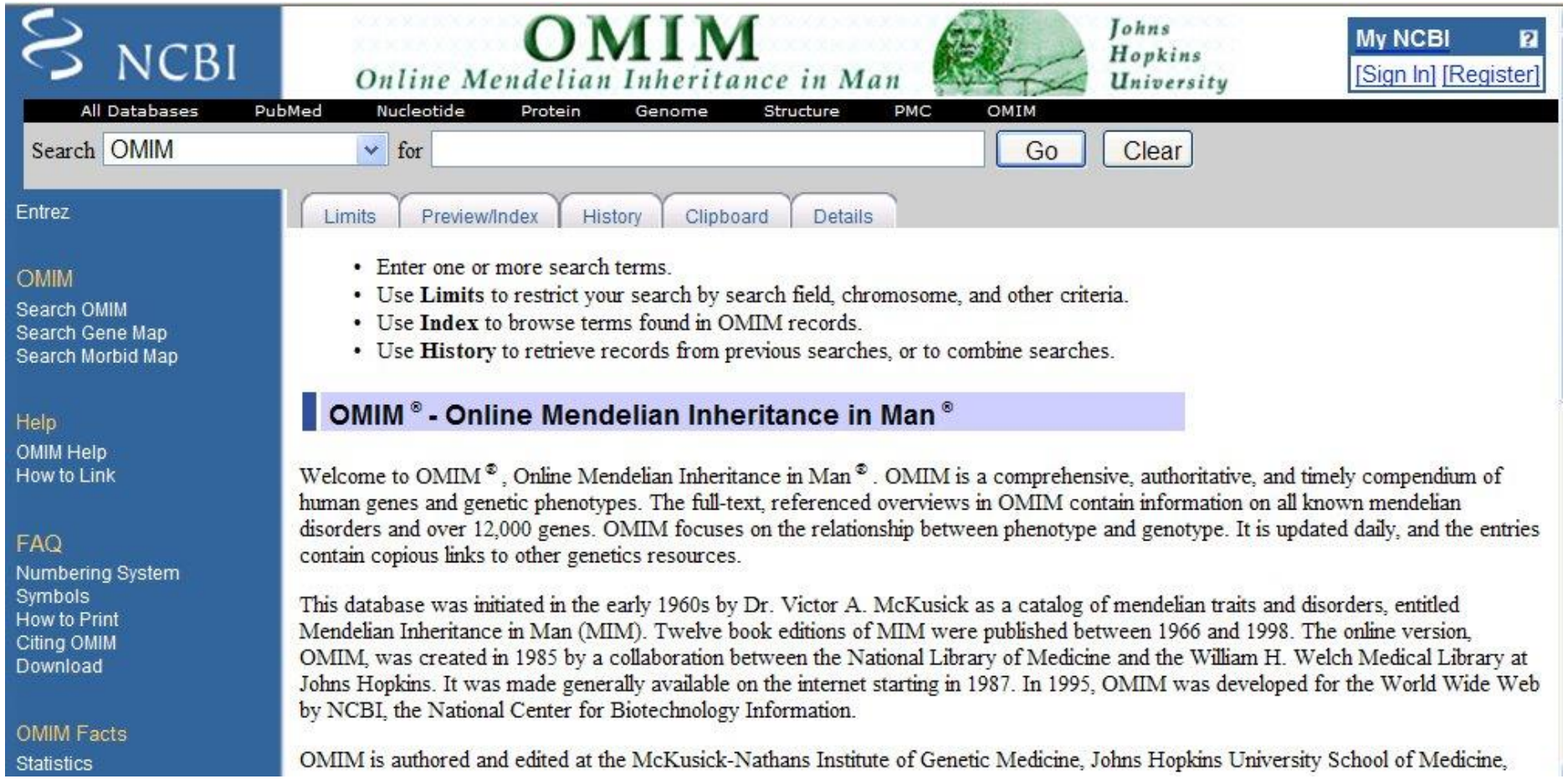

Published Genome-Wide Associations through 12/2012
Published GWA at p≤5X10-8 for 17 trait categories
NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/
National Human Genome Research Institute
EMBL-EBI

The future of GWAs?

(http://www.genome.gov/multimedia/illustrations/Published_GWA_Reports_6-2012.pdf)

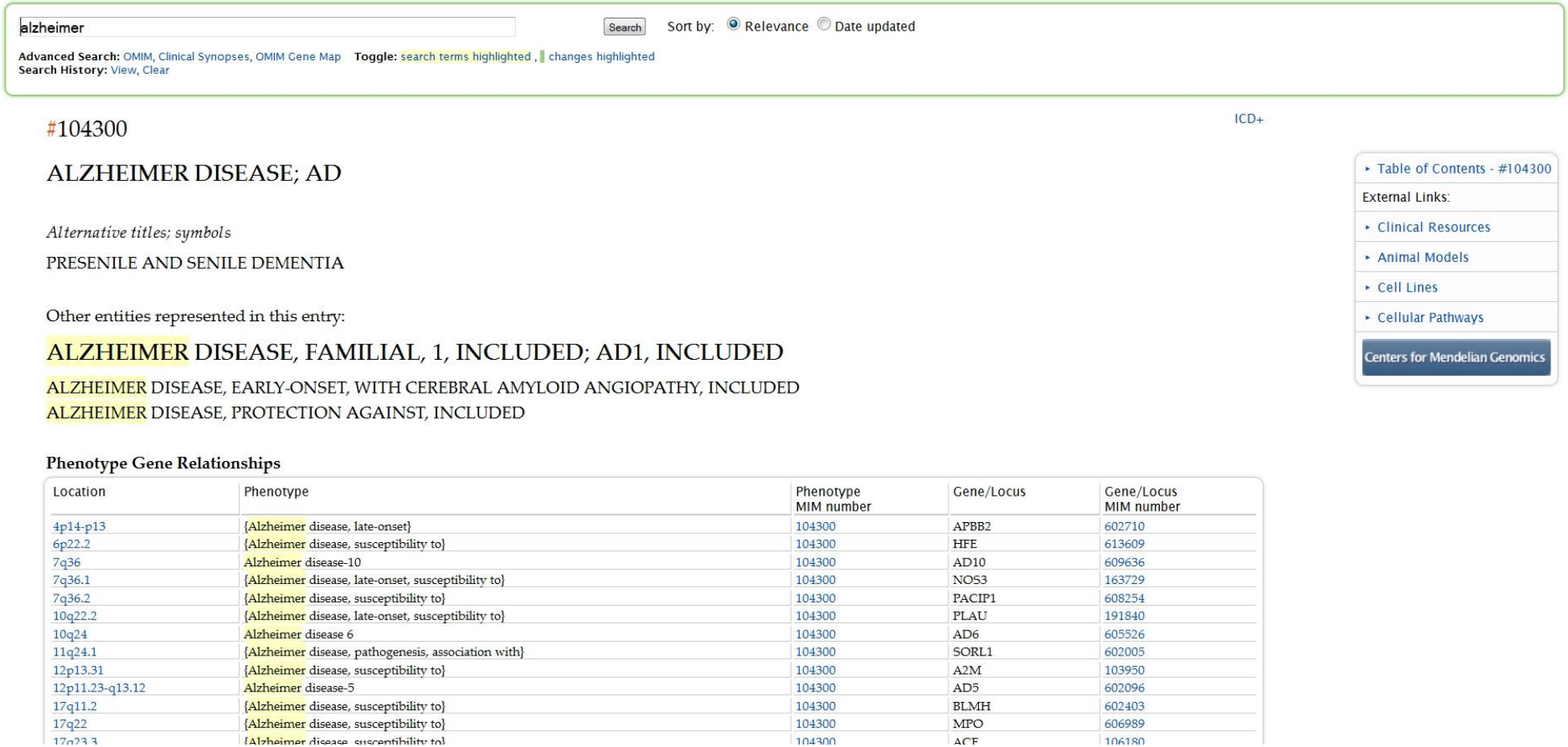
The pace of the molecular dissection of human disease can be measured by looking at the catalog of human genes and genetic disorders identified so far in *OMIM*, which is updated daily (www.ncbi.nlm.nih.gov/omim).

(V. A. McKusick, Mendelian Inheritance in Man (Johns Hopkins Univ. Press, Baltimore, ed. 12, 1998))


NCBI
OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC OMIM
Search OMIM for Go Clear
Entrez
Limits Preview/Index History Clipboard Details
OMIM
Search OMIM
Search Gene Map
Search Morbid Map
Help
OMIM Help
How to Link
FAQ
Numbering System
Symbols
How to Print
Citing OMIM
Download
OMIM Facts
Statistics
• Enter one or more search terms.
• Use Limits to restrict your search by search field, chromosome, and other criteria.
• Use Index to browse terms found in OMIM records.
• Use History to retrieve records from previous searches, or to combine searches.
OMIM ® - Online Mendelian Inheritance in Man ®
Welcome to OMIM ®, Online Mendelian Inheritance in Man ®. OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources.
This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the internet starting in 1987. In 1995, OMIM was developed for the World Wide Web by NCBI, the National Center for Biotechnology Information.
OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine,

What is OMIM?

- Online Mendelian Inheritance in Man (OMIM®) is a continuously updated catalog of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression.
- It is thus 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



alzheimer Search Sort by: Relevance Date updated

Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map Toggle: search terms highlighted , changes highlighted
Search History: View, Clear

#104300

ICD+

ALZHEIMER DISEASE; AD

- Table of Contents - #104300

External Links:

- Clinical Resources
- Animal Models
- Cell Lines
- Cellular Pathways

Centers for Mendelian Genomics

Alternative titles; symbols

PRESENILE AND SENILE DEMENTIA

Other entities represented in this entry:

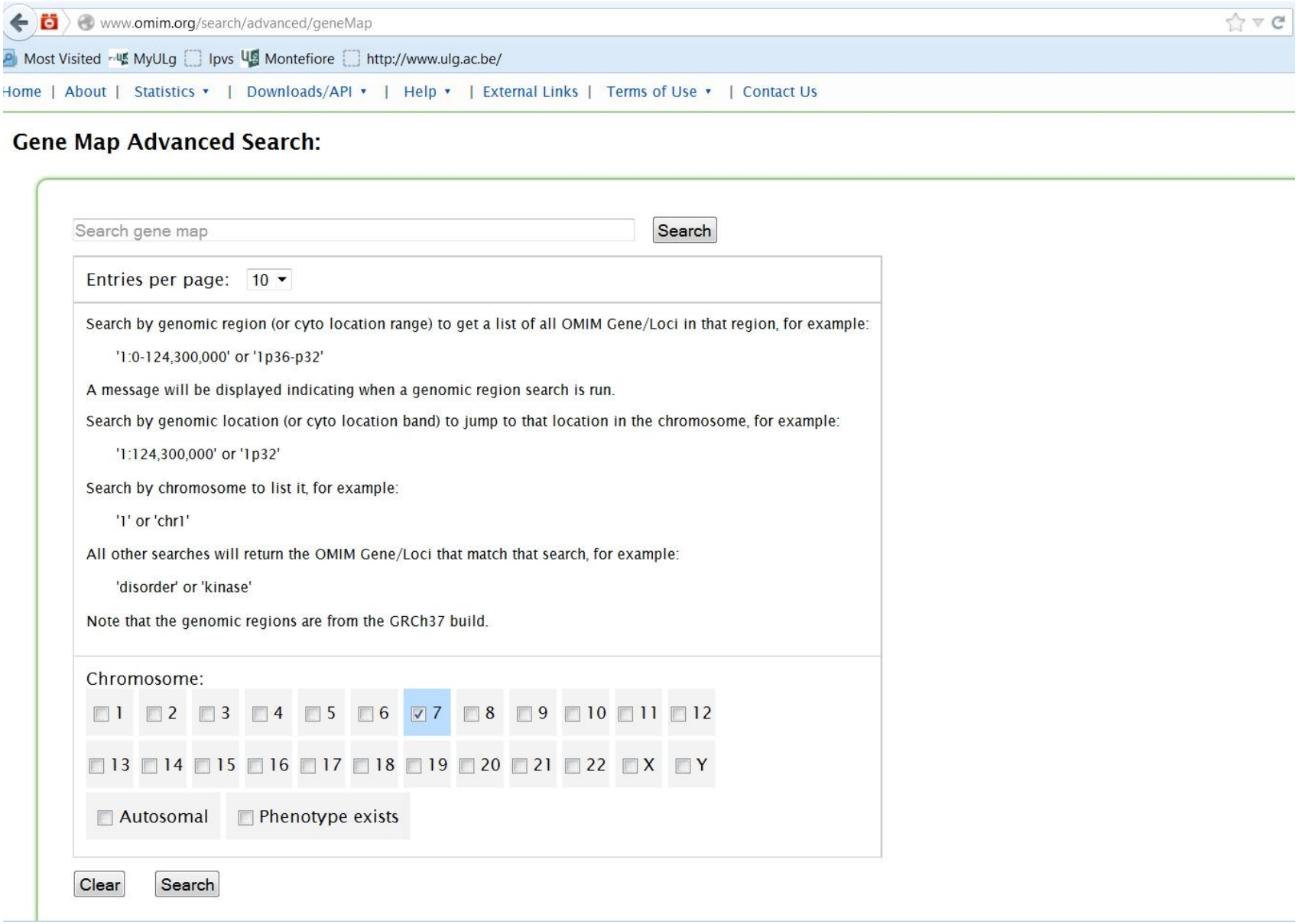
ALZHEIMER DISEASE, FAMILIAL, 1, INCLUDED; AD1, INCLUDED

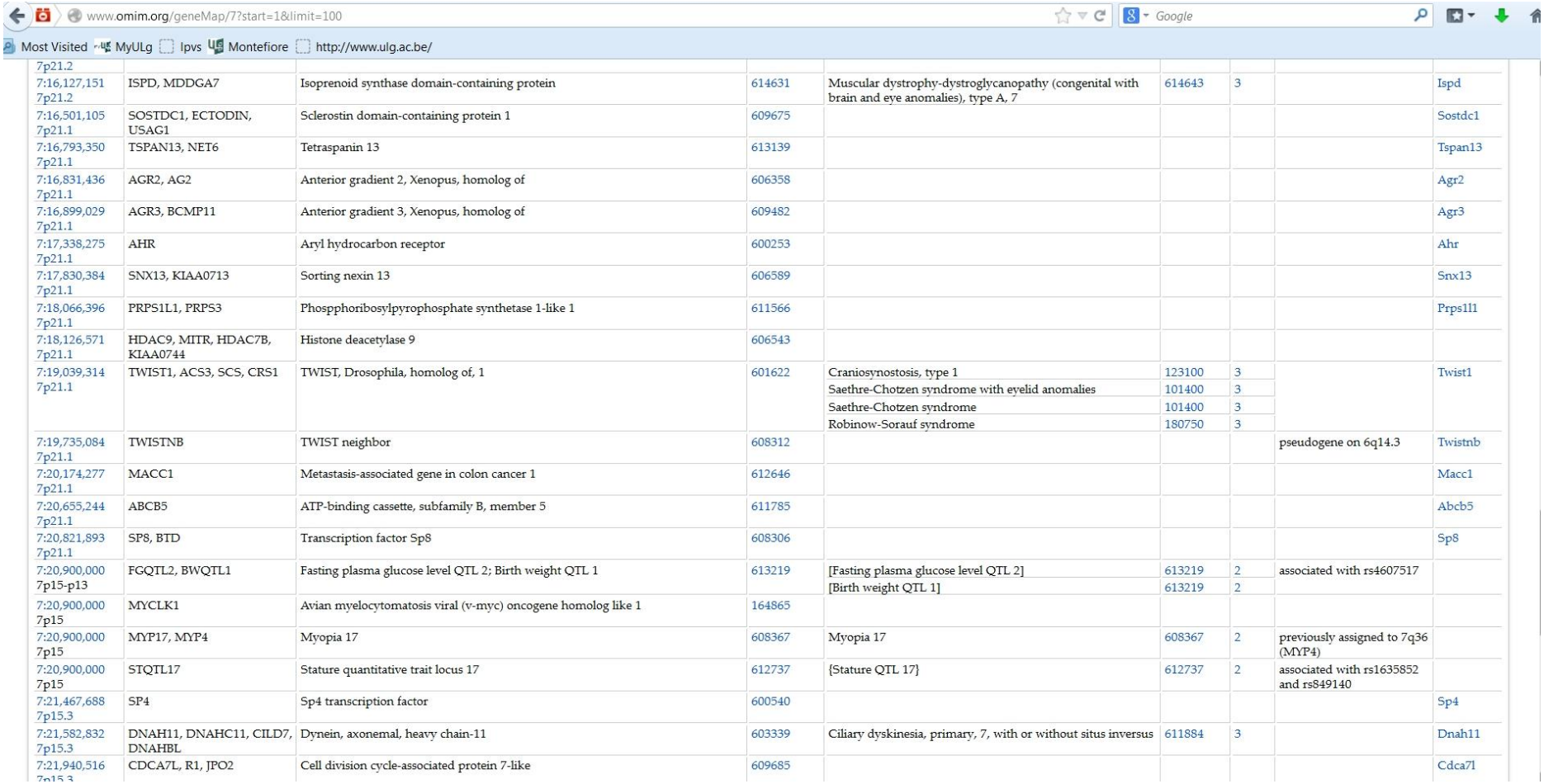
ALZHEIMER DISEASE, EARLY-ONSET, WITH CEREBRAL AMYLOID ANGIOPATHY, INCLUDED
ALZHEIMER DISEASE, PROTECTION AGAINST, INCLUDED

Phenotype Gene Relationships

Location	Phenotype	Phenotype MIM number	Gene/Locus	Gene/Locus MIM number
4p14-p13	{Alzheimer disease, late-onset}	104300	APBB2	602710
6p22.2	{Alzheimer disease, susceptibility to}	104300	HFE	613609
7q36	Alzheimer disease-10	104300	AD10	609636
7q36.1	{Alzheimer disease, late-onset, susceptibility to}	104300	NOS3	163729
7q36.2	{Alzheimer disease, susceptibility to}	104300	PACIP1	608254
10q22.2	{Alzheimer disease, late-onset, susceptibility to}	104300	PLAU	191840
10q24	Alzheimer disease 6	104300	AD6	605526
11q24.1	{Alzheimer disease, pathogenesis, association with}	104300	SORL1	602005
12p13.31	{Alzheimer disease, susceptibility to}	104300	A2M	103950
12p11.23-q13.12	Alzheimer disease-5	104300	AD5	602096
17q11.2	{Alzheimer disease, susceptibility to}	104300	BLMH	602403
17q22	{Alzheimer disease, susceptibility to}	104300	MPO	606989
17q23.3	{Alzheimer disease, susceptibility to}	104300	ACE	106180

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


www.omim.org/search/advanced/geneMap
Most Visited MyULg Ipvs Montefiore http://www.ulg.ac.be/
Home | About | Statistics | Downloads/API | Help | External Links | Terms of Use | Contact Us
Gene Map Advanced Search:
Search gene map
Search
Entries per page: 10
Search by genomic region (or cyto location range) to get a list of all OMIM Gene/Loci in that region, for example:
'1:0-124,300,000' or '1p36-p32'
A message will be displayed indicating when a genomic region search is run.
Search by genomic location (or cyto location band) to jump to that location in the chromosome, for example:
'1:124,300,000' or '1p32'
Search by chromosome to list it, for example:
'1' or 'chr1'
All other searches will return the OMIM Gene/Loci that match that search, for example:
'disorder' or 'kinase'
Note that the genomic regions are from the GRCh37 build.
Chromosome:
1 2 3 4 5 6 7 8 9 10 11 12
13 14 15 16 17 18 19 20 21 22 X Y
Autosomal
Phenotype exists
Clear
Search



