INTRODUCTION TO GENETIC EPIDEMIOLOGY

Prof. Dr. Dr. K. Van Steen
GENOME-WIDE ASSOCIATION STUDIES

1 Setting the pace
1.a A hype about GWA studies
1.b Genetic terminology revisited
1.c Genetic association studies

2 Study Designs
2.a Marker level
2.b Subject level
2.c Gender level
3 Preliminary analyses

3.a Quality Control: Hardy-Weinberg equilibrium and missingness

3.b Linkage disequilibrium, haplotypes and SNP tagging

3.c Confounding: population stratification

4 Tests of association

4.a Single SNP

4.b Repeated single SNP tests: Multiple testing correction

4.c Replication

5 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
How much (sequence) data are available?

• The complete genome sequence of humans and of many other species provides a new starting point for understanding our basic genetic makeup and how variations in our genetic instructions result in disease.
### Table 24.1: The history of human genetics discoveries up to the 50th anniversary of the discovery of the DNA helical structure in 1953.

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1866</td>
<td>Gregor Mendel proposes basic laws of heredity based on pea plants.</td>
</tr>
<tr>
<td>April 1953</td>
<td>Francis Crick and James Watson discover double helical nature of DNA.</td>
</tr>
<tr>
<td>1977</td>
<td>Maxam, Gilbert and Sanger develop DNA sequencing.</td>
</tr>
<tr>
<td>1985</td>
<td>First use of DNA “fingerprinting” in a criminal investigation.</td>
</tr>
<tr>
<td>1992</td>
<td>US Army begins collecting blood and tissue from all new recruits to create a “genetic dog tag” program to identify soldiers killed in combat.</td>
</tr>
<tr>
<td>1999</td>
<td>USA announces a 3 year mouse genome project. First human chromosome sequenced: chromosome 22.</td>
</tr>
<tr>
<td>1882</td>
<td>Walter Fleming (embryologist) discovers tiny threads in the nuclei of cells of salamander larvae that appeared to be dividing. These later turn out to be chromosomes.</td>
</tr>
<tr>
<td>1964</td>
<td>Charles Yanofsky and colleagues prove sequence of nucleotides in DNA correspond exactly to the sequence of amino acids in proteins.</td>
</tr>
<tr>
<td>1986</td>
<td>First automated sequencer developed. Approval for first genetically engineered vaccine for humans, for hepatitis B.</td>
</tr>
<tr>
<td>1993</td>
<td>First rough map of all 23 chromosomes produced. Gene for HD cloned.</td>
</tr>
<tr>
<td>1883</td>
<td>Francis Galton coins the term eugenics referring to improving the human race.</td>
</tr>
<tr>
<td>1969</td>
<td>First gene in a piece of bacterial DNA isolated. The gene plays a role in the metabolism of sugar.</td>
</tr>
<tr>
<td>1980</td>
<td>Mapping human genome proposed using RFLPs (restriction fragment length polymorphisms).</td>
</tr>
<tr>
<td>1989</td>
<td>Creation of the National Centre for Human Genome Research (headed by James Watson) which would oversee the Human Genome Project (HGP) to map and sequence the genes in human DNA by 2005.</td>
</tr>
<tr>
<td>1995</td>
<td>H. influenzae (virus) sequenced. Microarray (CHIP) technology developed.</td>
</tr>
<tr>
<td>2000 June</td>
<td>“Working draft” of human genome sequence announced.</td>
</tr>
<tr>
<td>Year</td>
<td>Event</td>
</tr>
<tr>
<td>------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>1910</td>
<td>Thomas Morgan’s experiments with the fruit fly (Drosophila) reveal some characteristics that are sex-linked; confirms genes reside on chromosomes</td>
</tr>
<tr>
<td>1970</td>
<td>Researchers at the University of Wisconsin synthesis a gene from scratch</td>
</tr>
<tr>
<td>1982</td>
<td>First genetically engineered drug approved: insulin</td>
</tr>
<tr>
<td>1990</td>
<td>Formal launch of the HGP</td>
</tr>
<tr>
<td></td>
<td>First human gene therapy experiment performed on a 4 yr old girl with an immune deficiency</td>
</tr>
<tr>
<td>1996</td>
<td>S. cerevisiae (yeast) sequenced</td>
</tr>
<tr>
<td>2001 February</td>
<td>Publication of initial working draft of the human genome published in Science &amp; Nature by the two rival private and public groups</td>
</tr>
<tr>
<td>1926</td>
<td>US biologist Hermann Muller discovers X-rays cause genetic mutations in fruit flies</td>
</tr>
<tr>
<td>1973</td>
<td>First genetic engineering experiment: Insertion of a gene from an African clawed toad into a bacterium</td>
</tr>
<tr>
<td>1983</td>
<td>Genetic marker for the genetic condition Huntington disease (HD) located on chromosome 4</td>
</tr>
<tr>
<td>1990</td>
<td>Publication of Michael Crichton’s novel “Jurassic Park” in which bioengineered dinosaurs roam a palaeontological theme park: the experiment goes awry</td>
</tr>
<tr>
<td>1997</td>
<td>Cloning of “Dolly”</td>
</tr>
<tr>
<td>2002</td>
<td>Genome of mouse completed</td>
</tr>
<tr>
<td>1944</td>
<td>Oswald Avery, Colin McLeod &amp; Maclyn McCarthy discover DNA, not protein, is the hereditary material in most living organisms</td>
</tr>
<tr>
<td>1975</td>
<td>First call for guidelines governing genetic engineering</td>
</tr>
<tr>
<td>1985</td>
<td>Kary Mullis develops PCR (polymerase chain reaction) to rapidly reproduce DNA from a very small sample that enables genetic testing for health and other applications such as forensics and paternity testing</td>
</tr>
<tr>
<td>1991</td>
<td>First gene involved in inherited predisposition to breast cancer and ovarian cancer (BRCA1) located on chromosome 17</td>
</tr>
<tr>
<td>1998</td>
<td>C. elegans (worm) sequenced</td>
</tr>
<tr>
<td>April 25th 2003</td>
<td>Completion of the mapping of the genes in the human genome announced setting the stage for determining the function of the then estimated 30,000 or so genes</td>
</tr>
</tbody>
</table>
Introduction to Genetic Epidemiology