www.omim.org/geneMap/7?start=1&limit=100

Most Visited MyULg Ipvs Montefiore http://www.ulg.ac.be/

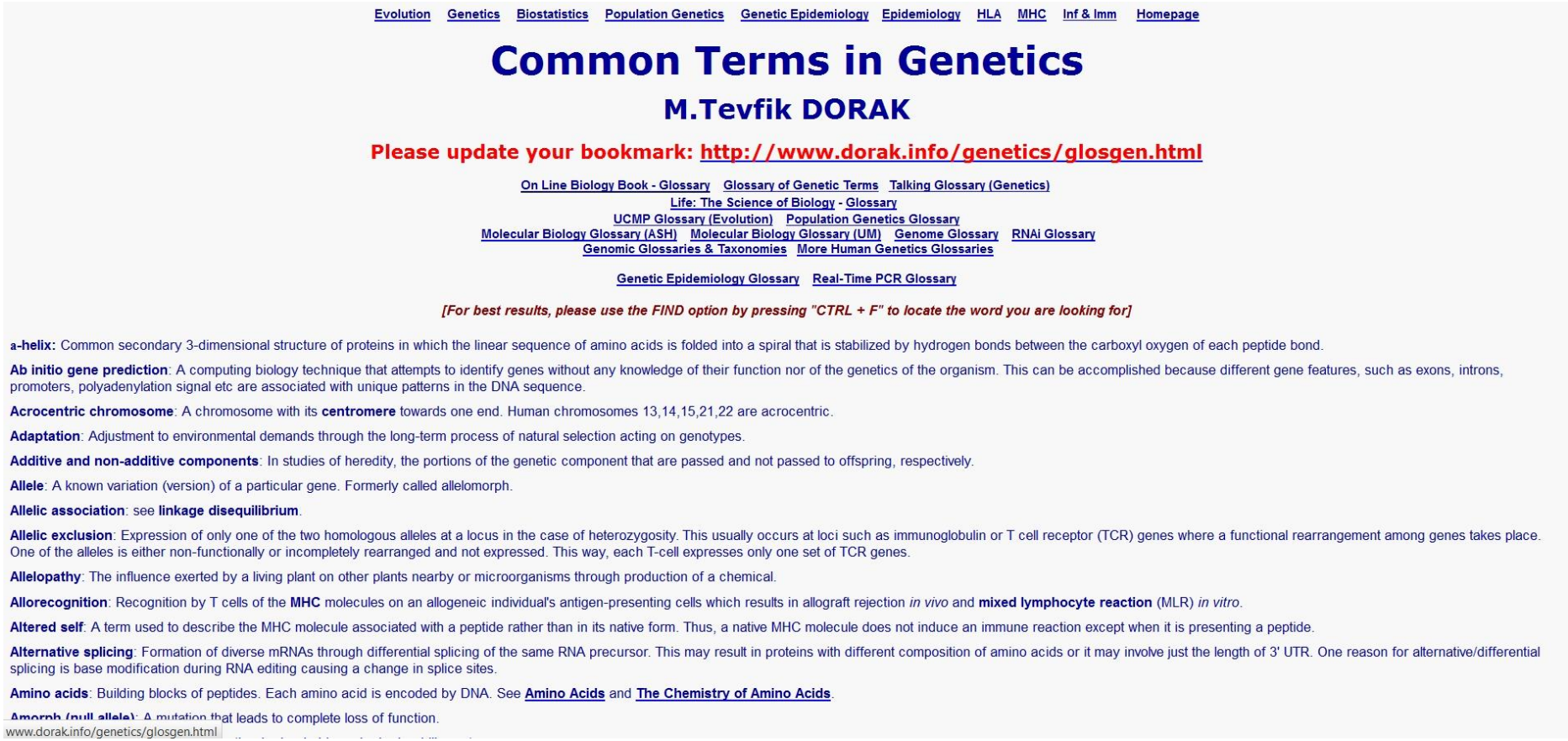
7p21.2								
7:16,127,151 7p21.2	ISPD, MDDGA7	Isoprenoid synthase domain-containing protein	614631	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 7	614643	3		Ispd
7:16,501,105 7p21.1	SOSTDC1, ECTODIN, USAG1	Sclerostin domain-containing protein 1	609675					Sostdc1
7:16,793,350 7p21.1	TSPAN13, NET6	Tetraspanin 13	613139					Tspan13
7:16,831,436 7p21.1	AGR2, AG2	Anterior gradient 2, Xenopus, homolog of	606358					Agr2
7:16,899,029 7p21.1	AGR3, BCMP11	Anterior gradient 3, Xenopus, homolog of	609482					Agr3
7:17,338,275 7p21.1	AHR	Aryl hydrocarbon receptor	600253					Ahr
7:17,830,384 7p21.1	SNX13, KIAA0713	Sorting nexin 13	606589					Snx13
7:18,066,396 7p21.1	PRPS1L1, PRPS3	Phospphoribosylpyrophosphate synthetase 1-like 1	611566					Prps1l1
7:18,126,571 7p21.1	HDAC9, MITR, HDAC7B, KIAA0744	Histone deacetylase 9	606543					
7:19,039,314 7p21.1	TWIST1, ACS3, SCS, CRS1	TWIST, Drosophila, homolog of, 1	601622	Craniosynostosis, type 1 Saethre-Chotzen syndrome with eyelid anomalies Saethre-Chotzen syndrome Robinow-Sorauf syndrome	123100 101400 101400 180750	3 3 3 3		Twist1
7:19,735,084 7p21.1	TWISTNB	TWIST neighbor	608312				pseudogene on 6q14.3	Twistnb
7:20,174,277 7p21.1	MACC1	Metastasis-associated gene in colon cancer 1	612646					Macc1
7:20,655,244 7p21.1	ABCB5	ATP-binding cassette, subfamily B, member 5	611785					Abcb5
7:20,821,893 7p21.1	SP8, BTD	Transcription factor Sp8	608306					Sp8
7:20,900,000 7p15-p13	FGQTL2, BWQTL1	Fasting plasma glucose level QTL 2; Birth weight QTL 1	613219	[Fasting plasma glucose level QTL 2] [Birth weight QTL 1]	613219 613219	2 2	associated with rs4607517	
7:20,900,000 7p15	MYCLK1	Avian myelocytomatosis viral (v-myc) oncogene homolog like 1	164865					
7:20,900,000 7p15	MYP17, MYP4	Myopia 17	608367	Myopia 17	608367	2	previously assigned to 7q36 (MYP4)	
7:20,900,000 7p15	STQTL17	Stature quantitative trait locus 17	612737	{Stature QTL 17}	612737	2	associated with rs1635852 and rs849140	
7:21,467,688 7p15.3	SP4	Sp4 transcription factor	600540					Sp4
7:21,582,832 7p15.3	DNAH11, DNAHC11, CILD7, DNAHBL	Dynein, axonemal, heavy chain-11	603339	Ciliary dyskinesia, primary, 7, with or without situs inversus	611884	3		Dnah11
7:21,940,516 7p15.3	CDCA7L, R1, JPO2	Cell division cycle-associated protein 7-like	609685					Cdca7l

(http://www.omim.org/) on the map using statistics

(2) or molecular basis of the disorder is known (3)

1.b Speaking the language

What is ….?



Evolution Genetics Biostatistics Population Genetics Genetic Epidemiology Epidemiology HLA MHC Inf & Imm Homepage

Common Terms in Genetics

M.Tevfik DORAK

Please update your bookmark: http://www.dorak.info/genetics/glosgen.html

On Line Biology Book - Glossary Glossary of Genetic Terms Talking Glossary (Genetics)
Life: The Science of Biology - Glossary
UCMP Glossary (Evolution) Population Genetics Glossary
Molecular Biology Glossary (ASH) Molecular Biology Glossary (UM) Genome Glossary RNAi Glossary
Genomic Glossaries & Taxonomies More Human Genetics Glossaries

Genetic Epidemiology Glossary Real-Time PCR Glossary

[For best results, please use the FIND option by pressing "CTRL + F" to locate the word you are looking for]

a-helix: Common secondary 3-dimensional structure of proteins in which the linear sequence of amino acids is folded into a spiral that is stabilized by hydrogen bonds between the carboxyl oxygen of each peptide bond.

Ab initio gene prediction: A computing biology technique that attempts to identify genes without any knowledge of their function nor of the genetics of the organism. This can be accomplished because different gene features, such as exons, introns, promoters, polyadenylation signal etc are associated with unique patterns in the DNA sequence.

Acrocentric chromosome: A chromosome with its **centromere** towards one end. Human chromosomes 13,14,15,21,22 are acrocentric.

Adaptation: Adjustment to environmental demands through the long-term process of natural selection acting on genotypes.

Additive and non-additive components: In studies of heredity, the portions of the genetic component that are passed and not passed to offspring, respectively.

Allele: A known variation (version) of a particular gene. Formerly called allelomorph.

Allelic association: see **linkage disequilibrium**.

Allelic exclusion: Expression of only one of the two homologous alleles at a locus in the case of heterozygosity. This usually occurs at loci such as immunoglobulin or T cell receptor (TCR) genes where a functional rearrangement among genes takes place. One of the alleles is either non-functionally or incompletely rearranged and not expressed. This way, each T-cell expresses only one set of TCR genes.

Allelopathy: The influence exerted by a living plant on other plants nearby or microorganisms through production of a chemical.

Allorecognition: Recognition by T cells of the **MHC** molecules on an allogeneic individual's antigen-presenting cells which results in allograft rejection *in vivo* and **mixed lymphocyte reaction** (MLR) *in vitro*.

Altered self: A term used to describe the MHC molecule associated with a peptide rather than in its native form. Thus, a native MHC molecule does not induce an immune reaction except when it is presenting a peptide.

Alternative splicing: Formation of diverse mRNAs through differential splicing of the same RNA precursor. This may result in proteins with different composition of amino acids or it may involve just the length of 3' UTR. One reason for alternative/differential splicing is base modification during RNA editing causing a change in splice sites.

Amino acids: Building blocks of peptides. Each amino acid is encoded by DNA. See **Amino Acids** and **The Chemistry of Amino Acids**.

Amorph (null allele): A mutation that leads to complete loss of function.

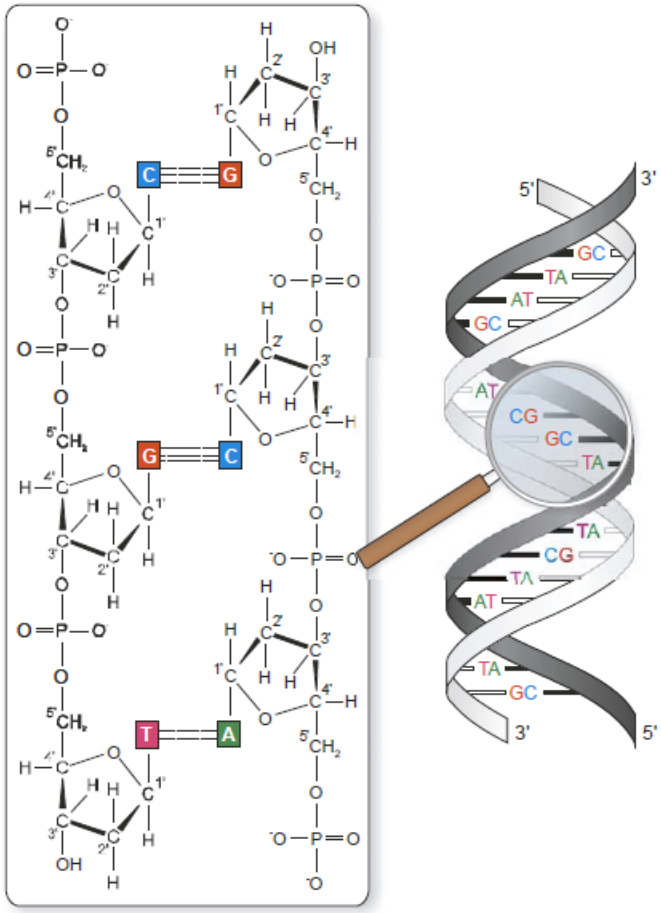
www.dorak.info/genetics/glosgen.html

Where is the genetic information located?

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

(Ziegler and Van Steen, Brazil 2010)

Where is the genetic information located?



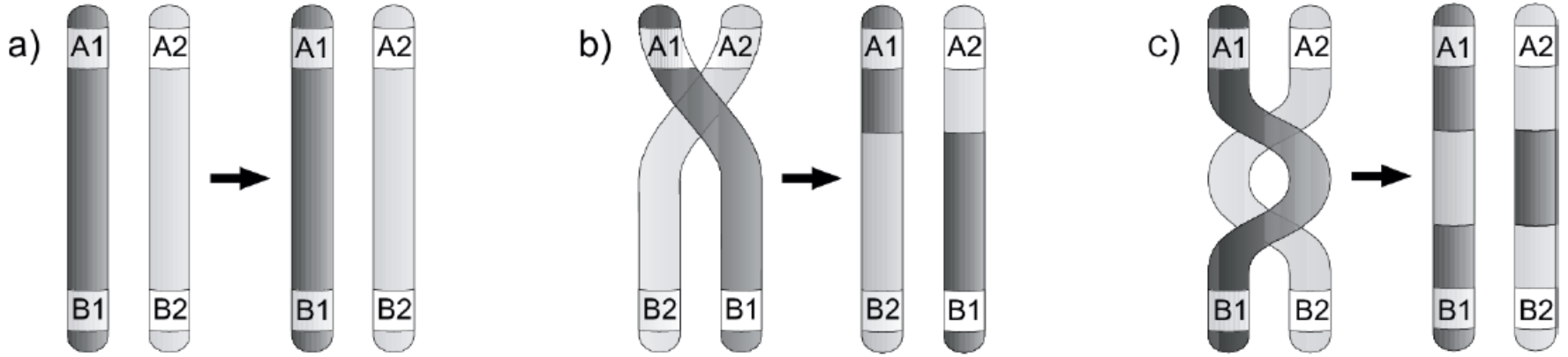
(Ziegler and Van Steen, Brazil 2010)

Where is the genetic information located?

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

(Ziegler and Van Steen, Brazil 2010)

What is 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)

How much do individuals differ with respect to genetic information?

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

(Ziegler and Van Steen, Brazil 2010)

How much do individuals differ with respect to genetic information?

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

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

Are haplotypes always better in association studies for "disease"?

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

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

Test 1 vs. 2 for M1: D + d vs. d
Test 1 vs. 2 for M2: D + d vs. d
Test haplotype H1 vs. all others: D vs. d

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

What are microsatellite markers?

- 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)

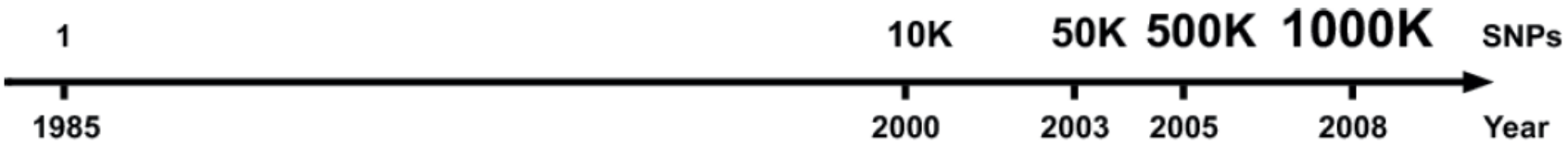
What are single nucleotide polymorphisms?

- 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)

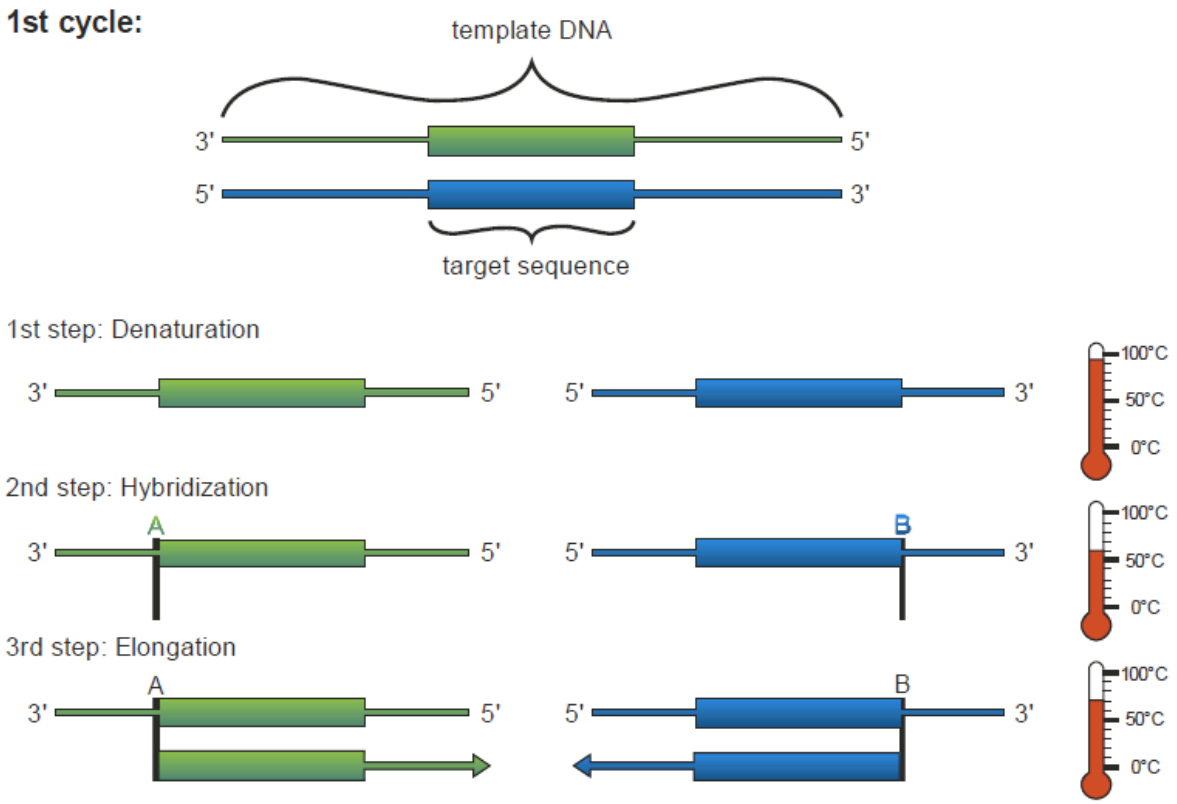
Why are SNPs preferred over STRs?

- 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)

Recall



(Ziegler and Van Steen, Brazil 2010)

Which genotyping methods are currently being used?

Method	Principle	Thru-put
Allele-specific PCR	1 common reverse primer, 2 forward allele-specific primers with different tails, amplification of two allele-specific PCR products of different lengths, separation by gel electrophoresis	Low
RFLP analysis	DNA sample digested by restriction enzymes, resulting restriction fragments separated according to their lengths by gel electrophoresis	Low
Pyrosequencing	Single strand sequencing, enzymatic synthesizing of complementary strand	Middle
SNPstream	Single-base primer extension technology	Middle / High

(Ziegler and Van Steen, Brazil 2010)

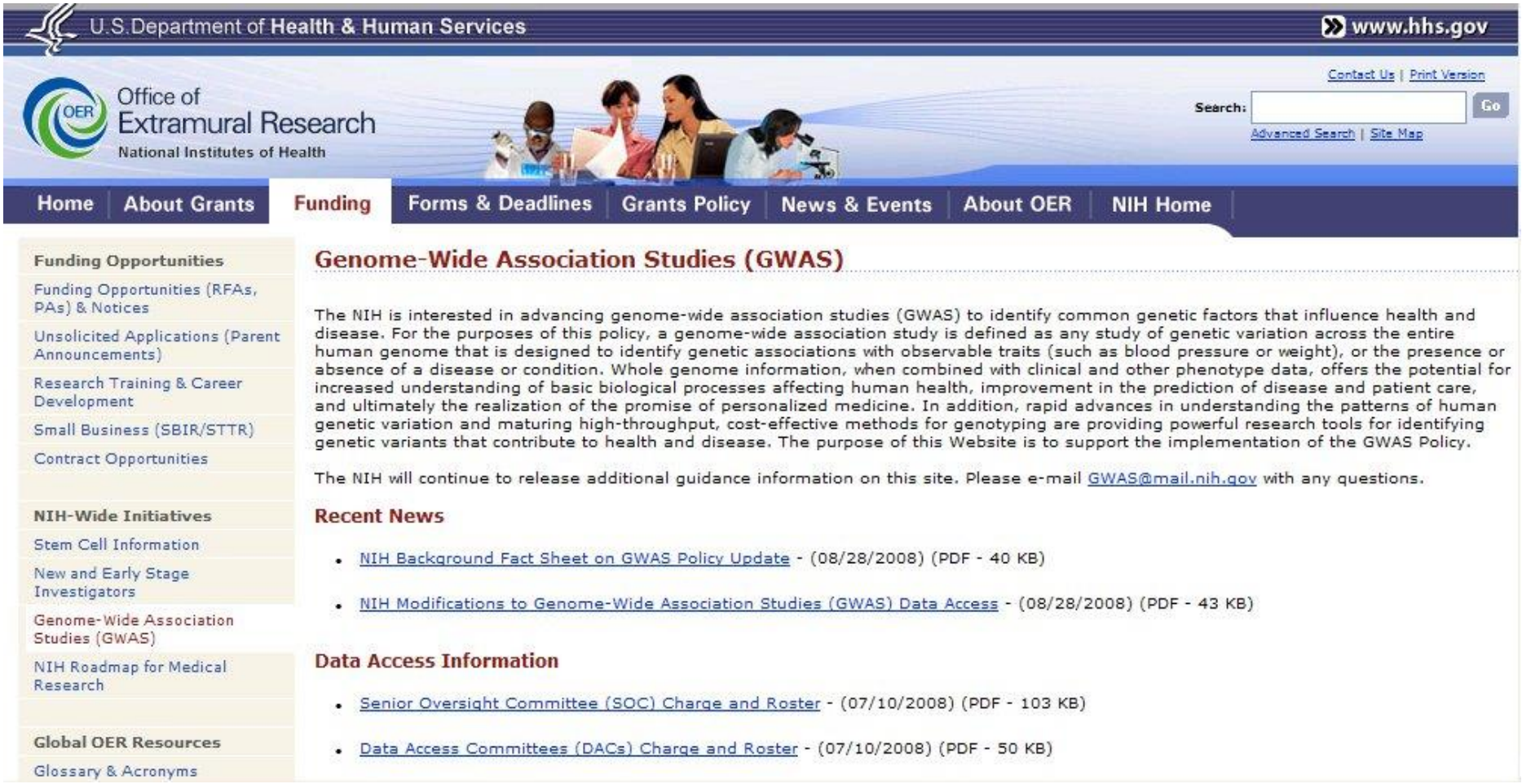
Which genotyping methods are currently being used?

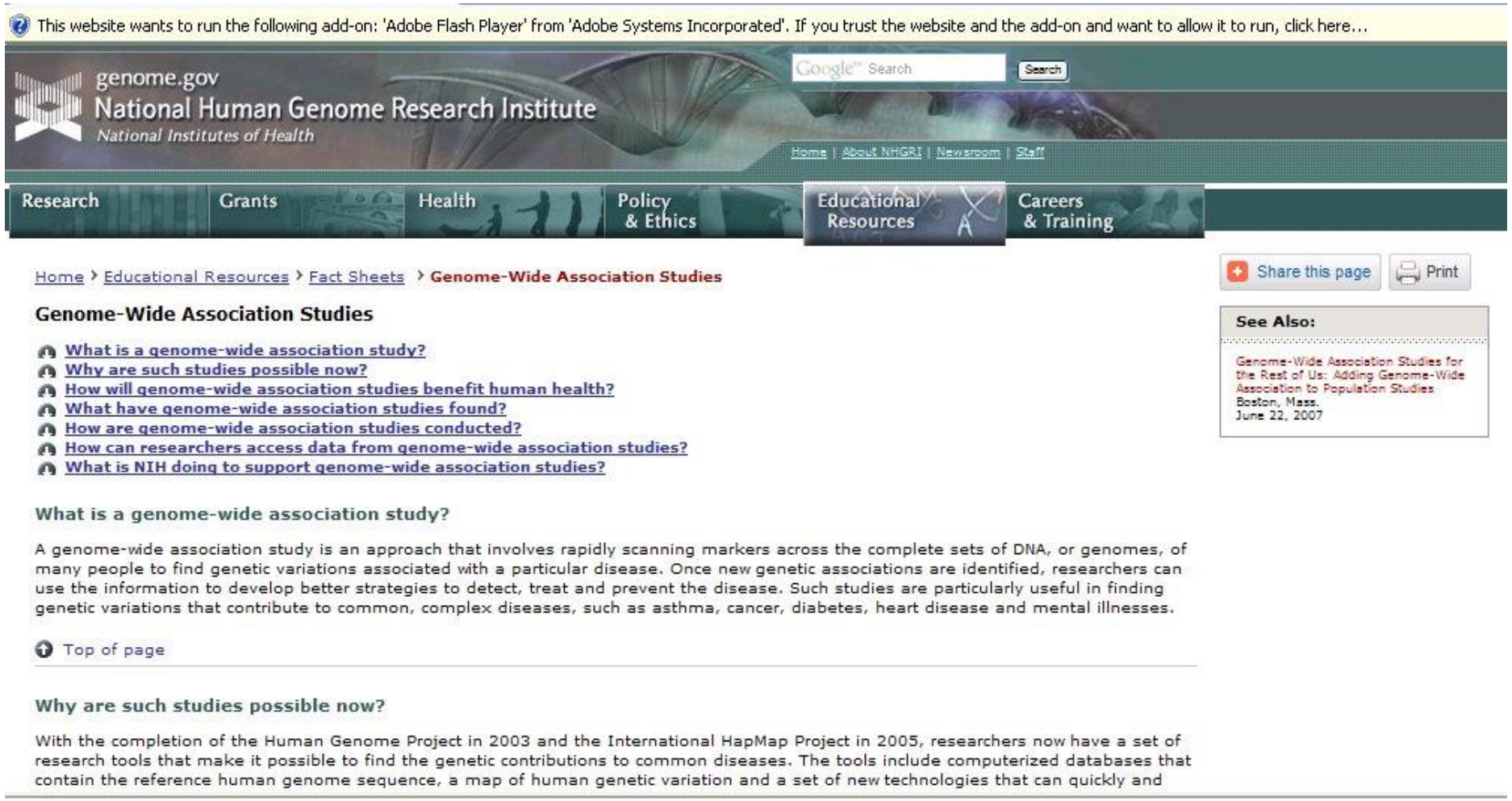
Method	Principle	Thru-put
TaqMan	Quantitative real-time PCR, allele-specific TaqMan probes	Middle
SNPlex	Oligonucleotide ligation/PCR and capillary electrophoresis	Middle
Affymetrix	Microarray based, fluorescence labeled DNA	Ultra-high
Illumina	Microarray based, fluorescence labeled DNA	Ultra-high

(Ziegler and Van Steen, Brazil 2010)

1.c Genome-wide association studies

- Note: From –etic to –omic: scale explosion
- A genome-wide association study refers to a method / methodology for interrogating all 10 million variable points across the human genome.
- Since variation is inherited in groups, or blocks, not all 10 million points have to be tested.
- Blocks are shorter though (so need for testing more points) the less closely people are related.


U.S. Department of Health & Human Services
www.hhs.gov
Office of Extramural Research
National Institutes of Health
Contact Us | Print Version
Search:
Go
Advanced Search | Site Map
Home
About Grants
Funding
Forms & Deadlines
Grants Policy
News & Events
About OER
NIH Home
Funding Opportunities
Funding Opportunities (RFAs, PAs) & Notices
Unsolicited Applications (Parent Announcements)
Research Training & Career Development
Small Business (SBIR/STTR)
Contract Opportunities
NIH-Wide Initiatives
Stem Cell Information
New and Early Stage Investigators
Genome-Wide Association Studies (GWAS)
NIH Roadmap for Medical Research
Global OER Resources
Glossary & Acronyms
Genome-Wide Association Studies (GWAS)
The NIH is interested in advancing genome-wide association studies (GWAS) to identify common genetic factors that influence health and disease. For the purposes of this policy, a genome-wide association study is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition. Whole genome information, when combined with clinical and other phenotype data, offers the potential for increased understanding of basic biological processes affecting human health, improvement in the prediction of disease and patient care, and ultimately the realization of the promise of personalized medicine. In addition, rapid advances in understanding the patterns of human genetic variation and maturing high-throughput, cost-effective methods for genotyping are providing powerful research tools for identifying genetic variants that contribute to health and disease. The purpose of this Website is to support the implementation of the GWAS Policy.
The NIH will continue to release additional guidance information on this site. Please e-mail GWAS@mail.nih.gov with any questions.
Recent News
NIH Background Fact Sheet on GWAS Policy Update - (08/28/2008) (PDF - 40 KB)
NIH Modifications to Genome-Wide Association Studies (GWAS) Data Access - (08/28/2008) (PDF - 43 KB)
Data Access Information
Senior Oversight Committee (SOC) Charge and Roster - (07/10/2008) (PDF - 103 KB)
Data Access Committees (DACs) Charge and Roster - (07/10/2008) (PDF - 50 KB)


This website wants to run the following add-on: 'Adobe Flash Player' from 'Adobe Systems Incorporated'. If you trust the website and the add-on and want to allow it to run, click here...
genome.gov
National Human Genome Research Institute
National Institutes of Health
Google Search
Search
Home | About NHGRI | Newsroom | Staff
Research
Grants
Health
Policy & Ethics
Educational Resources
Careers & Training
Home › Educational Resources › Fact Sheets › Genome-Wide Association Studies
Share this page
Print
Genome-Wide Association Studies
What is a genome-wide association study?
Why are such studies possible now?
How will genome-wide association studies benefit human health?
What have genome-wide association studies found?
How are genome-wide association studies conducted?
How can researchers access data from genome-wide association studies?
What is NIH doing to support genome-wide association studies?
See Also:
Genome-Wide Association Studies for the Rest of Us: Adding Genome-Wide Association to Population Studies
Boston, Mass.
June 22, 2007
What is a genome-wide association study?
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 disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.
Top of page
Why are such studies possible now?
With the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005, researchers now have a set of research tools that make it possible to find the genetic contributions to common diseases. 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

What is a genome-wide association study?

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

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

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

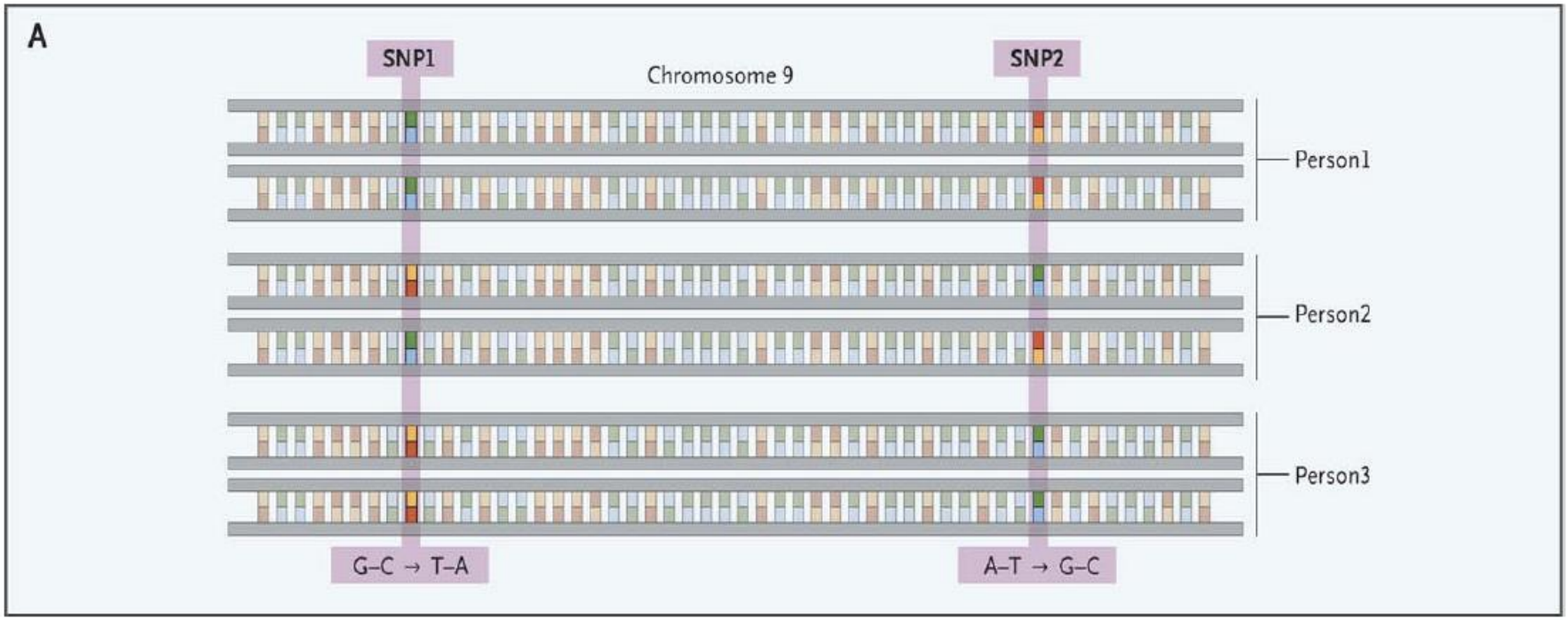
What do we need to carry out a genome-wide association study?

- 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.

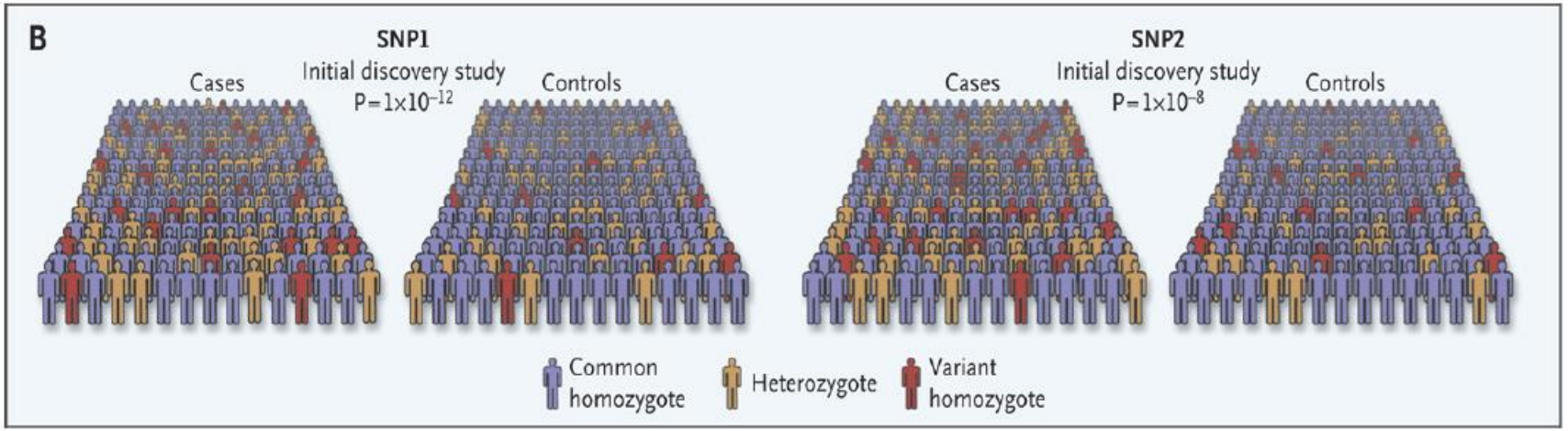
(http://www.genome.gov/pfv.cfm?pageID=20019523)

What is the flow of a genome-wide association study?

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)



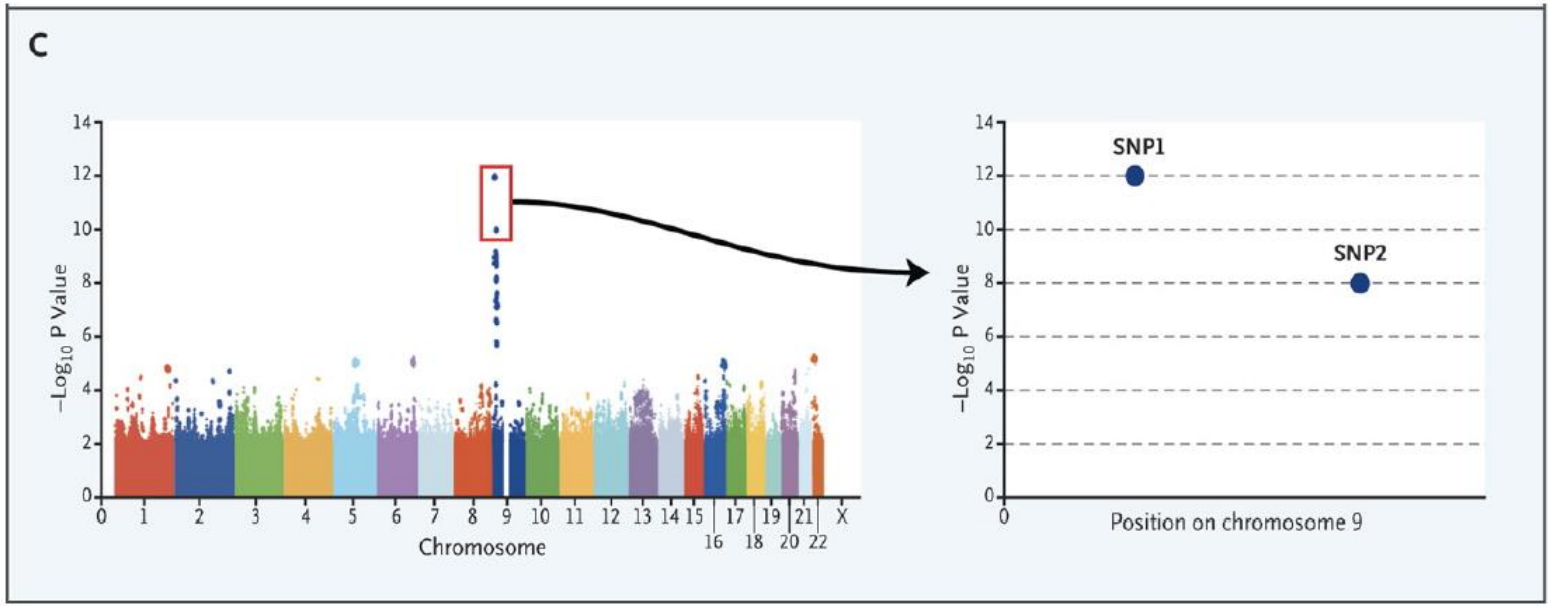
What is the flow of a genome-wide association study?



- 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

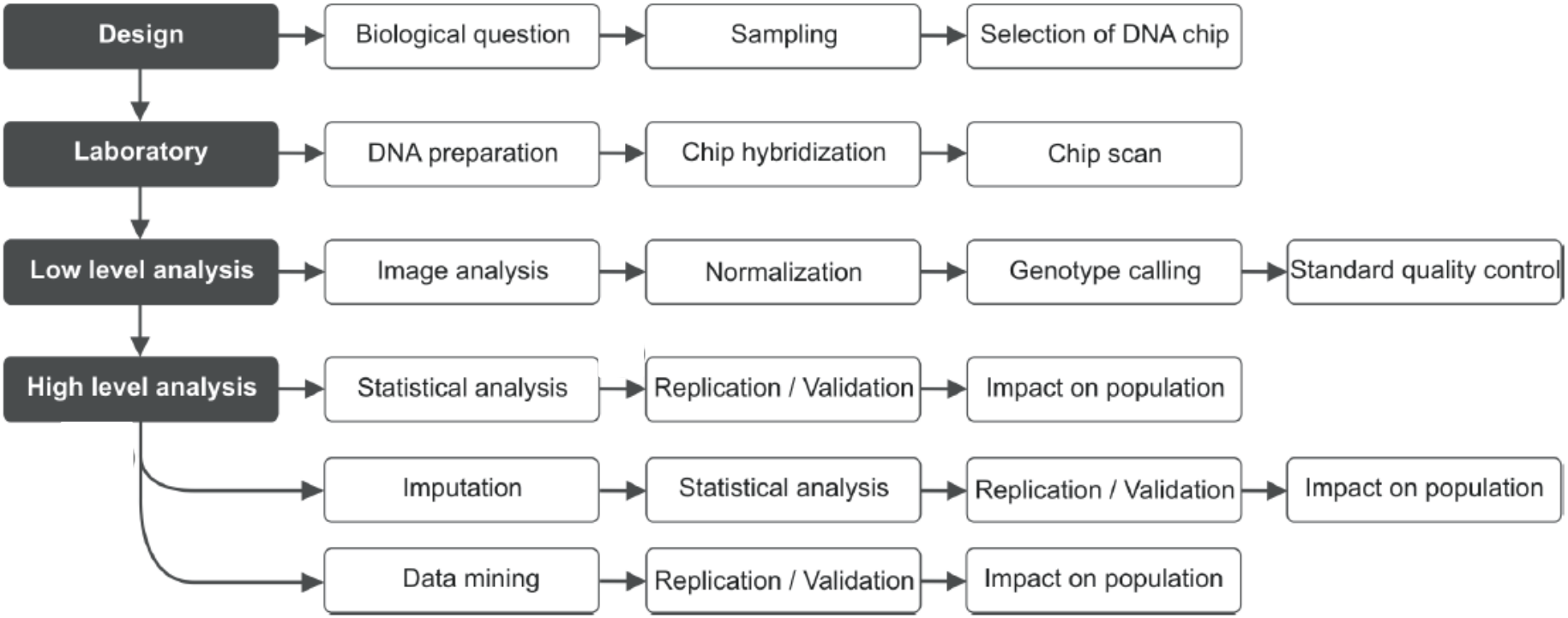
(Manolio 2010)

What is the flow of a genome-wide association study?



- 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)

What is the flow of a genome-wide association study?



(Ziegler 2009)

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

PERSPECTIVE

DRINKING FROM THE FIRE HOSE — STATISTICAL ISSUES IN GENOMEWIDE ASSOCIATION STUDIES

STATISTICS AND MEDICINE

Drinking from the Fire Hose — Statistical Issues in Genomewide Association Studies

David J. Hunter, M.B., B.S., and Peter Kraft, Ph.D.

Related article, page 443

The past 3 months have seen the publication of a series of studies examining the inherited genetic underpinnings of common diseases such as prostate cancer, breast cancer, diabetes, and in this issue of the *Journal*, coronary artery disease (reported by Samani et al., pages 443–453). These genomewide association studies have been able to examine interpatient differences in inherited genetic variability at an unprecedented level of resolution, thanks to the development of microarrays, or chips, capable of assaying the need for guessing which genes are likely to harbor variants affecting risk. Most of the robust associations seen in this type of study have not been with genes previously suspected of being related to the disease. Some of these associations have been found in regions not even known to harbor genes, such as the 8q24 region, in which multiple variants have been found to be associated with prostate cancer.[2] Such findings promise to open up new avenues of research, through both the discovery of new genes rele-

The main problem with this strategy is that, because of the high cost of SNP chips, most studies are somewhat constrained in terms of the number of samples and thus have limited power to generate P values as small as 10^{-7}. In addition, most variants identified recently have been associated with modest relative risks (e.g., 1.3 for heterozygotes and 1.6 for homozygotes), and many true associations are not likely to exceed P values as extreme as 10^{-7} in an initial study. On the other hand, a "statistically significant" finding

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

- To distinguish between true and chance effects, there are several routes to be taken:
  - Set **tight standards** for statistical significance
  - Only consider patterns of polymorphisms that could plausibly have been generated by causal genetic variants (**use** understanding of and **insights** into human genetic history or evolutionary processes such as recombination or mutation)
  - Adequately deal with distorting factors, including missing data and genotyping errors (**quality control** measures)

Are GWAs part of the Bioinformatics discipline?

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

Are GWAs part of the Bioinformatics discipline?

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

Are GWAs part of the Bioinformatics discipline?

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

Are GWAs part of the Bioinformatics discipline?

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

(Martin et al 2009)

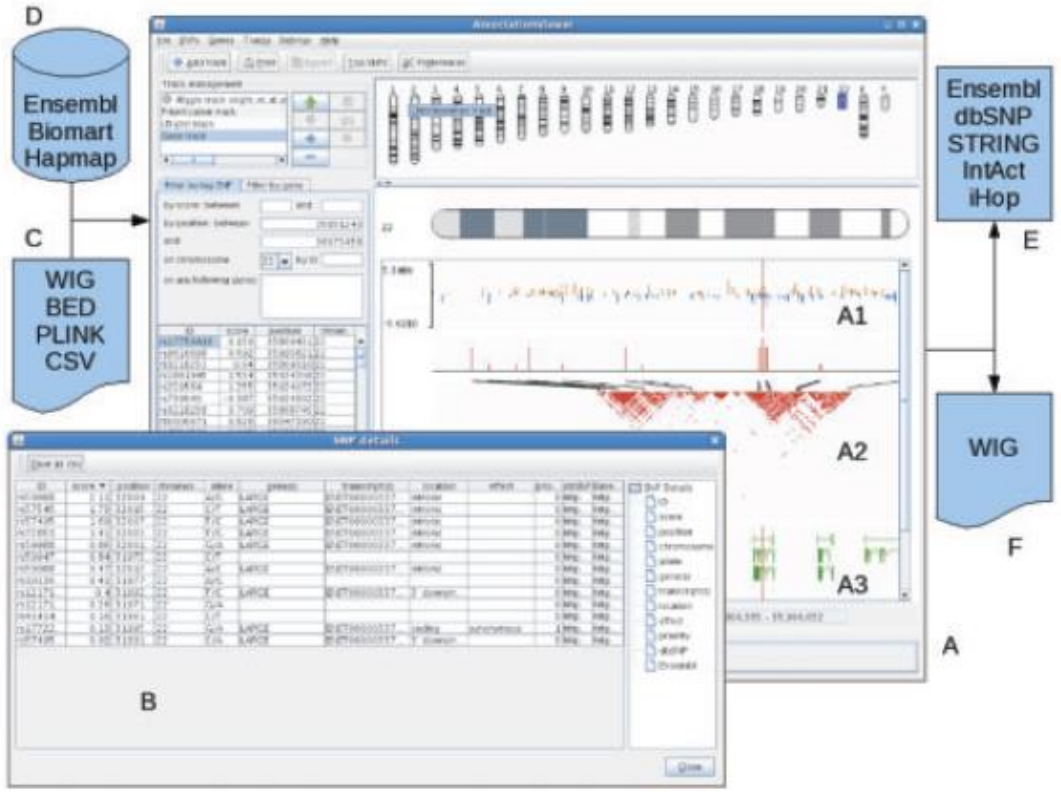


Fig. 1. General view of AssociationViewer (**A** and **B**). Also displayed are the input files (**C**), annotation data downloaded (**D**), cross-references (**E**) and export format (**F**).

2 GWA Details: Study Design

What are the components of a study design for GWA studies?

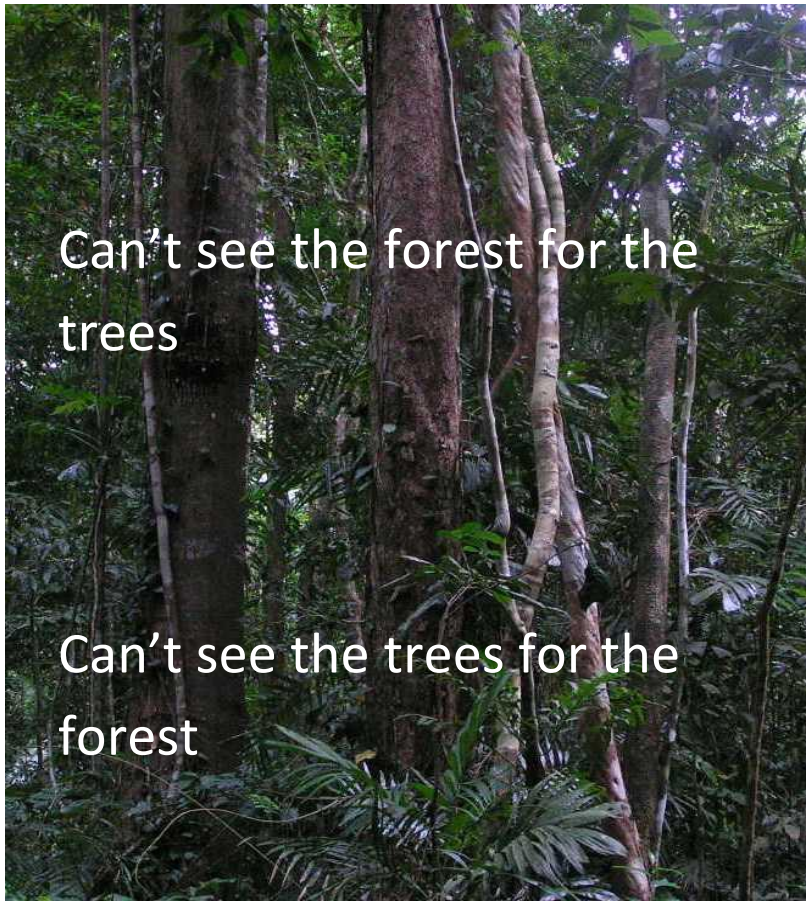
- The design of a genetic association study may refer to
  - study scale:
    - Genome-wide
    - Genomic
  - marker design:
    - Which markers are most informative? Microsatellites? SNPs? CNVs?
    - Which platform is the most promising?
  - subject design

Does scale matter?