The different faces of quality control in genetic association studies

2005: HapMap Project Completed

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

More Information

References:


To view the PDFs on this page, you will need Adobe Reader.
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 genetic landscape in even more detail. In 2007, researchers came to appreciate the extent to which our genomes differ from person to person and the implications of this variation for deciphering the genetics of complex diseases and personal traits.

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

Pennisi 2007 Science 318:1842-3
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?
• 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))
The different faces of quality control in genetic association studies

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 1968 and 1998. The online version, OMIM, was created in 1983 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
The different faces of quality control in genetic association studies

K Van Steen
The different faces of quality control in genetic association studies

OMIM Statistics

The blue sidebar of the Online Mendelian Inheritance in Man (OMIM) home page includes a link to OMIM statistics. That shows the total number of records in the database, as well as the breakdown of the number of records in categories that correspond to the OMIM number prefixes:

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>Genes with known sequence</td>
</tr>
<tr>
<td>+</td>
<td>Genes with known sequence and phenotype</td>
</tr>
<tr>
<td>#</td>
<td>Phenotype description, molecular basis known</td>
</tr>
<tr>
<td>%</td>
<td>Mendelian phenotype or locus, molecular basis unknown</td>
</tr>
<tr>
<td>no</td>
<td>Other mainly phenotypes with suspected mendelian basis</td>
</tr>
</tbody>
</table>

RefSeq Statistics

The NCBI FTP site for RefSeq includes statistics for the current release and past releases.

Taxonomy Statistics

The NCBI Taxonomy home page includes a link to taxonomy statistics. By default, the cumulative, current statistics are shown for the number of higher taxa, genera, species, and lower taxa represented in NCBI's taxonomy database. The number of taxa that were added in any particular year can be viewed by following the link for the year of interest.

As noted in the Taxonomy database summary description in the Resource Guide, the NCBI Taxonomy Database contains the names and lineages of living and extinct organisms that are represented in the genetic databases with at least one nucleotide or protein sequence. New organisms are added to the database as sequence data are deposited for them.