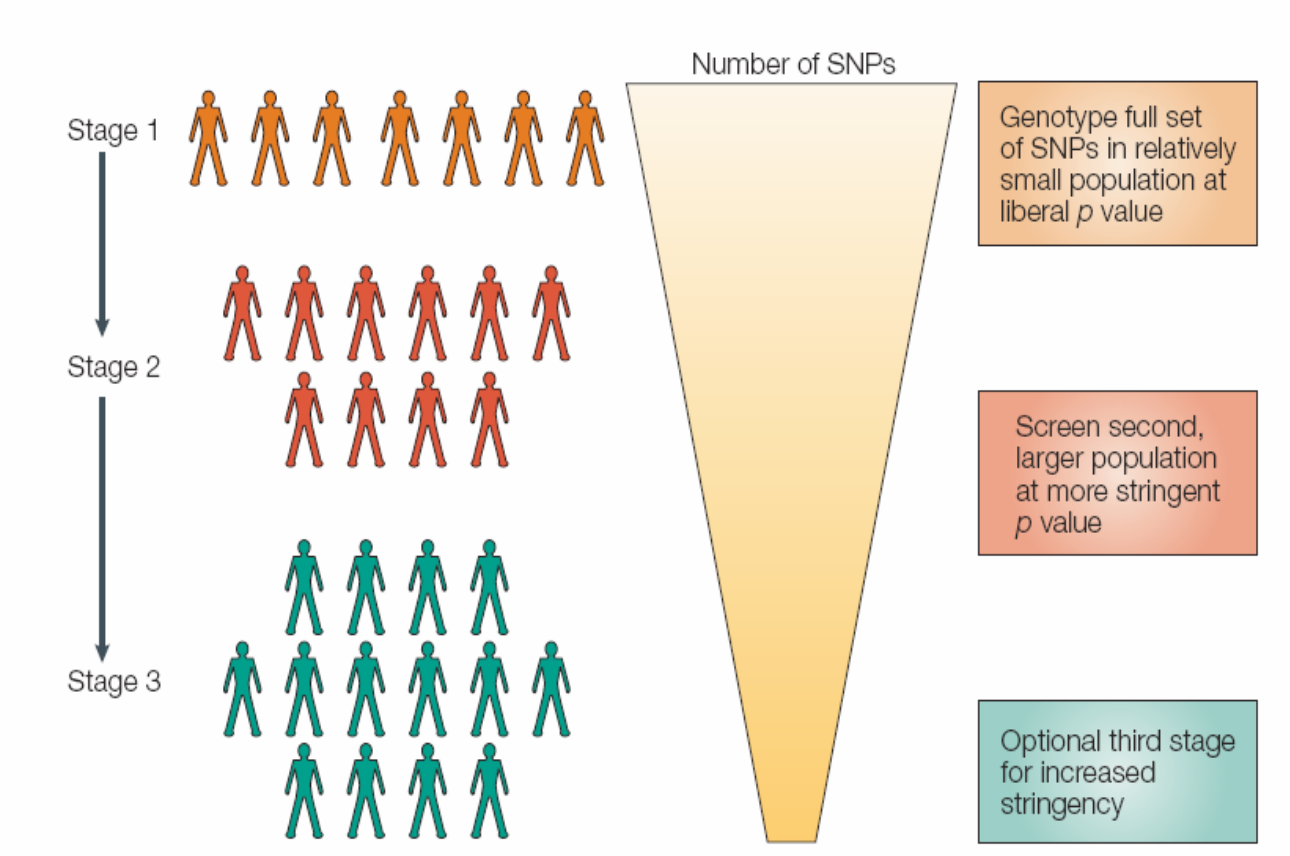
candidate gene approach

vs

genome-wide screening approach



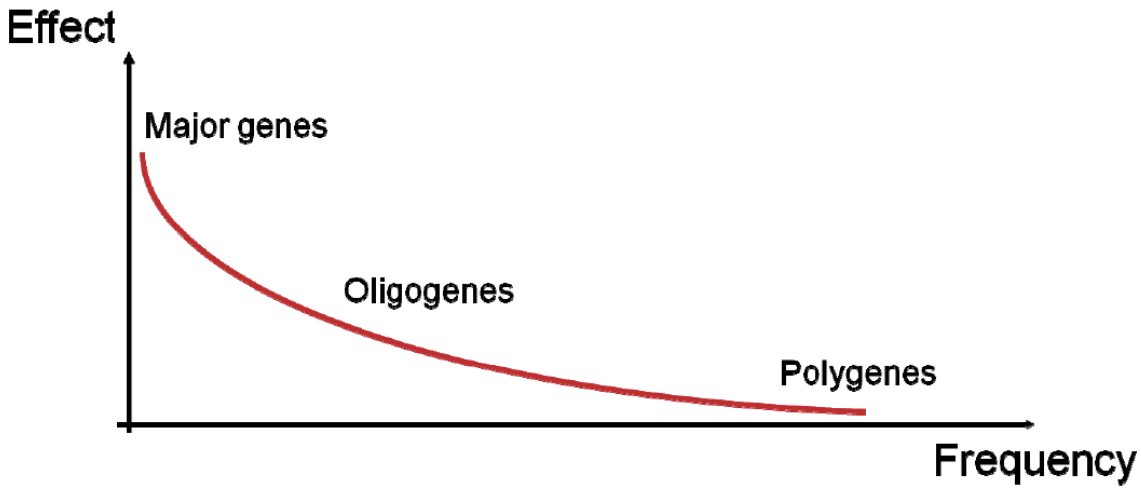
Does scale matter?



2.a Marker Level

Which genetic markers to select?

- Continuous distribution of genetic variants, shaped by mutation and selection
- The **Common Disease/Common Variant** hypothesis (CDCV)



(Ziegler and Van Steen, 2010)

Types of genetic diseases: Mendelian, oligogenic, polygenic

- **Monogenic diseases** are those in which defects in a single gene produce disease. Often these disease are severe and appear early in life, e.g., cystic fibrosis. For the population as a whole, they are relatively rare. In a sense, these are pure genetic diseases: They do not require any environmental factors to elicit them. Although nutrition is not involved in the causation of monogenic diseases, these diseases can have implications for nutrition. They reveal the effects of particular proteins or enzymes that also are influenced by nutritional factors

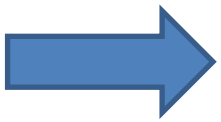
(http://www.utsouthwestern.edu)

- **Oligogenic diseases** are conditions produced by the combination of two, three, or four defective genes. Often a defect in one gene is not enough to elicit a full-blown disease; but when it occurs in the presence of other moderate defects, a disease becomes clinically manifest. It has long been the expectation of human geneticists that many chronic diseases can be explained by the combination of defects in a few (major) genes.

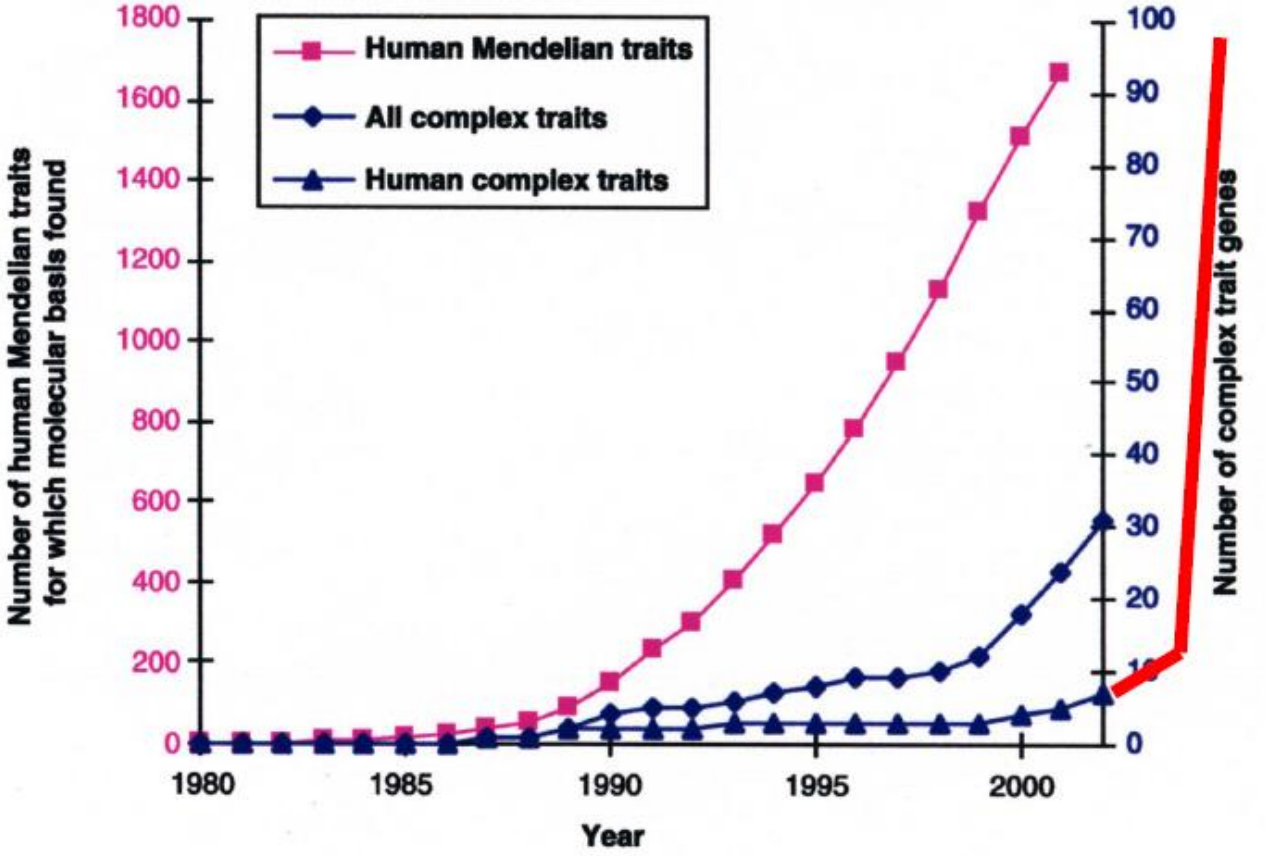
- A third category of genetic disorder is **polygenic disease**. According to the polygenic hypothesis, many mild defects in genes conspire to produce some chronic diseases. To date the full genetic basis of polygenic diseases has not been worked out; multiple interacting defects are highly complex !!!

(http://www.utsouthwestern.edu)

- **Complex diseases** refer to conditions caused by many contributing factors. Such a disease is also called a multifactorial disease.
  - Whereas some disorders, such as sickle cell anemia and cystic fibrosis, are caused by mutations in a single gene,
  - common medical problems such as heart disease, diabetes, and obesity likely associated with the effects of multiple genes in combination with lifestyle and environmental factors, all of them possibly interacting.



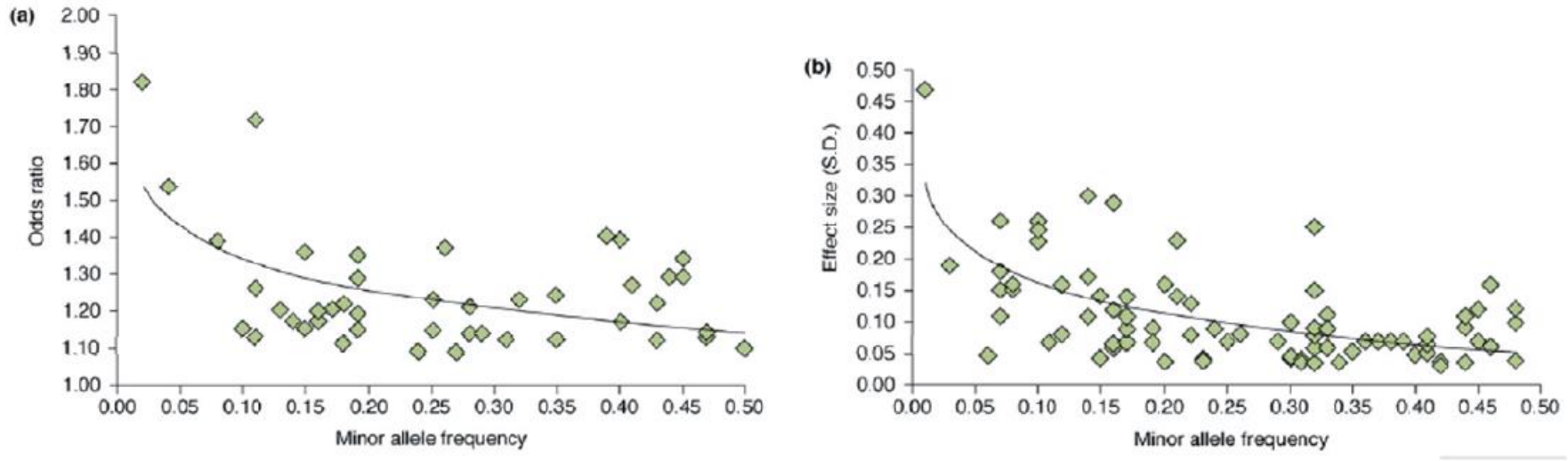
Challenge for many years to come …



(Glazier et al 2002)

Dichotomous Traits

Quantitative Traits

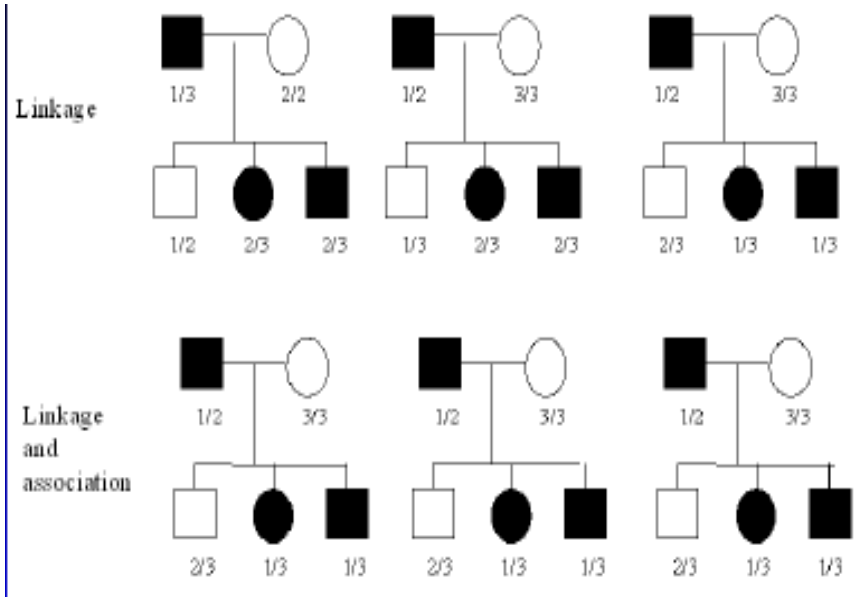


Arking & Chakravarti 2009 Trends Genet

Food for thought:

- The higher the MAF (minor allele frequency), the higher the detection rate?
- The higher the MAF, the lower the penetrance?

Which genetic markers to select?



(Figure: courtesy of Ed Silverman)

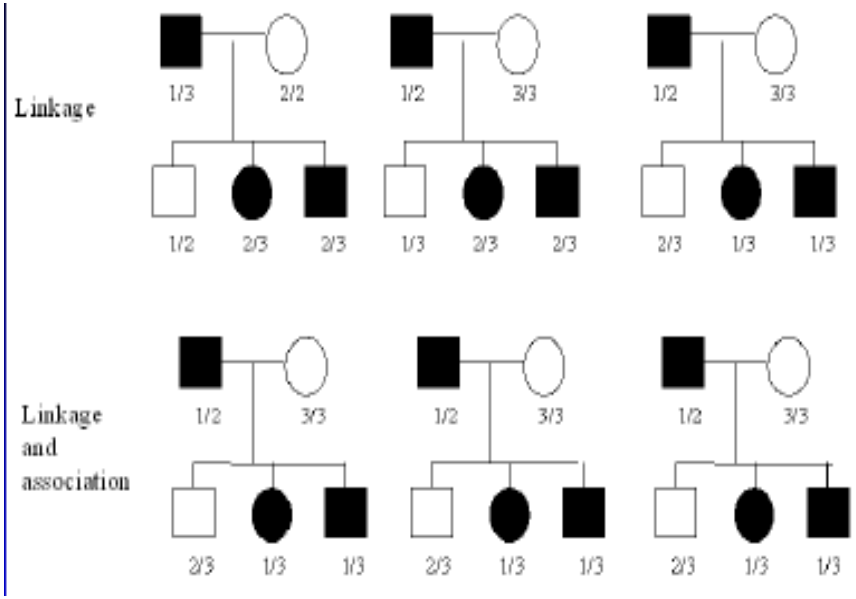
- Linkage exists over a very broad region, entire chromosome can be done using data on only 400-800 DNA markers
- Broad linkage regions imply studies must be followed up with more DNA markers in the region
- Must have family data with more than one affected subject

E.g., microsatellites

Which genetic markers to select?

- Association exists over a narrow region; markers must be close to disease gene
  - The basic concept is linkage disequilibrium (LD) – see later in this chapter
- Initially used for candidate genes or in linked regions
- Can use population-based (unrelated cases) or family-based design

E.g., SNPs

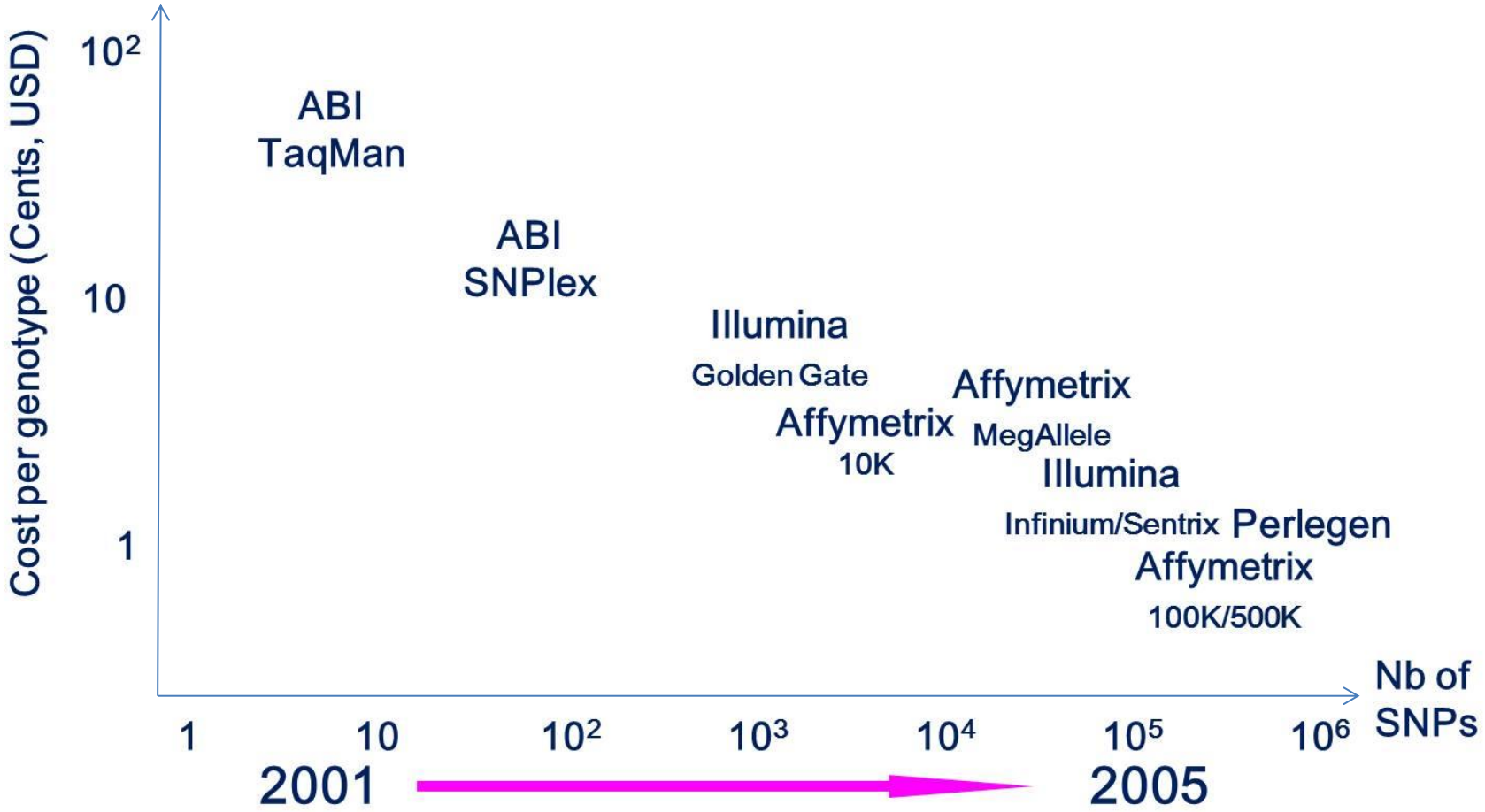


The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

SCIENCE • VOL. 273 • 13 SEPTEMBER 1996

Which DNA SNPs to select? (adapted from Manolio 2010)



How can technology bias be avoided?

- Standard experimental design problems
  - Cases and controls not balanced / randomized across plates
  - Controls borrowed from other studies
  - Trios/families split across plates
  - Genotyping performed at different sites and / or using different technologies and / or chips
- Consequences of design problems
  - Batch effects
  - High type I error fractions
  - Up to 50% of top hits discarded
  - Analyses of copy number variation extremely compromised

(Ziegler and Van Steen, Brazil 2010)

How can technology bias be avoided?

- DNA extraction
  - Same site
  - Same tissue (e.g., blood only)
  - Same extraction kit
  - Same time between freezing
  - Same collection time of cases and controls
  - Avoid cell lines
  - Avoid whole genome amplification (if necessary do it in both cases and controls)

(Ziegler and Van Steen, Brazil 2010)

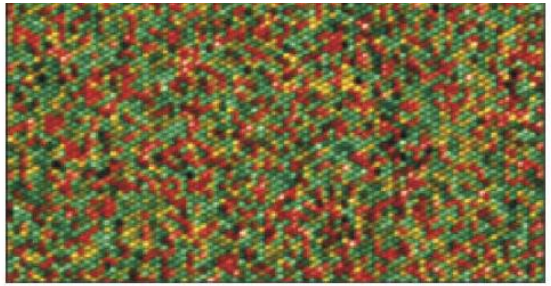
How can technology bias be avoided?

- Plating
  - Randomize phenotype/s across plates using statistical design
  - Stratify by gender
  - Run technical duplicates within and across plates to assess variability
  - Keep families together
  - Do it yourself, do not leave it to the laboratory
- Genotyping
  - All chips from single manufacturing lot
  - Genotype at single site
  - Genotype over shortest period of time possible
  - Avoid day effects, e.g., by using same technician over time
  - Re-genotype bad samples

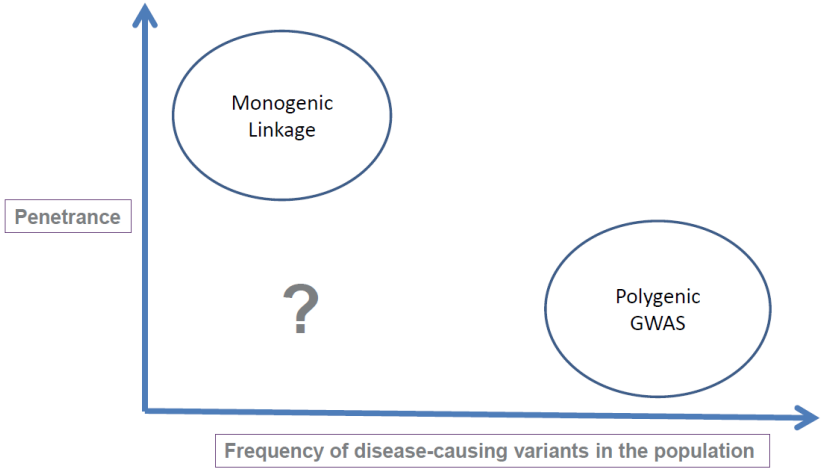
(Ziegler and Van Steen, Brazil 2010)

Choosing SNPs for GWAs: summary

- Costs may play a role, but a balance is needed between costs and chip performance as well as coverage (e.g., exonic regions only?)
- Old array technologies versus Next Generation Sequencing efforts to include rare variants into the game



Illumina 610S Quad Beadchip
Ragoussis 2009 Annu Rev Genomics Hum Genet



(Gut 2012)

2.b Subject Level

Which study subjects to select?

	Details	Advantages	Disadvantages	Statistical analysis method
Cross-sectional	Genotype and phenotype (ie, note disease status or quantitative trait value) a random sample from population	Inexpensive. Provides estimate of disease prevalence	Few affected individuals if disease rare	Logistic regression, χ^2 tests of association or linear regression
Cohort	Genotype subsection of population and follow disease incidence for specified time period	Provides estimate of disease incidence	Expensive to follow-up. Issues with drop-out	Survival analysis methods
Case-control	Genotype specified number of affected (case) and unaffected (control) individuals. Cases usually obtained from family practitioners or disease registries, controls obtained from random population sample or convenience sample	No need for follow-up. Provides estimates of exposure effects	Requires careful selection of controls. Potential for confounding (eg, population stratification)	Logistic regression, χ^2 tests of association
Extreme values	Genotype individuals with extreme (high or low) values of a quantitative trait, as established from initial cross-sectional or cohort sample	Genotype only most informative individuals hence save on genotyping costs	No estimate of true genetic effect sizes	Linear regression, non-parametric, or permutation approaches
Case-parent triads	Genotype affected individuals plus their parents (affected individuals determined from initial cross-sectional, cohort, or disease-outcome based sample)	Robust to population stratification. Can estimate maternal and imprinting effects	Less powerful than case-control design	Transmission/disequilibrium test, conditional logistic regression or log-linear models
Case-parent-grandparent septets	Genotype affected individuals plus their parents and grandparents	Robust to population stratification. Can estimate maternal and imprinting effects	Grandparents rarely available	Log-linear models
General pedigrees	Genotype random sample or disease-outcome based sample of families from general population. Phenotype for disease trait or quantitative trait	Higher power with large families. Sample may already exist from linkage studies	Expensive to genotype. Many missing individuals	Pedigree disequilibrium test, family-based association test, quantitative transmission/disequilibrium test
Case-only	Genotype only affected individuals, obtained from initial cross-sectional, cohort, or disease-outcome based sample	Most powerful design for detection of interaction effects	Can only estimate interaction effects. Very sensitive to population stratification	Logistic regression, χ^2 tests of association
DNA-pooling	Applies to variety of above designs, but genotyping is of pools of anywhere between two and 100 individuals, rather than on an individual basis	Potentially inexpensive compared with individual genotyping (but technology still under development)	Hard to estimate different experimental sources of variance	Estimation of components of variance

Table 2: Study designs for genetic association studies

(Cordell and Clayton 2005)

Which study subjects to select?

- Cohort studies
  - Assumption I: Participants under study representative for population of interest
  - Assumption II: Phenotypes ascertained similarly in subjects with and without the relevant genetic variants
  - Advantage I: Incident cases, free of survival bias
  - Advantage II: If prevalent cases available, too, comparison of incident and prevalent cases possible
  - Advantage III: Availability of intermediate phenotypes (quantitative traits) with distribution as in population
  - Advantage IV: Direct measure of risk
  - Advantage V: Fewer bias than case-control studies
  - Disadvantage I: Long follow-up required

(Ziegler and Van Steen, 2010)

Which study subjects to select?

- Cohort studies (continued)
  - Disadvantage II: Large sample size required
  - Disadvantage III: Expensive
  - Disadvantage IV: Poorly suited for studying rare diseases
  - Disadvantage VII: Unbalanced distribution of cases and controls
  - Disadvantage V: Consent for GWA genotyping often required
  - Disadvantage VI: Consent for data sharing often required
  - Disadvantage VIII: DNA quality

(Ziegler and Van Steen, 2010)

Which study subjects to select?

- Family-based association studies
  - Assumption I: Families representative for population of interest
  - Assumption II: Same genetic background in both parents
  - Advantage I: Controls immune to population stratification, i.e., no spurious associations, i.e., no association without linkage
  - Advantage II: Checks for Mendelian inheritance possible, i.e., fewer genotyping errors
  - Advantage III: Parental phenotyping not required
  - Advantage IV: Simple logistics for diseases in children
  - Advantage V: Allows investigation of imprinting
  - Disadvantage I: Cost inefficient
  - Disadvantage II: Lower power when compared with case-control studies
  - Disadvantage III: Sensitive to genotyping errors

(Ziegler and Van Steen, 2010)

Which study subjects to select?

- Case-control studies
  - Assumption I: Cases and controls drawn from same population
  - Assumption II: Cases representative for all cases in population
  - Assumption III: All data collected similarly in cases and controls
  - Advantage I: Simple
  - Advantage II: Cheap
  - Advantage III: Large number of cases and controls available
  - Advantage IV: Optimal for studying rare diseases
  - Disadvantage I: Prone to population stratification
  - Disadvantage II: Prone to batch effects
  - Disadvantage III: Prone to other biases
  - Disadvantage IV: Cases usually prevalent ↓ fatal, short episodes, mild cases …
  - Disadvantage V: Overestimation of risk for common disease

(Ziegler and Van Steen, 2010)

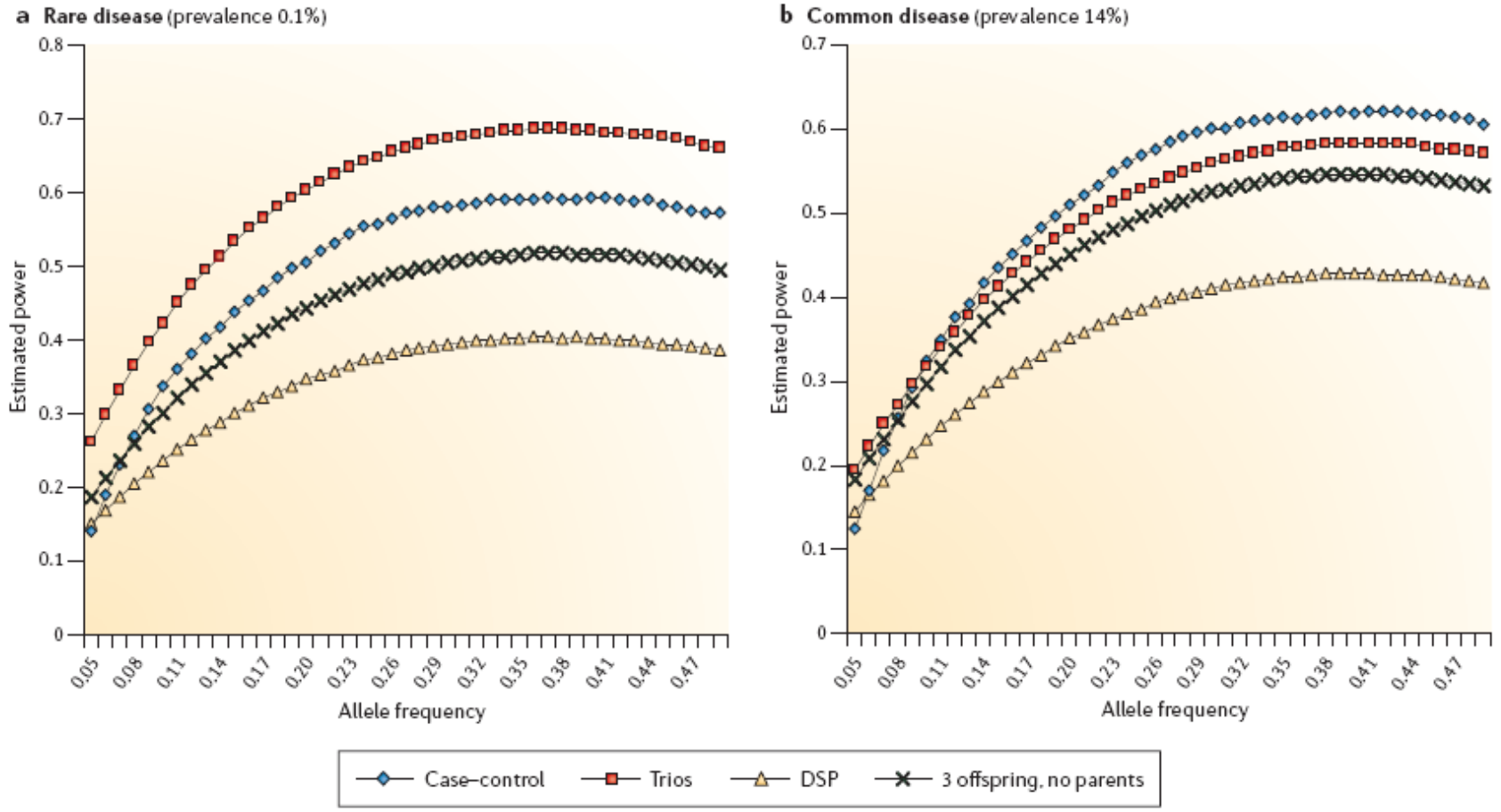
Which study subjects to select?

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

Which study subjects to select?

Rare versus common diseases (Lange and Laird 2006)



3 GWA Details: Prior Analyses

Is there a 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

Is there a standard file format for GWA studies?

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

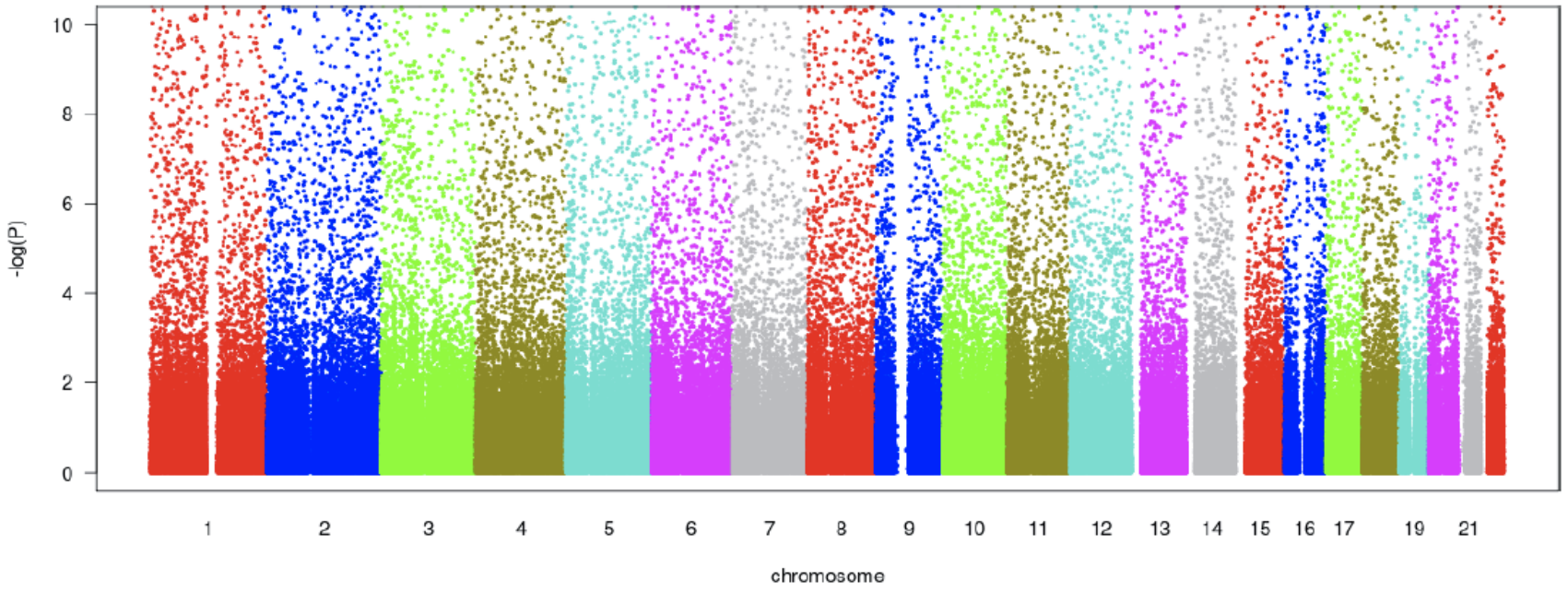
FamID	PID	FID	MID	SEX	AFF
1	1	0	0	1	1
2	1	0	0	1	1
3	1	0	0	1	1
4	1	0	0	1	2
5	1	0	0	1	2
6	1	0	0	1	2

tfam file

3.a Quality control

Why is quality control important?

BEFORE (false positives !!!!):

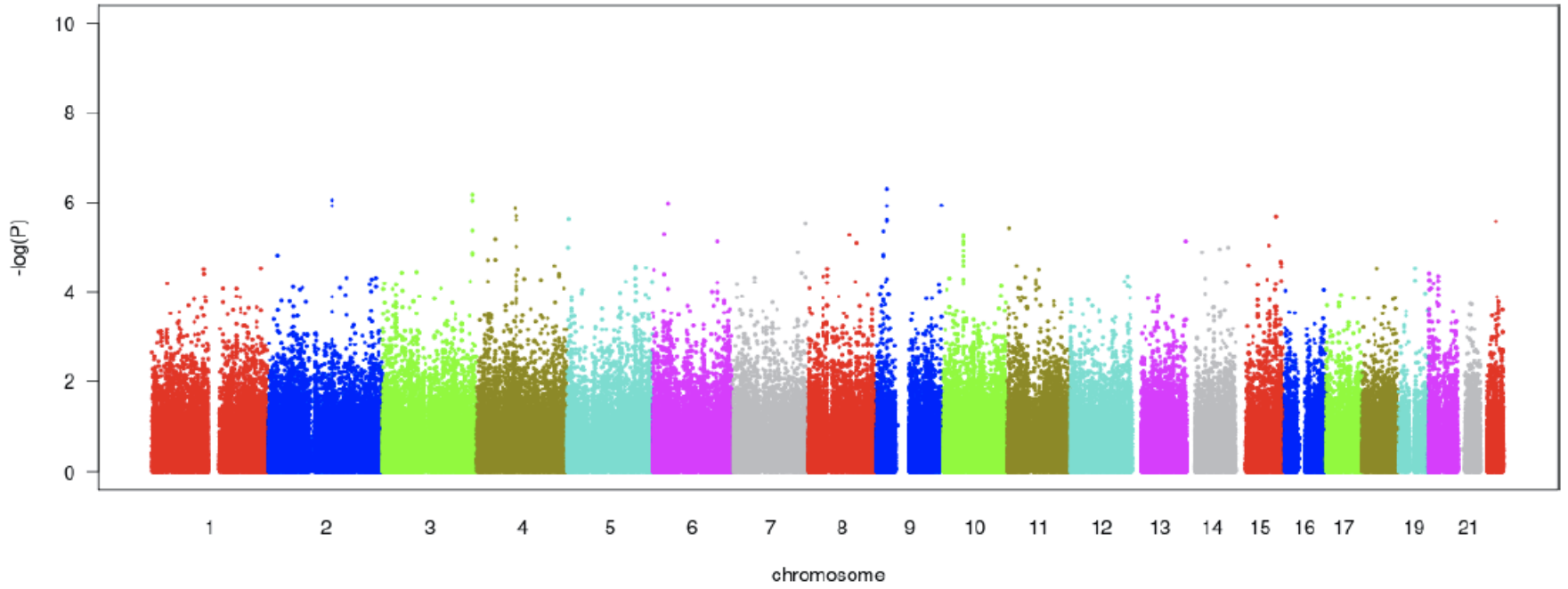


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(Ziegler and Van Steen 2010)

Why is quality control important?

AFTER:



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 on different levels:
  - Subject or sample level
  - SNP level
  - X-chromosomal SNP level

What are standard filters on the sample level?

- Call fraction as high as possible
- Cryptic relatedness: if identity by state (IBS) too high, subjects closely related
- Ethnic origin (principal component, multidimensional scaling, non-metric multidimensional scaling): homogeneous study populations required
- No excess or deficiency of heterozygosity (contamination of DNA, hybridization failure)

(Ziegler and Van Steen 2010)

What are standard filters on the SNP level?

- Minor allele frequency (MAF)
  - Genotype calling algorithms perform poorly for SNPs with low MAF
  - Power low for detecting associations to SNPs with low MAF,
- Missing frequency (MiF)
  - Also termed 1 minus SNP call rate
  - Indicator for cluster separation
  - Investigate MiF separately in cases and in controls because of differential missingness
- Hardy-Weinberg equilibrium (HWE)
  - SNPs excluded if substantially more or fewer subjects heterozygous at a SNP than expected (excess heterozygosity or heterozygote deficiency)

(Ziegler and Van Steen 2010)

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

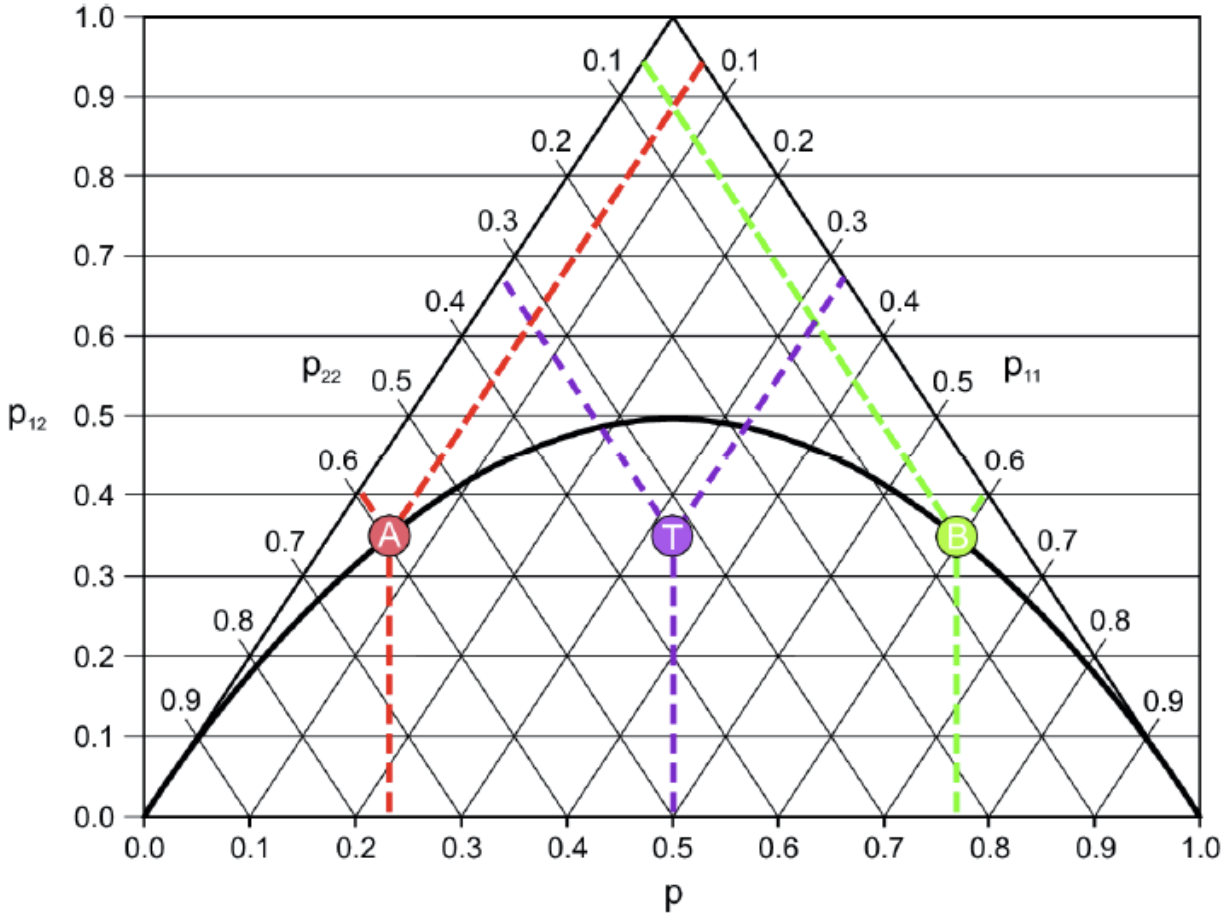
(Ziegler and Van Steen 2010)

What are the assumptions of HWE?