The purpose of the taxonomy project at NCBI is to build a consistent phylogenetic taxonomy for the sequence databases.
The different faces of quality control in genetic association studies

K Van Steen
Published Genome-Wide Associations through 12/2010,
1212 published GWA at \( p \leq 5 \times 10^{-8} \) for 210 traits

NHGRI GWA Catalog
www.genome.gov/GWASudies
1.b Genetic terminology

What is genetic epidemiology?

“... Examining the role of genetic factors, along with the environmental contributors to disease, and at the same time giving equal attention to the differential impact of environmental agents, non-familial as well as familial, on different genetic backgrounds”

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

(Ziegler and Van Steen, Brazil 2010)
Where is the genetic information located?

- Cell has nucleus
- Nucleus carries genetic information in chromosomes
- Chromosomes composed of deoxyribonucleic 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 $\geq 2$ 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 $10^{-6} \implies$ 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*...

<table>
<thead>
<tr>
<th>M1</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSL</td>
<td>D</td>
<td>d</td>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>M2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

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*
How can individual differences be detected?

(Ziegler and Van Steen, Brazil 2010)
What are microsatellite markers?

• Synonymous: short tandem repeat, STR
• Number of repeats varies between individuals
  o 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 $\Rightarrow$ dense marker map
- Some SNPs functionally relevant $\Rightarrow$ candidate variations for disease
- SNPs more stable, i.e., lower mutation rate
- Genotyping in highly automated fashion

(Ziegler and Van Steen, Brazil 2010)
Which genotyping methods are currently being used?

<table>
<thead>
<tr>
<th>Method</th>
<th>Principle</th>
<th>Thru-put</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele-specific PCR</td>
<td>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</td>
<td>Low</td>
</tr>
<tr>
<td>RFLP analysis</td>
<td>DNA sample digested by restriction enzymes, resulting restriction fragments separated according to their lengths by gel electrophoresis</td>
<td>Low</td>
</tr>
<tr>
<td>Pyrosequencing</td>
<td>Single strand sequencing, enzymatic synthesizing of complementary strand</td>
<td>Middle</td>
</tr>
<tr>
<td>SNPstream</td>
<td>Single-base primer extension technology</td>
<td>Middle / High</td>
</tr>
</tbody>
</table>

(Ziegler and Van Steen, Brazil 2010)
Which genotyping methods are currently being used?

<table>
<thead>
<tr>
<th>Method</th>
<th>Principle</th>
<th>Thru-put</th>
</tr>
</thead>
<tbody>
<tr>
<td>TaqMan</td>
<td>Quantitative real-time PCR, allele-specific TaqMan probes</td>
<td>Middle</td>
</tr>
<tr>
<td>SNPlex</td>
<td>Oligonucleotide ligation/PCR and capillary electrophoresis</td>
<td>Middle</td>
</tr>
<tr>
<td>Affymetrix</td>
<td>Microarray based, fluorescence labeled DNA</td>
<td>Ultra-high</td>
</tr>
<tr>
<td>Illumina</td>
<td>Microarray based, fluorescence labeled DNA</td>
<td>Ultra-high</td>
</tr>
</tbody>
</table>

(Ziegler and Van Steen, Brazil 2010)
1.c Genetic association studies

What is a genome-wide association study?

- It 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.
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 GWASMail.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 Committee (DACs) Charge and Roster](#) - (07/10/2008) (PDF - 50 KB)
Introduction to Genetic Epidemiology

The different faces of quality control in genetic association studies

Genome-Wide Association Studies

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.

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 disease.
• 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)

• The impact on medical care from genome-wide association studies could potentially be substantial. Such research is laying the groundwork for the era of personalized medicine, in which the current one size-fits-all approach to medical care will give way to more customized strategies.
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 do we need to carry out a genome-wide association study?

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 of 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 assessing 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. Such findings promise to open up new avenues of research, through both the discovery of new genes related to the disease and the identification of novel pathways.

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)
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, with each chromosome shown in 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)
2 Study Designs

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

Can’t see the forest for the trees

Can’t see the trees for the forest
Does scale matter?