- Random mating
- No selection or migration
- No mutation
- No population stratification
- Infinite population size

One of the signs of deviations from HWE?

Increased HOM (e.g., in case of population stratification; Wahlund effect)



How can HWE be measured?

A simple calculator to determine whether observed genotype frequencies are consistent with Hardy-Weinberg equilibrium

Genotypes	Observed #	Expected #
Homozygote reference:	68	67,2
Heterozygote:	5	6,7
Homozygote variant:	1	0,2
	^Put your values here^	
Var allele freq:	0,05	

X^2 = 4,634376581

X^2 test *P* value = 0,031338 with 1 degree of freedom.

1. If $P < 0.05$ - not consistent with HWE.
2. Not accurate if <5 individuals in any genotype group.

Michael H. Court (2005-2008)

How can HWE be measured?

- The χ2 approximation can be poor when there are low genotype counts, in which case it is better to use a Fisher exact test.
- Discard loci that, for example, deviate from HWE among controls at significance level $\alpha = 10^{-3}$ or 10^{-4}. But be flexible!
- The open-source data-analysis software R includes the "*SNPassoc*" package that implements an exact SNP test of Hardy-Weinberg Equilibrium for you (http://www.sph.umich.edu/csg/abecasis/Exact/snp_hwe.r)

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Expectations computed under the null of HWE

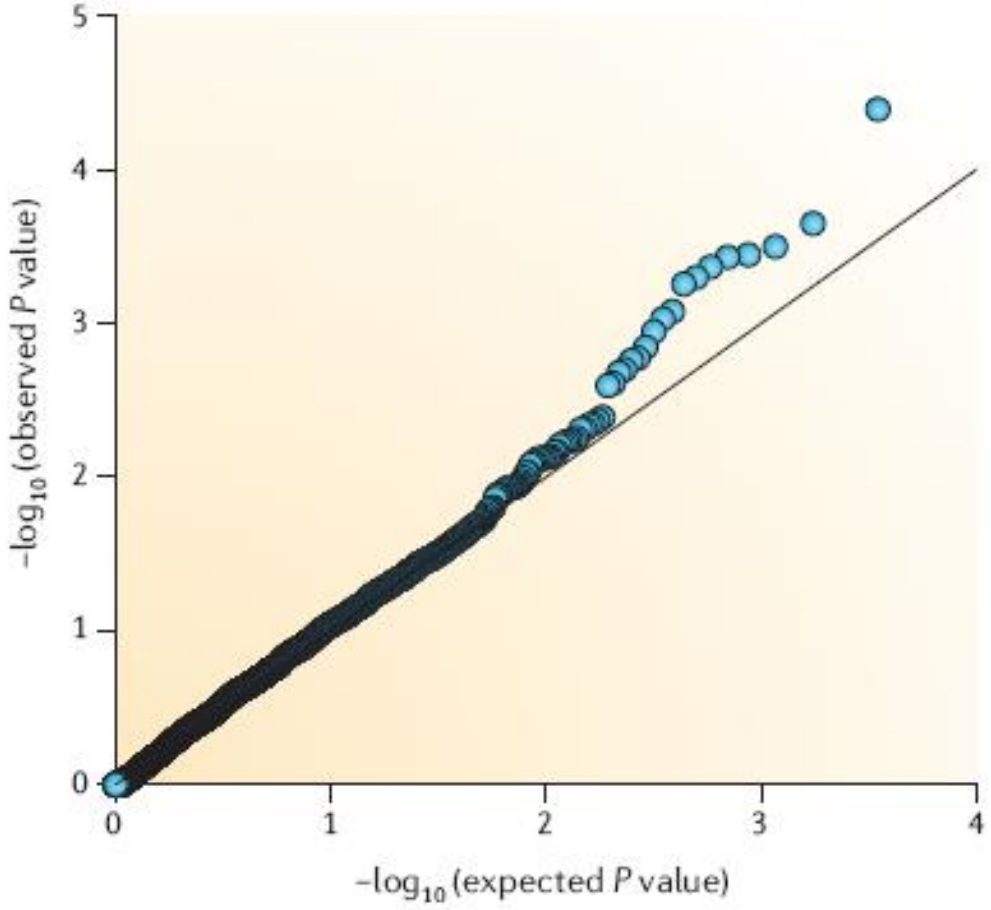
Nr of degrees of freedom is 1 (p+q=1)

How can extreme HWD be visualized?

- A useful tool for interpreting the results of HWE and other tests on many SNPs is the **log quantile–quantile (QQ) *p*-value plot**:
  - the negative logarithm of the *i*-th smallest *p*-value is plotted against $-\log(i / (L + 1))$, where L is the number of SNPs.
  - The 0.45 (or 45%) quantile is the point at which 45% percent of the data fall below and 55% fall above that value.

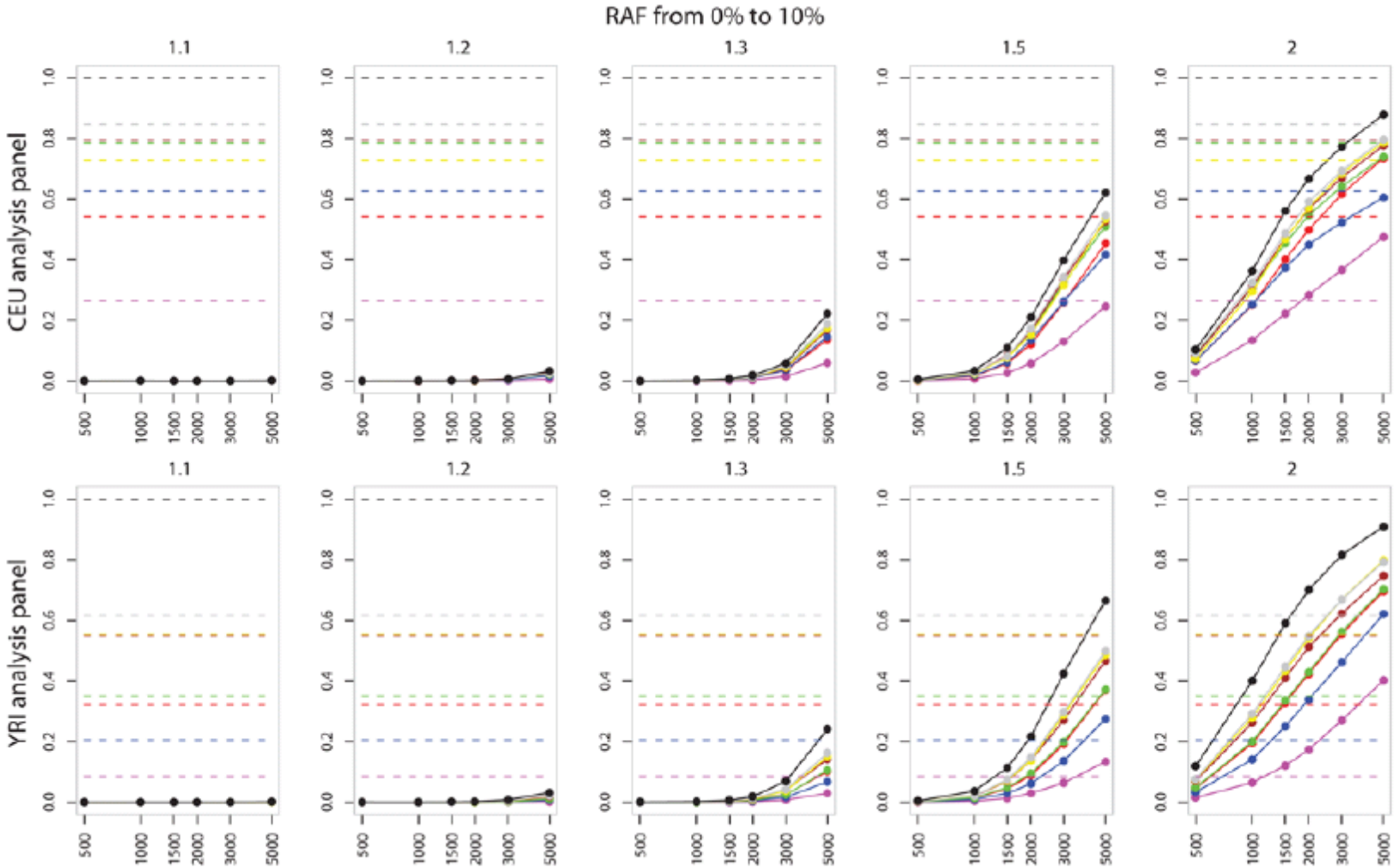
- A 45-degree reference line is also plotted as visualization tool:
  - If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
  - The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

How can extreme HWD be visualized?



(Balding 2006)

Is there a power advantage in imputing? (Spencer et al 2009)



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)

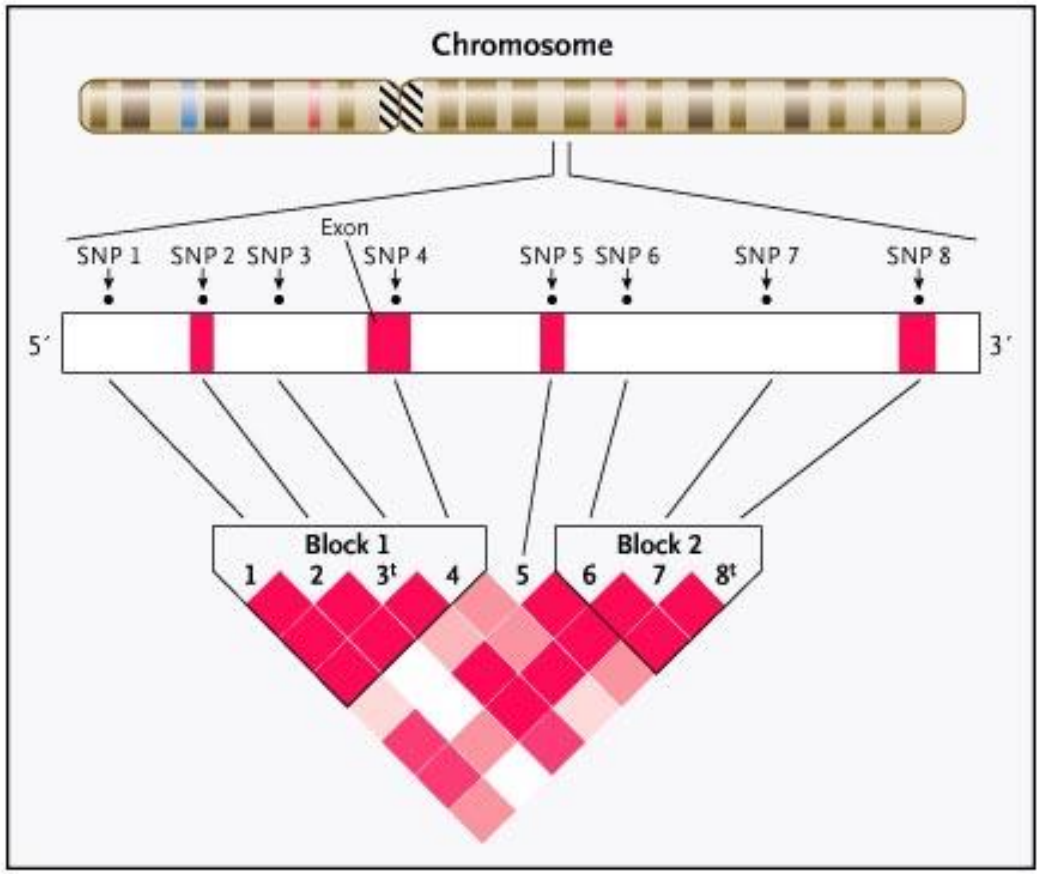
What are the Travemünde criteria?

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

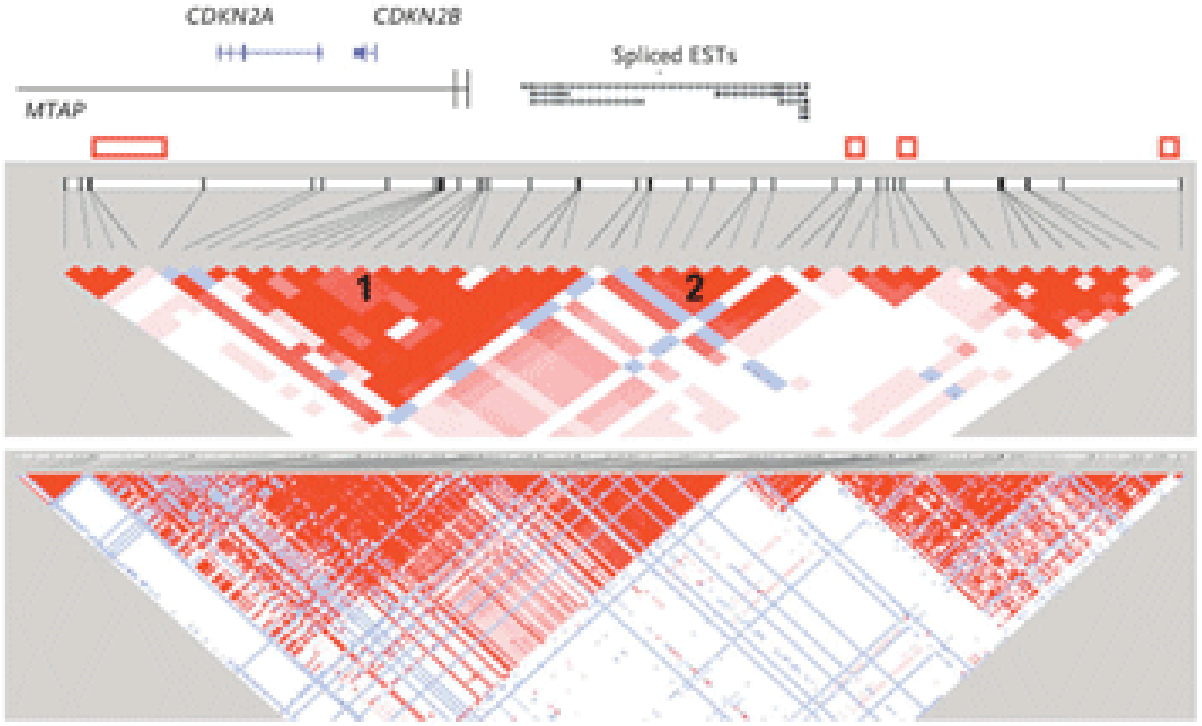
(Ziegler 2009)

3.b Linkage disequilibrium and SNP tagging

Mapping the relationships among SNPs (Christensen and Murray 2007)



Relationships among SNPs induce multiple signals



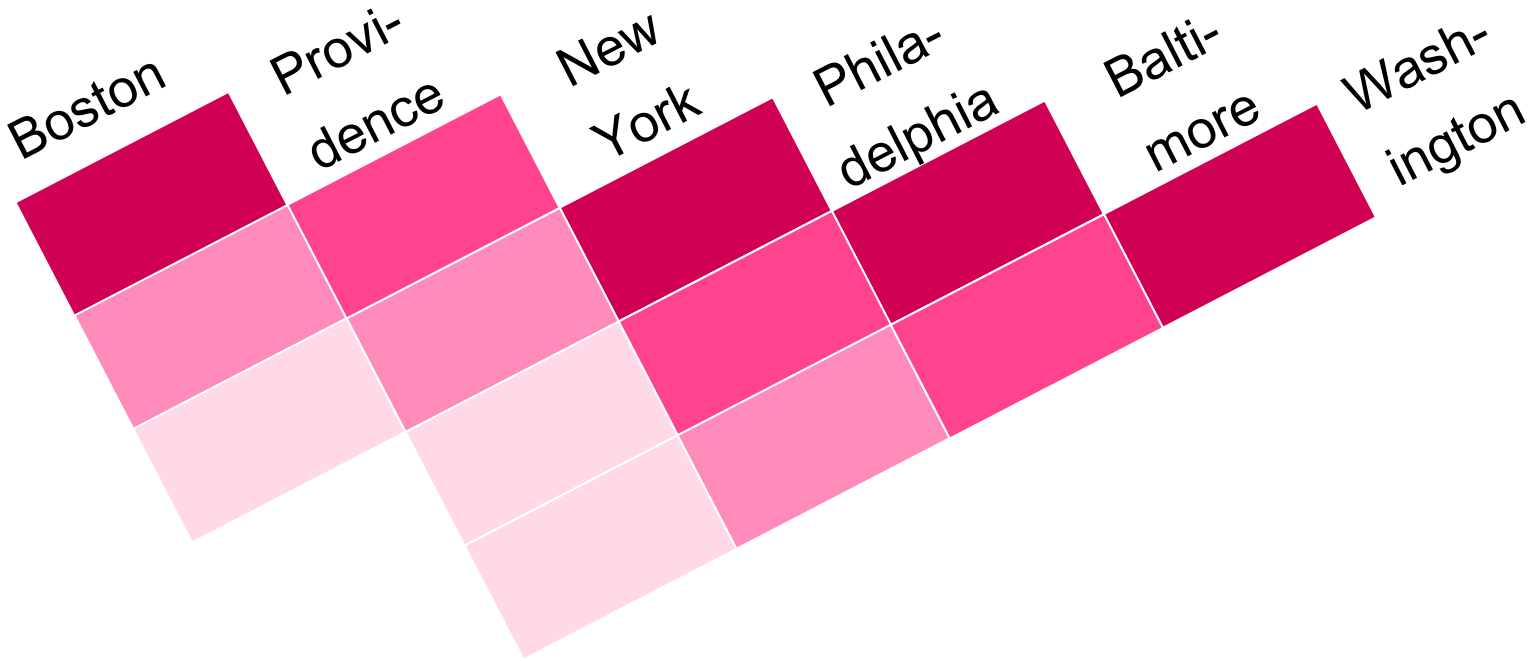
(Samani et al 2007))

- These plots can be generated using the free software "*Haploview*", but also in R!

Distances among cities

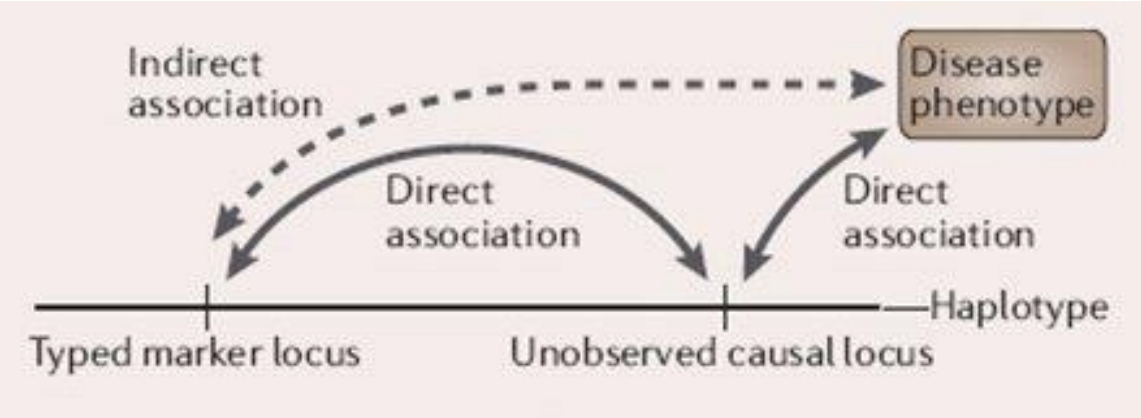
	Boston	Provi-dence	New York	Phila-delphia	Balti-more
Providence	59				
New York	210	152			
Philadelphia	320	237	86		
Baltimore	430	325	173	87	
Washington	450	358	206	120	34

Distances among cities


Boston
Provi-
dence
New
York
Phila-
delphia
Balti-
more
Wash-
ington

Distances among SNPs

- **Linkage Disequilibrium** (LD) is a measure of co-segregation of alleles in a population: Two alleles at different loci that occur together on the same chromosome (or gamete) more often than would be predicted by random chance.
- Hence, in general, LD is taken to be a **measure of allelic association**.
- 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}

  - D’ (Lewontin’s *D* prime) is the absolute ratio of D compared with its maximum value.
  - D’ =1 : complete LD

Distances among SNPs

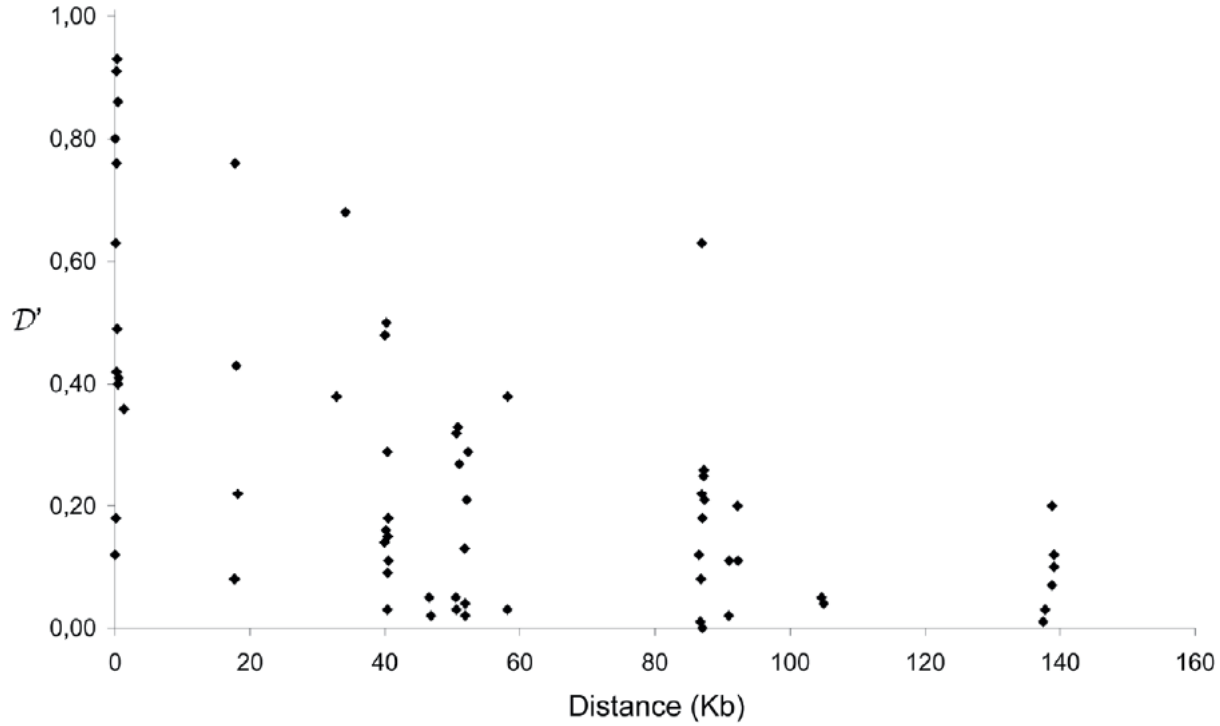
- The most popular measure of allelic association in GWA contexts is r^2 (the square correlation coefficient between the two loci under study).
- When r^2=1, knowing the genotypes of alleles of one SNP is directly predictive of genotype of another SNP
- r^2 relates to D in the following way:

$$R^2 = \frac{D^2}{P(A)P(a)P(B)P(b)}$$

- Sample size must be increased by a factor of $1/r^2$ to detect an unmeasured variant, compared with the sample size for testing the variant itself.