- Stage 1: Genotype full set of SNPs in relatively small population at liberal p value
- Stage 2: Screen second, larger population at more stringent p value
- Stage 3: Optional third stage for increased stringency
Which genetic markers to select?

The **Common Disease/Common Variant** hypothesis (CDCV)

- Continuous distribution of genetic variants, shaped by mutation and selection

(Ziegler and Van Steen, 2010)
Observations:

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

- **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 is 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.
  - 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 ...
The different faces of quality control in genetic association studies

(Glazier et al 2002)
Which genetic markers to select?

- 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
Which DNA SNPs to select?

- Costs may play a role, but a balance is needed between costs and chip performance as well as coverage (e.g., exonic regions only?)

- Some of the fundamental principles of array technology (see future class)
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)
Next generation sequencing will overtake array technology?

- The competing hypothesis to the CDCV hypothesis is the **Common Disease/Rare Variant (CDRV)** hypothesis.
- It argues that multiple rare DNA sequence variations, each with relatively high penetrance, are the major contributors to genetic susceptibility to common diseases.
- Although some common variants that underlie complex diseases have been identified, and given the recent huge financial and scientific investment in GWA studies, there is no longer a great deal of evidence in support of the CDCV hypothesis and much of it is equivocal...
- Hence, nowadays, both CDCV and CDRV hypotheses have their place in current research efforts.
Next generation sequencing will overtake array technology?

Sanger Sequencing
ABI 3730
Electrophoresis
1000 base reads

Roche GS-FLX, LifeTechnologies SOLID5500, Illumina HiSeq2500
Clonal amplification
400 base, 75 base, 100 base reads

1st Generation
1977 -

2nd Generation
2005 -
Next generation sequencing will overtake array technology?

- **IonTorrent, Proton**
  - Pacific Biosciences, RS
  - Enzyme/Readout
  - 400 base, 1000 base reads, and strobing

- **Oxford Nanopore Technologies, GridION**
  - Direct Readout
  - Very long ~10s kbase reads

---

2.5th Generation 2011 -

3rd Generation 2013 -

K Van Steen
Crucial question: How to best capture disease predisposition?

Frequency of disease-causing variants in the population

Penetrance

Monogenic Linkage

Polygenic GWAS

(Gut 2012)
## Which study subjects to select?

<table>
<thead>
<tr>
<th>Details</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Statistical analysis method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-sectional</td>
<td>Inexpensive. Provides estimate of disease prevalence</td>
<td>Few affected individuals if disease rare</td>
<td>Logistic regression, ( \chi^2 ) tests of association or linear regression</td>
</tr>
<tr>
<td>Cohort</td>
<td>Provides estimate of disease incidence</td>
<td>Expensive to follow-up. Issues with drop-out</td>
<td>Survival analysis methods</td>
</tr>
<tr>
<td>Case-control</td>
<td>No need for follow-up. Provides estimates of exposure effects</td>
<td>Requires careful selection of controls. Potential for confounding (e.g., population stratification)</td>
<td>Logistic regression, ( \chi^2 ) tests of association</td>
</tr>
<tr>
<td>Extreme values</td>
<td>Genotype individuals with extreme (high or low) values of a quantitative trait, as established from initial cross-sectional or cohort sample</td>
<td>No estimate of true genetic effect sizes</td>
<td>Linear regression, non-parametric, or permutation approaches</td>
</tr>
<tr>
<td>Case-parent triads</td>
<td>Robust to population stratification. Can estimate maternal and imprinting effects</td>
<td>Less powerful than case-control design</td>
<td>Transmission/disequilibrium test, conditional logistic regression or log-linear models</td>
</tr>
<tr>
<td>Case-parent-grandparent septets</td>
<td>Robust to population stratification. Can estimate maternal and imprinting effects</td>
<td>Grandparents rarely available</td>
<td>Log-linear models</td>
</tr>
<tr>
<td>General pedigrees</td>
<td>Higher power with large families. Sample may already exist from linkage studies</td>
<td>Expensive to genotype. Many missing individuals</td>
<td>Pedigree disequilibrium test, family-based association test, quantitative transmission/disequilibrium test</td>
</tr>
<tr>
<td>Case-only</td>
<td>Most powerful design for detection of interaction effects</td>
<td>Can only estimate interaction effects. Very sensitive to population stratification</td>
<td>Logistic regression, ( \chi^2 ) tests of association</td>
</tr>
<tr>
<td>DNA-pooling</td>
<td>Potentially inexpensive compared with individual genotyping (but technology still under development)</td>
<td>Hard to estimate different experimental sources of variance</td>
<td>Estimation of components of variance</td>
</tr>
</tbody>
</table>