(Jorgenson and Witte 2006)

How far does linkage disequilibrium extend?



(Hecker et al 2003)

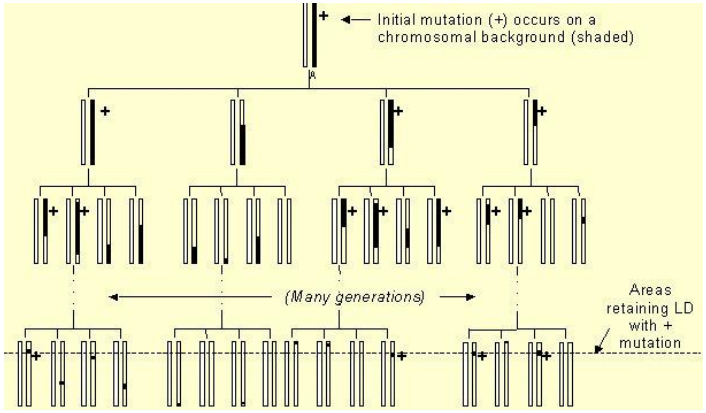
- LD is usually a function of distance between the two loci. This is mainly because recombination acts to break down LD in successive generations (Hill, 1966).

How to interpret LD data?

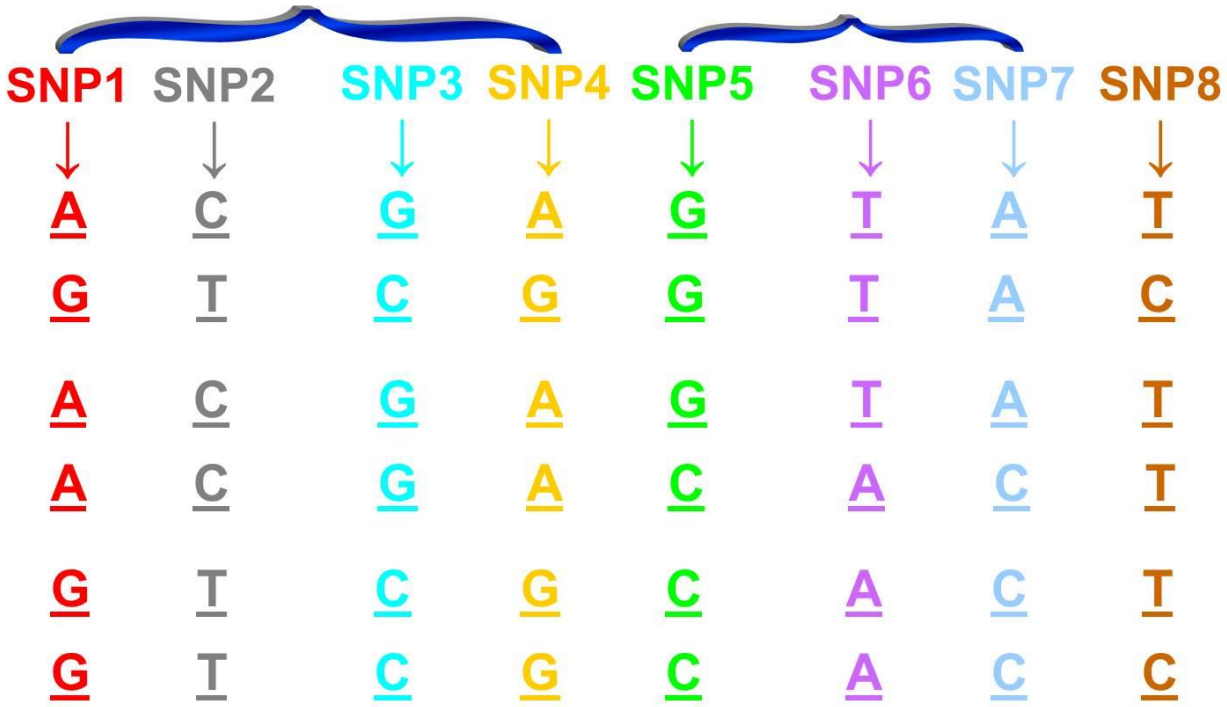
- The patterns of LD observed in natural populations are the result of a complex interplay between genetic factors and the population's demographic history (Pritchard, 2001).

- When a mutation first occurs it is in complete LD with the nearest marker (D' = 1.0). Given enough time and as a function of the distance between the mutation and the marker, LD tends to decay and in complete equilibrium reached D' = 0 value.

- Thus, it decreases at every generation of random mating unless some process is opposing to the approach to linkage 'equilibrium'.

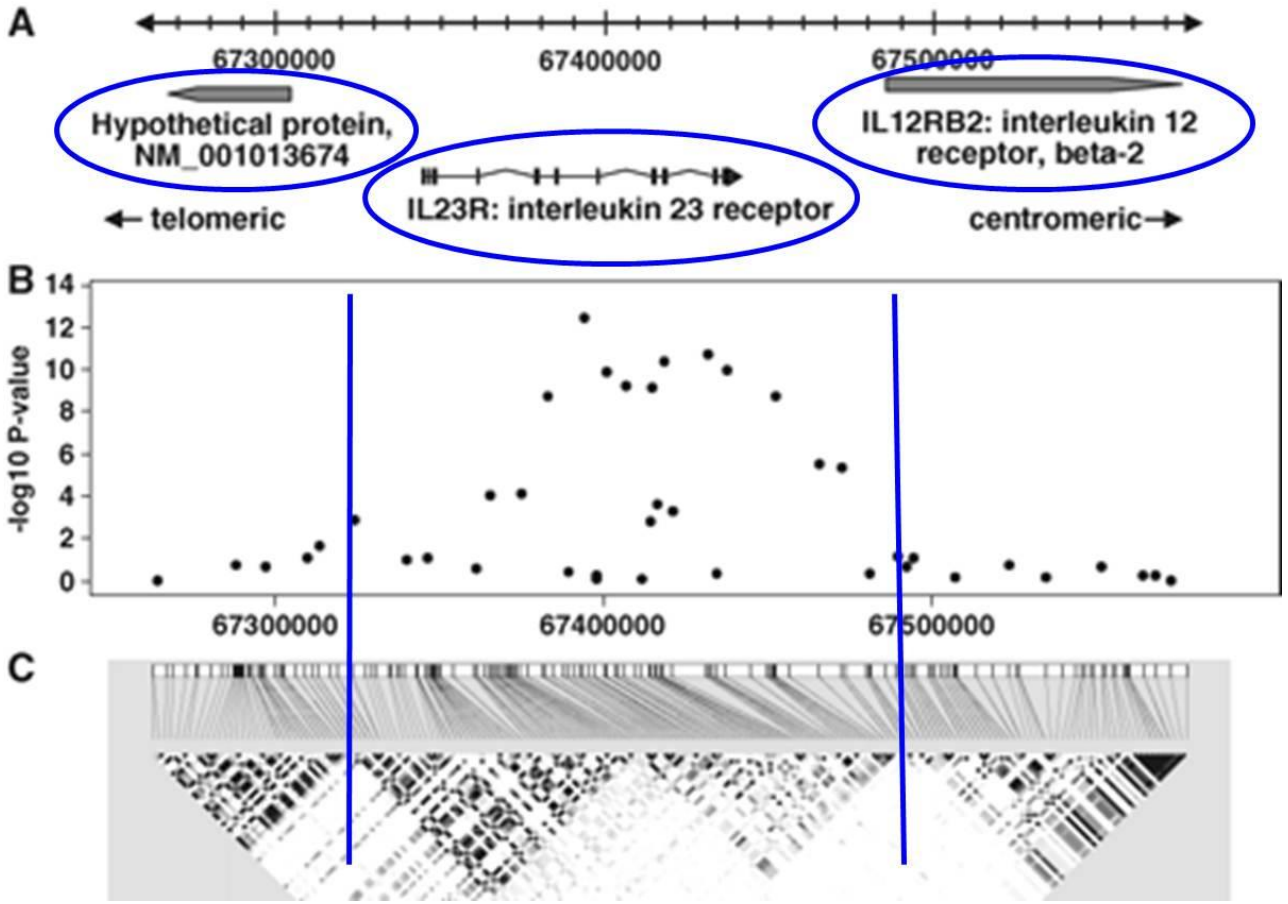


How can one tag SNP serve as proxy for many? (adapted from Manolio 2010)



How can one tag SNP serve as proxy for many? (adapted from Manolio 2010)

Where is the true causal variant? One of our proxy's? …

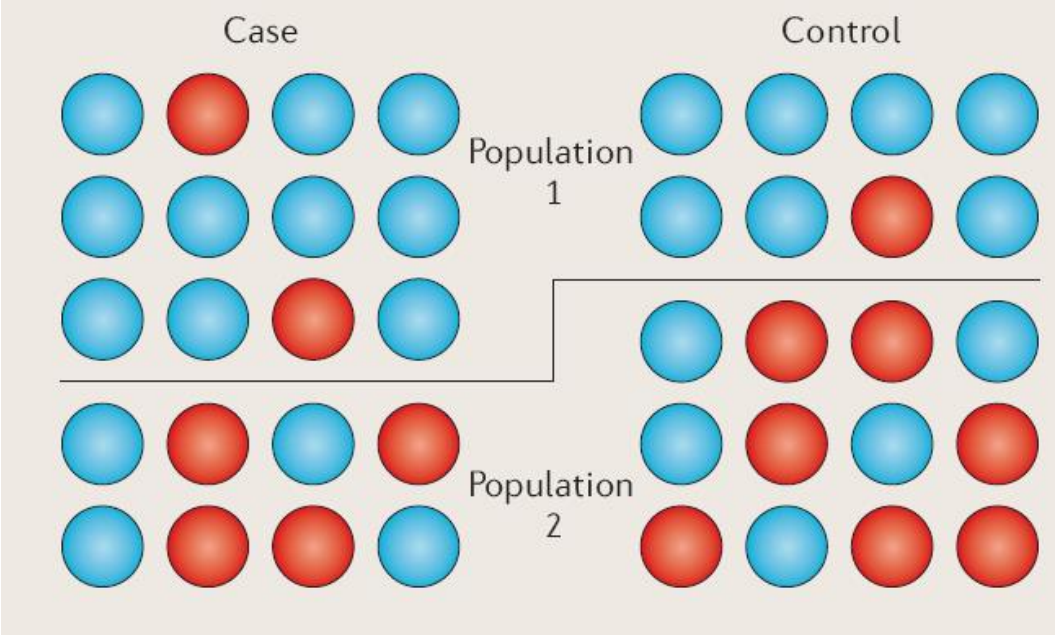


(Duerr et al 2006)

3.c Confounding

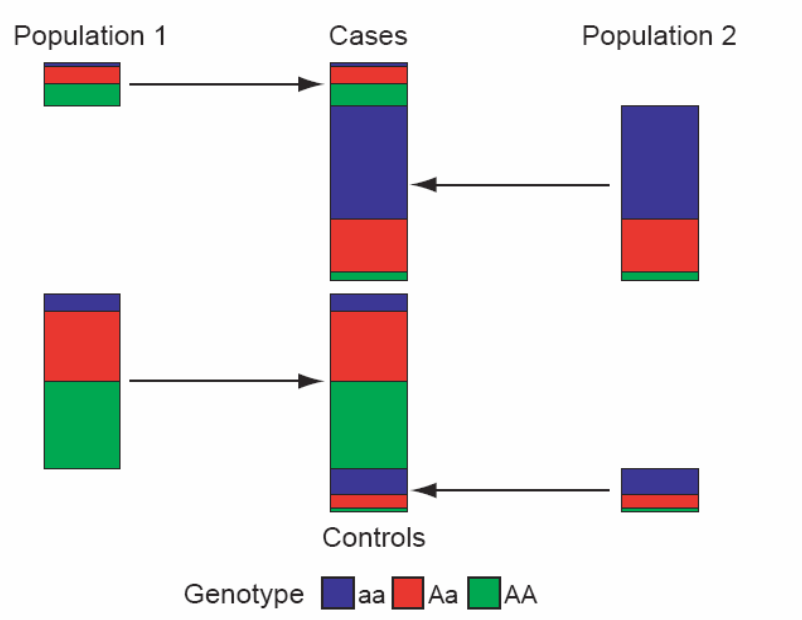
What is spurious association?

- 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:
  - Genomic control methods
  - Structured association methdos
  - Principal component-based methods

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

What is genomic control?

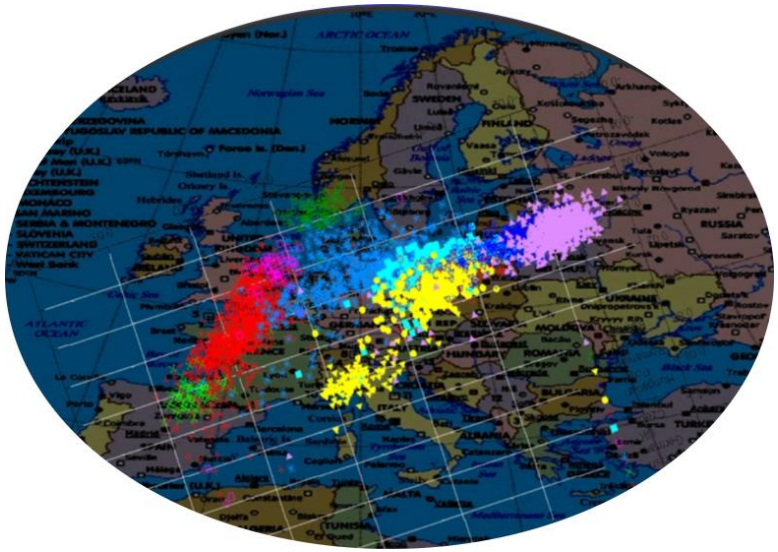
- The motivation for GC is that, as we expect few if any of the null SNPs to be associated with the phenotype, a value of $\lambda > 1$ is likely to be due to the effect of population stratification, and dividing by λ cancels this effect for the candidate SNPs.
- GC performs well under many scenarios, but can be conservative in extreme settings (and anti-conservative if insufficient null SNPs are used).
- There is an analogous procedure for a general (2 df) test; The method can also be applied to other testing approaches.

What is a structured association method?

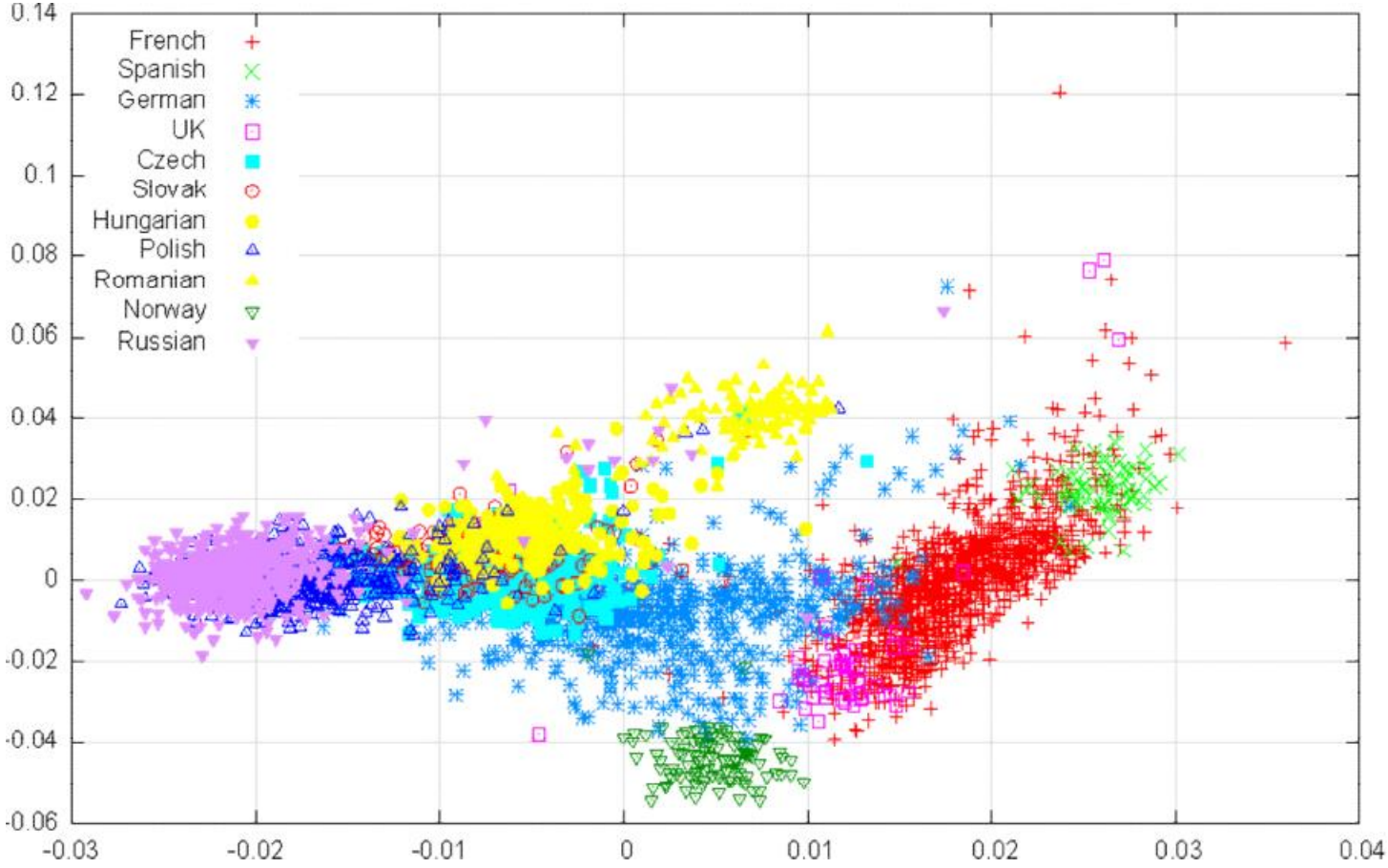
- Structured association (SA) approaches are based on the idea of attributing the genomes of study individuals to hypothetical subpopulations, and testing for association that is conditional on this subpopulation allocation.
- Several clustering algorithms exist to estimate the number of subpopulations.
- These approaches (such as Bayesian clustering approaches) are computationally demanding, and because the notion of subpopulation is a theoretical construct that only imperfectly reflects reality, the question of the correct number of subpopulations can never be fully resolved….

What is principal components analysis?

- When many null markers are available, principal components analysis provides a fast and effective way to diagnose population structure.
- Principal components are linear combinations of the original “variables” (here SNPs) that optimized in such a way that as much of the variation in the data is retained.

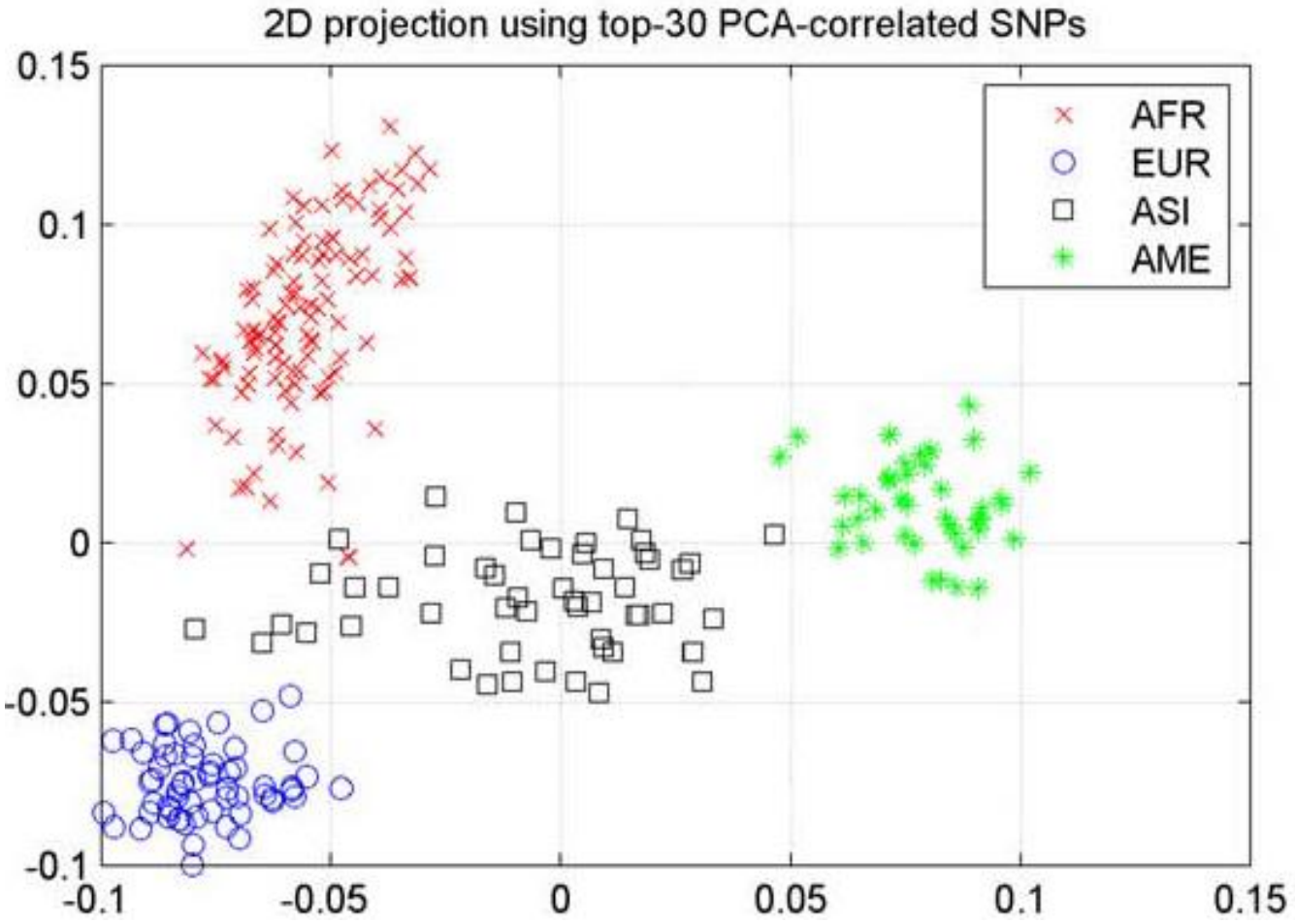


- In European data, the first 2 principal components “nicely” reflect the N-S and E-W axes !



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

- Does the same hold on a “global” (world) level?



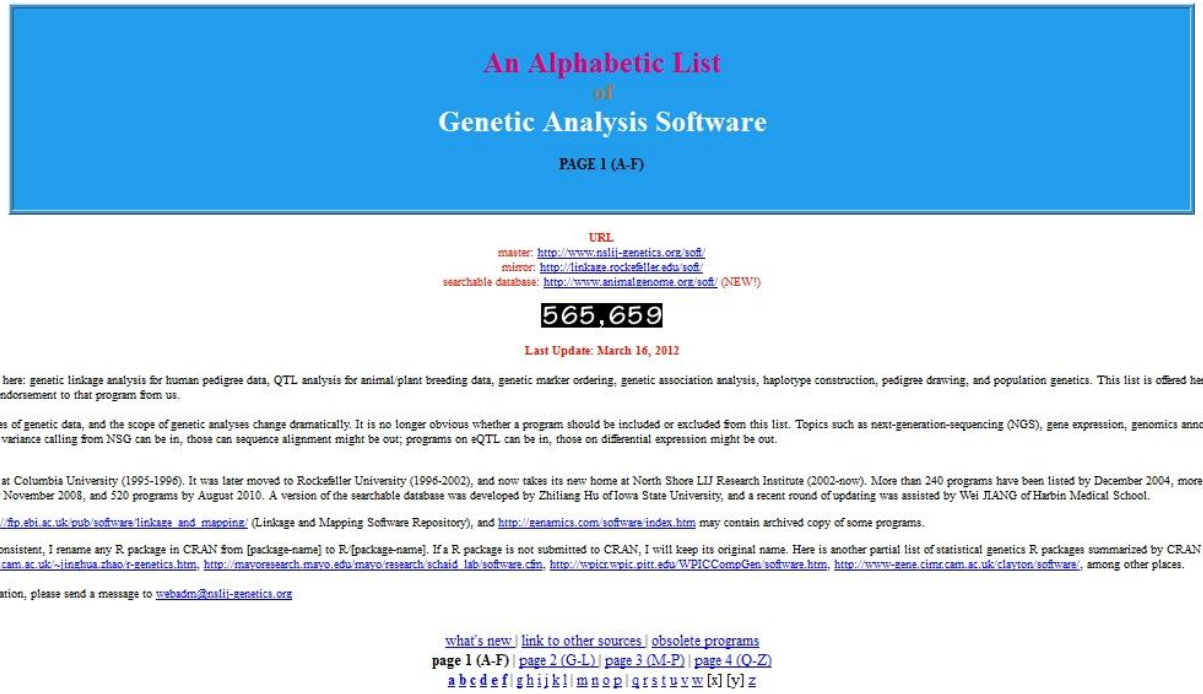
(Paschau 2007)

4 GWA Details: Testing for Associations

Can screening for 1000nds of SNPs be performed automatically in R?

- **GenAbel** is designed for the efficient storage and handling of GWAS data with fast analysis tools for quality control, association with **binary and quantitative traits**, as well as tools for visualizing results.
- *pbatR* provides a GUI to the powerful PBAT software which performs family and population based family and population based studies. The software has been implemented to take advantage of parallel processing, which vastly reduces the computational time required for GWAS.
- *SNPassoc* provides another package for carrying out GWAS analysis. It offers descriptive statistics of the data (including patterns of missing data!) and tests for Hardy-Weinberg equilibrium. Single-point analyses with binary or quantitative traits are implemented via generalized linear models, and multiple SNPs can be analyzed for haplotypic associations or epistasis.

Is there one tool that fits it all? NO



An Alphabetic List
of
Genetic Analysis Software

PAGE 1 (A-F)

URL
master: http://www.nslij-genetics.org/soft/
mirror: http://linkage.rockefeller.edu/soft/
searchable database: http://www.animalgenome.org/soft/ (NEW!)

565,659

Last Update: March 16, 2012

Computer software on the following topics are included here: genetic linkage analysis for human pedigree data, QTL analysis for animal/plant breeding data, genetic marker ordering, genetic association analysis, haplotype construction, pedigree drawing, and population genetics. This list is offered here as a service to the gene mapping community. The inclusion of a program should not be interpreted as an endorsement to that program from us.

In the last few years, new technology produces new types of genetic data, and the scope of genetic analyses change dramatically. It is no longer obvious whether a program should be included or excluded from this list. Topics such as next-generation-sequencing (NGS), gene expression, genomics annotation, etc. can all be relevant to a genetic study, yet be specialized topics by themselves. Though programs on variance calling from NSG can be in, those can sequence alignment might be out; programs on eQTL can be in, those on differential expression might be out.

This page was created by Dr. Wentian Li, when he was at Columbia University (1995-1996). It was later moved to Rockefeller University (1996-2002), and now takes its new home at North Shore LIJ Research Institute (2002-now). More than 240 programs have been listed by December 2004, more than 350 programs by August 2005, close to 400 programs by December 2006, close to 480 programs by November 2008, and 520 programs by August 2010. A version of the searchable database was developed by Zhiliang Hu of Iowa State University, and a recent round of updating was assisted by Wei JIANG of Harbin Medical School.

Some earlier software can be downloaded from EBI: ftp://ftp.ebi.ac.uk/pub/software/linkage_and_mapping/ (Linkage and Mapping Software Repository), and http://genamics.com/software/index.htm may contain archived copy of some programs.

More and more packages are now written in R. To be consistent, I rename any R package in CRAN from [package-name] to R/[package-name]. If a R package is not submitted to CRAN, I will keep its original name. Here is another partial list of statistical genetics R packages summarized by CRAN (http://cran.r-project.org/web/views/Genetics.html). Yet more R packages can be found in: http://www.mrc-epid.cam.ac.uk/~jinghua.zhao/r-genetics.htm, http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm, http://wpicr.wpic.pitt.edu/WPICCompGen/software.htm, http://www-gene.cimr.cam.ac.uk/clayton/software/, among other places.

If you have new programs to add or any updated information, please send a message to webadm@nslij-genetics.org

what's new | link to other sources | obsolete programs
page 1 (A-F) | page 2 (G-L) | page 3 (M-P) | page 4 (Q-Z)
a b c d e f | g h i j k l | m n o p | q r s t u v w [x] [y] z

(http://linkage.rockefeller.edu/soft/)

Other analytic methods

- Recursive Partitioning (CART; Breiman 1984, Foulkes 2005)
- Random Forests (Pavolov 1997)
- Combinatorial Partitioning (Nelson 2001)
- **Multifactor-Dimensionality Reduction (Ritchie 2001) → interactions !**
- Permutation-Based Procedures (Trimming/Weighting; Hoh 2000)
- Multivariate Adaptive Regression Splines (Friedman 1991)
- Boosting (Schapire 1990)
- Support Vector Machines (Vapnik 2000)
- Neural Networks (Friedman & Tukey 1974, Friedman & Stuetzle 1981)
- Bayesian Pathway Modeling (Conti 2003, Cortessis & Thomas 2004)
- Clique-Finding (Mushlin 2006)

What is a multiple testing correction?

- Simultaneously test *m* null hypotheses, one for each SNP *j*

  H_{0j}: no association between SNP *j* and the trait
- Every statistical test comes with an inherent false positive, or type I error rate—which is equal to the threshold set for statistical significance, generally 0.05.
- However, this is just the error rate for one test. When more than one test is run, the overall type I error rate is much greater than 5%.

What is a multiple testing correction?

- Suppose 100 statistical tests are run when (1) there are no real effects and (2) these tests are independent, then the probability that no false positives occur in 100 tests is $0.95^{100} = 0.006$. So the probability that at least one false positive occurs is 1-0.006=0.994 or 99.4%
- There is not a single measure to quantify false positives (Hochberg et al 1987): FEW (family-wise error); FDR (false discovery rate); ...

What is a multiple testing correction?

- Several multiple testing corrections have been developed and curtailed to a genome-wide association context, when deemed necessary
- *Bonferroni* (highly conservative) [divide each single SNP-based p-value by the nr of tests before comparing to the nominal sign level 0.05] vs
- *permutation-based* (highly computational demanding) [keep the LD structure, but swap the trait labels among the subjects]

Note: To reduce the multiple testing burden one can exploit the LD structure in the data (e.g., perform multilocus tests, or haplotype tests, or take a limited number of tagging SNPs to be tested one at a time).

4.c Replication

nature genetics

Freely associating

Editorial: Once and Again—Issues Surrounding Replication in Genetic Association Studies

May

J. Hirschhorn

PERSPECTIVE

The Future of Association Studies: Gene-Based Analysis and Replication

Benjamin M. Neale[1] and Pak C. Sham[1,2]

Am J Hum Genet July

Editorial

Replication Publication

Mark Patterson

Statistical false positive or true disease pathway?

John A Todd

Nat Genet July 2006

What does replication mean?

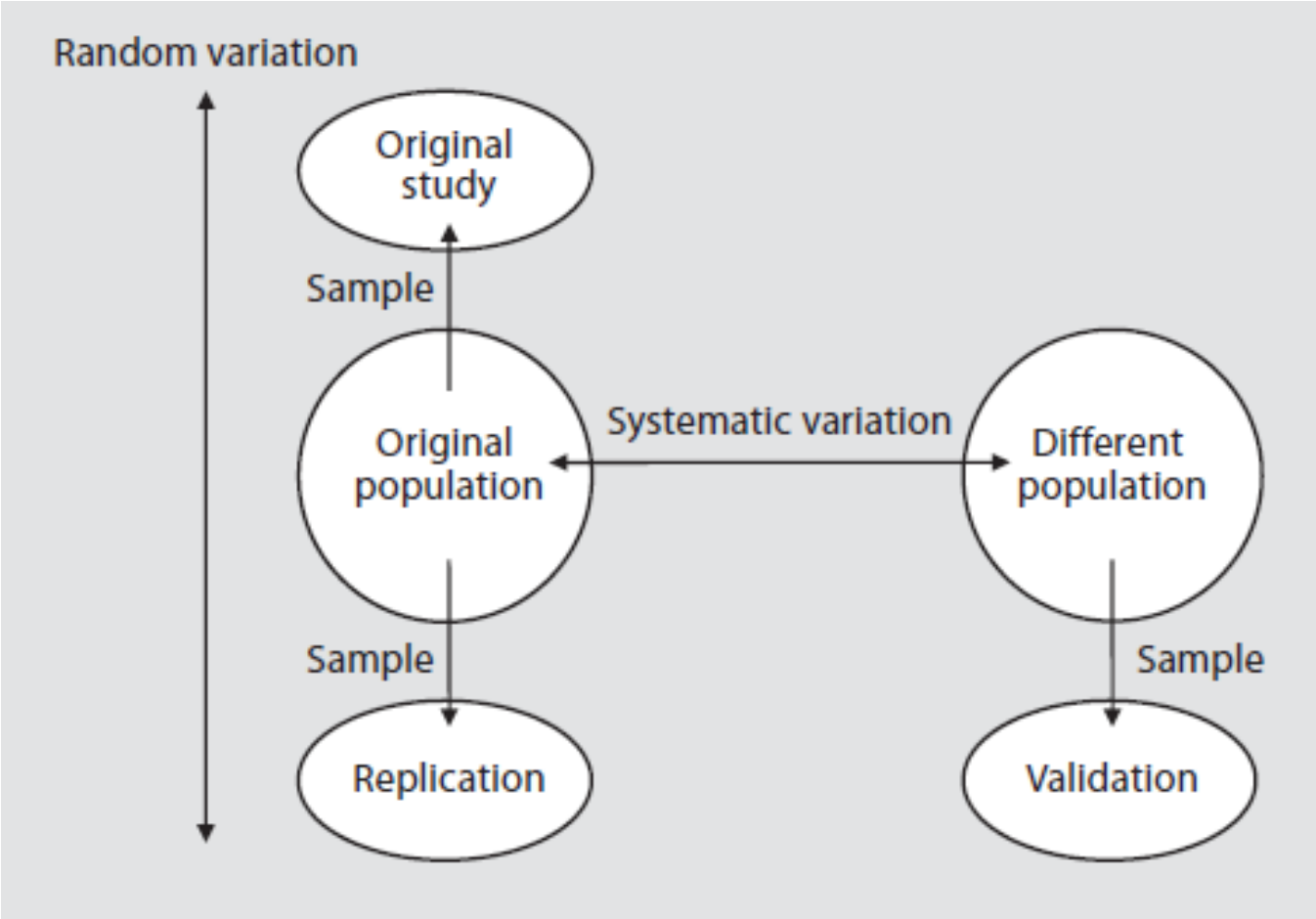
- Replicating the genotype-phenotype association is the “gold standard” for “proving” an association is genuine
- Most loci underlying complex diseases will not be of large effect. It is unlikely that a single study will unequivocally establish an association without the need for replication → think about what this means when targeting gene-gene or gene-environment interactions!!!
- SNPs most likely to replicate:
  - Showing modest to strong statistical significance
  - Having common minor allele frequency
  - Exhibiting modest to strong genetic effect size

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

➔ **check the NHGRI Catalog of GWA studies**
www.genome.gov/gwastudies/

What does validation mean?

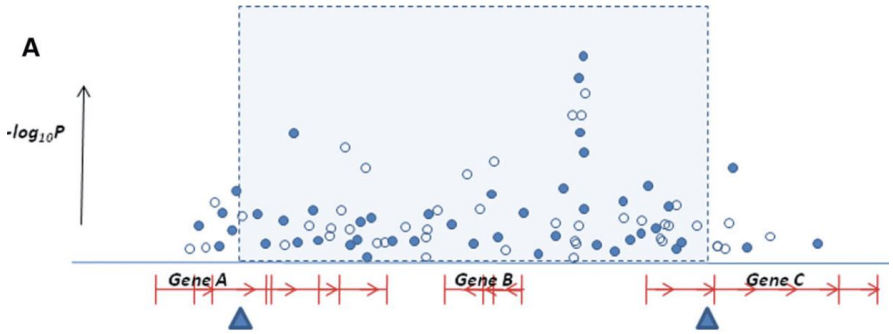


(Igl et al. 2009)

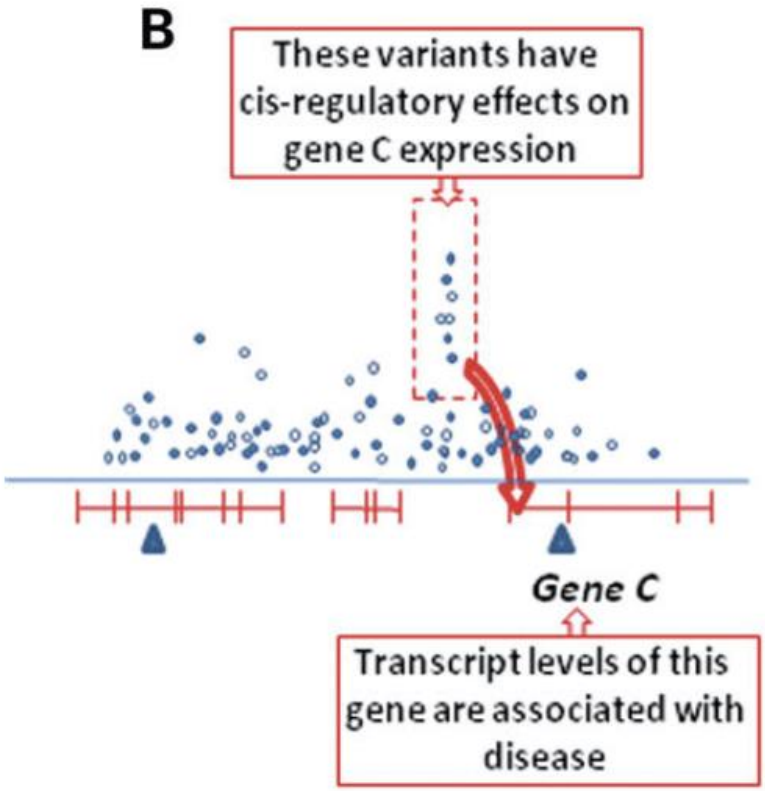
5 GWAI Interpretation and Follow-Up

Strategies for using functional data to support causal variant and causal gene identification

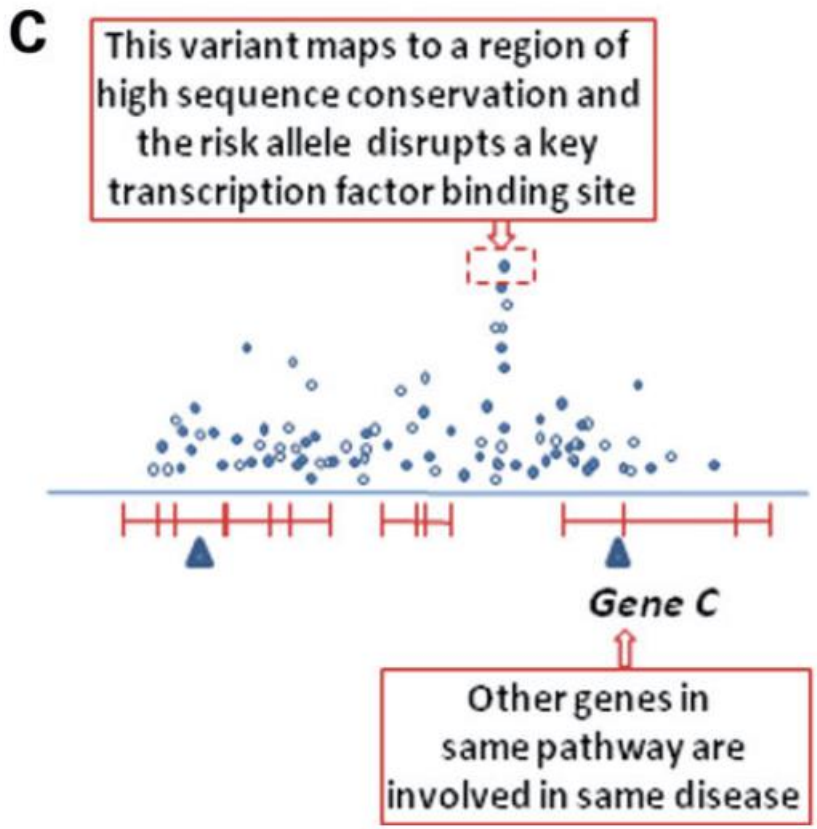
- **(A)** consider a locus at which GWA analysis (complemented by replication data—not shown) has revealed a highly significant association mapping between the coding regions of genes B and C. Directly typed SNPs are shown in the filled symbols, imputed SNPs in open symbols. Flanking recombination hotspots (blue triangles) define an interval within which the variant causal for that signal is most likely to reside. This interval contains the entire coding sequence of gene B, and portions of genes A and C. The causal variant turns out to be the typed SNP with the strongest association; it exerts its effect on disease through altering expression of gene C;



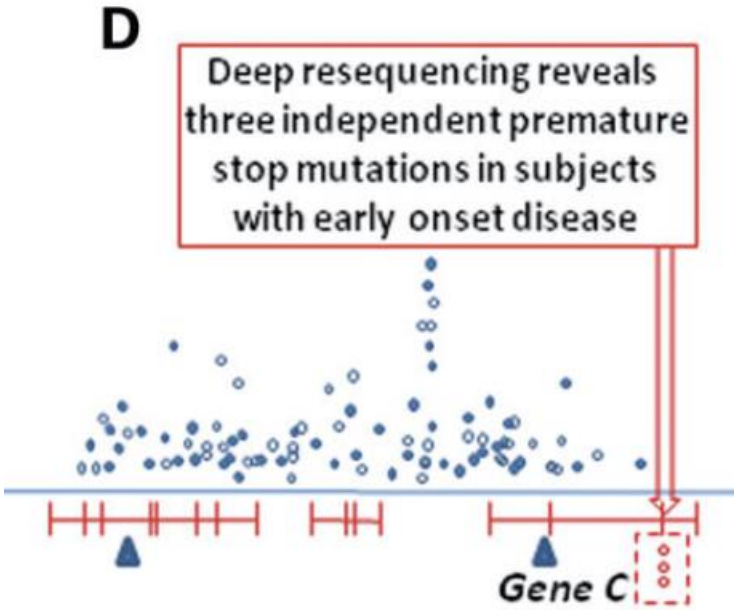
- (**B**) clues to the identity of the causal gene are derived by expression QTL studies in a tissue relevant to disease: not only is the expression of gene C associated with the same cluster of variants which shows the disease association; but there are also directionally-consistent associations between gene C transcript levels and disease state;



- **(C)** clues to the identity of the causal gene are derived from analysis of genome annotations: not only does gene C code for a member of a pathway previously implicated in the disease, but the associated variants are predicted to have strong functional credibility;



- **(D)** clues to the identity of the causal gene are derived from deep exon resequencing of genes A–C: three independent premature stop-codon mutations in gene C (predicted to lead to generation of a truncated protein product with dominant-negative effects) are found in subjects with severe, early-onset forms of the disease of interest.



What are characteristics of the hit (all) SNPs? (Manolio 2010)

- Intergenic (**) [in between genes]
- Intronic (**) [in the intronic regions within a gene]
- Synonymous [silent]
- Missense [non-synonymous which involves creation of different amino acid]
- 5' UTR [5' untranslated region on mRNA strand]
- 3' UTR
- …

**: most common!

Are there criteria for assessing the functional significance of a variant?

Criterion	Strong Support	Moderate Support	Neutral Information	Evidence Against
Nucleotide Sequence	Variant disrupts a known functional motif	missense change, disrupts putative functional motif	-	Non-functional change
Evolutionary Conservation	Strong conservation across species, multigene family	Some conservation across species or multigene family	Not known	No conservation
Population Genetics	Strong deviations from expected frequencies	Some deviations from expected frequencies	Not known	No deviations from expected frequencies
Experimental	Consistent evidence in human target tissue	Some evidence	No data available	No functional effect
Exposures	Variant affects relevant metabolism in target tissue	Variant affects metabolism	No data available	Variant does not affect metabolism
Epidemiology	Consistent and reproducible reports	Reports without replication	No data available	No association

"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 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

Background reading / supporting document to this chapter

- Anderson et al. 2010. Data quality control in genetic case-control association studies. Nat Protoc. 5(9): 1564–1573. doi:10.1038/nprot.2010.116
- Balding 2006. A tutorial on statistical methods for population association studies. Nature Reviews Genetics 5:63-70

Guiding questions are provided on the Theory Course Website.