(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 down to fatal, short episodes, mild cases...
  - Disadvantage V: Overestimation of risk for common disease

(Ziegler and Van Steen, 2010)
### Which study subjects to select?

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

*Morton & Collins 1998 Proc Natl Acad Sci USA 95:11389*
Which study subjects to select?

Rare versus common diseases (Lange and Laird 2006)
3 Preliminary analyses

Is there a standard file format for GWA studies?

Standard data format: tped = transposed ped format file

<table>
<thead>
<tr>
<th>FamID</th>
<th>PID</th>
<th>FID</th>
<th>MID</th>
<th>SEX</th>
<th>AFF</th>
<th>SNP1</th>
<th>SNP1</th>
<th>SNP2</th>
<th>SNP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

ped file

<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP name</th>
<th>Genetic distance</th>
<th>Chromosomal position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SNP1</td>
<td>0</td>
<td>123456</td>
</tr>
<tr>
<td>1</td>
<td>SNP2</td>
<td>0</td>
<td>123654</td>
</tr>
</tbody>
</table>

map file
Is there a standard file format for GWA studies?

<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP</th>
<th>Gen. dist.</th>
<th>Pos</th>
<th>PID 1</th>
<th>PID 2</th>
<th>PID 3</th>
<th>PID 4</th>
<th>PID 5</th>
<th>PID 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SNP1</td>
<td>0</td>
<td>123456</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>1</td>
<td>SNP2</td>
<td>0</td>
<td>123654</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>T</td>
</tr>
</tbody>
</table>

tfam file: First 6 columns of standard ped file

<table>
<thead>
<tr>
<th>FamID</th>
<th>PID</th>
<th>FID</th>
<th>MID</th>
<th>SEX</th>
<th>AFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

tped 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
What are signs of deviations from HWE?

Decreased or increased HET

<table>
<thead>
<tr>
<th>Decrease in HET caused by</th>
<th>Increase in HET caused by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection against heterozygotes</td>
<td>Selection favoring heterozygotes</td>
</tr>
<tr>
<td>Inbreeding</td>
<td>Outbreeding</td>
</tr>
<tr>
<td>Positive assortative mating</td>
<td>Negative assortative mating</td>
</tr>
<tr>
<td>Null allele</td>
<td>Copy number variation</td>
</tr>
<tr>
<td>Wahlund effect</td>
<td>Amplification artifact of new alleles</td>
</tr>
<tr>
<td>Allele dropout in old samples</td>
<td>Misclassification of alleles at different loci in multigene families</td>
</tr>
</tbody>
</table>

Ziegler & König 2010 ISBN-13 978-3-527-32389-0
Hedrick 2009 9780763757373
What are signs of deviations from HWE?

$F_{IT}$ is the inbreeding coefficient of an individual (I) relative to the total (T) population, as above; $F_{IS}$ is the inbreeding coefficient of an individual (I) relative to the subpopulation (S), using the above for subpopulations and averaging them; and $F_{ST}$ is the effect of subpopulations (S) compared to the total population (T)
What are signs of deviations from HWE?

Increased HOM (e.g., in case of population stratification; Wahlund effect)
How can HWE be measured?

- Disequilibrium coefficient: $D_{A_1} = p_{11} - p^2 = p_{22} - q^2 = -p_{12} + 2pq$
- Inbreeding coefficient:
  - Assume $P(A_1) = p$, probability of 2\textsuperscript{nd} allele to be identical
  - Prob. of two $A_1$ alleles equal to $p \cdot f$
  - Prob. for two independent $A_1$ alleles $p^2$
  - Ergo: $P(A_1A_1) = p^2(1 - f) + pf = p^2 + fpq$
    $$P(A_1A_2) = 2pq - 2fpq = 2pq(1 - f)$$
- Excess heterozygosity: $\gamma = p_{12}/\left(2\sqrt{p_{11}p_{22}}\right)$

- Standard procedure in GWA studies: $\chi^2$ lack of fit test
Introduction to Genetic Epidemiology

The different faces of quality control in genetic association studies

How can HWE be measured?

• The Pearson test is easy to compute, but the $\chi^2$ approximation can be poor when there are low genotype counts, in which case it is better to use a Fisher exact test, which does not rely on the $\chi^2$ approximation.
• 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 (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 HWE be measured?

- 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 
    \[-\log \left( \frac{i}{L+1} \right)\], where *L* is the number of SNPs.
- The 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% 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 HWE be measured?

(Balding 2006)
Why is cluster plot reading important?

**dbGaP study phs000001: rs203674 Genotype Summary**

<table>
<thead>
<tr>
<th></th>
<th>Genotype</th>
<th>Genotype Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA</td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>AC</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>75%</td>
</tr>
<tr>
<td>Case</td>
<td>51</td>
<td>38.8%</td>
</tr>
<tr>
<td>Control</td>
<td>82</td>
<td>45.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Allele</th>
<th>Allele Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>75%</td>
</tr>
<tr>
<td>Case</td>
<td>259</td>
<td>67.1%</td>
</tr>
<tr>
<td>Control</td>
<td>252</td>
<td>33.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Number of Samples</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genotyped</td>
<td>Total</td>
</tr>
<tr>
<td>Case</td>
<td>394</td>
<td>395</td>
</tr>
<tr>
<td>Control</td>
<td>193</td>
<td>198</td>
</tr>
</tbody>
</table>

Case pHWE: 0.053
Control pHWE: 1.000
Odds ratio of minor allele 'A': 0.260
Chi-square: 109.447
p-value of Chi-square test: 1.29e-25

**AREDS (Illumina 100K) Normalized Allele Intensity for SNP rs203674 (623 individual samples)**
What are standard filters on the gender level?

- Absolute difference in call fractions for males and females
- Proportion of heterozygotes in males and females in all samples
- Missing data by gender
- Test of allelic association by gender among controls

(Ziegler and Van Steen 2010)
Is there a power advantage in imputing?

During the late stages of the Human Genome Project and in the years thereafter, several important private and public spin-off projects have been launched. These projects are capitalizing on the new biology and new technologies brought about by the HGP that have enabled their research to go forward. They include:

**Genomic Science Program**

http://genomicsgtr.energy.gov/

GSP research conducts explorations of microbes and plants at the molecular, cellular, and community levels. The goal is to gain insights about fundamental biological processes and, ultimately, a predictive understanding of how living systems operate. The resulting knowledgebase—all linked through DNA sequence and freely available—will catalyze the translation of science into new technologies for energy and environmental applications.

**1000 Genomes Project**

http://www.1000genomes.org/page.php

The 1000 Genomes Project is sequencing the genomes of at least a thousand people from around the world. The project is funded by Wellcome Trust Sanger Institute in Hinxton, England; the Beijing Genomics Institute Shenzhen in China; and the US NIH National Human Genome Research Institute. The project's goal is to develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

**Human Microbiome Project**
Is there a power advantage in imputing?

1000 Genomes Project data available on Amazon Cloud

Project is exemplar of new White House big data initiative

Bethesda, Md., Thurs., March 29, 2012 — The world’s largest set of data on human genetic variation — produced by the international 1000 Genomes Project — is now publicly available on the Amazon Web Services (AWS) cloud, the National Institutes of Health (NIH) and AWS jointly announced today.

The public-private collaboration demonstrates the kind of solutions that may emerge from the Big Data Research and Development Initiative announced today by the White House Office of Science and Technology Policy (OSTP) during an event at the American Association for the Advancement of Science in Washington, D.C.

“The explosion of biomedical data has already significantly advanced our understanding of health and disease. Now we want to find new and better ways to make the most of these data to speed discovery, innovation and improvements in the nation’s health and economy,” said NIH Director Francis S. Collins, M.D., Ph.D. Dr. Collins is among agency leaders speaking in support of the initiative at the launch event.

The Big Data initiative will initially engage at least six federal science agencies — including the NIH, the National Science Foundation, and the Department of Defense and the Department of Energy — committing more than $200 million to a collaborative effort to develop core technologies and other resources needed by researchers to manage and analyze enormous data sets.

Among the NIH components participating in the Big Data initiative are the National Human Genome Research Institute (NHGRI) and the
Is there a power advantage in imputing?
Is there a power advantage in imputing? (Spencer et al 2009)
What are the Travemünde criteria?

<table>
<thead>
<tr>
<th>Level</th>
<th>Filter criterion</th>
<th>Standard value for filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample level</td>
<td>Call fraction</td>
<td>$\geq 97%$</td>
</tr>
<tr>
<td></td>
<td>Cryptic relatedness</td>
<td>Study specific</td>
</tr>
<tr>
<td></td>
<td>Ethnic origin</td>
<td>Study specific; visual inspection of principal components</td>
</tr>
<tr>
<td></td>
<td>Heterozygosity</td>
<td>Mean $\pm 3$ std.dev. over all samples</td>
</tr>
<tr>
<td></td>
<td>Heterozygosity by gender</td>
<td>Mean $\pm 3$ std.dev. within gender group</td>
</tr>
<tr>
<td>SNP level</td>
<td>MAF</td>
<td>$\geq 1%$</td>
</tr>
<tr>
<td></td>
<td>MiF</td>
<td>$\leq 2%$ in any study group, e.g., in both cases and controls</td>
</tr>
<tr>
<td></td>
<td>MiF by gender</td>
<td>$\leq 2%$ in any gender</td>
</tr>
<tr>
<td></td>
<td>HWE</td>
<td>$p &lt; 10^{-4}$</td>
</tr>
</tbody>
</table>

(Ziegler 2009)
What are the Travemünde criteria?

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

(Ziegler 2009)
3.b Linkage disequilibrium, haplotypes 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Providence</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New York</td>
<td>210</td>
<td>152</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philadelphia</td>
<td>320</td>
<td>237</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baltimore</td>
<td>430</td>
<td>325</td>
<td>173</td>
<td>87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Washington</td>
<td>450</td>
<td>358</td>
<td>206</td>
<td>120</td>
<td>34</td>
<td></td>
</tr>
</tbody>
</table>
Distances among cities
Distances among SNPs

- If a causal polymorphism is not genotyped, we can still hope to detect its effects through **Linkage Disequilibrium (LD)** with polymorphisms that are typed (key principle behind doing genetic association analysis ...).
- 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. In general, LD is taken to be a measure of allelic association.
- Among the measures that have been proposed for two-locus haplotype data, the two most important are $D'$ (Lewontin’s $D$ prime) and $r^2$ (the square correlation coefficient between the two loci under study).
- 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)
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: $p_{AB} - p_A p_B$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>$p_{AB}$</td>
<td>$p_{aB}$</td>
</tr>
<tr>
<td>b</td>
<td>$p_{Ab}$</td>
<td>$p_{ab}$</td>
</tr>
</tbody>
</table>

- $D'$ is the absolute ratio of D compared with its maximum value.
- $D' = 1$ : complete LD

- $R^2$ is the statistical correlation of two markers :
  - When $R^2 = 1$, knowing the genotypes of alleles of one SNP is directly predictive of genotype of another SNP

$$R^2 = \frac{D^2}{P(A)P(a)P(B)P(b)}$$
How far does linkage disequilibrium extend?

(Hecker et al 2003)
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).

- 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).

- 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 to interpret LD data?

• Therefore, the key concept in a (population-based) genetic association study is linkage disequilibrium.
How to interpret LD data?

- It gives the rational for performing genetic association studies

**Phenotype:** The visible or measurable (expressed) characteristic of an organism

**Trait:** Coded phenotype
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?

(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 methods
  - 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 $\lambda$ 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 $\lambda > 1$ the test statistics are divided by $\lambda$.

- 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 $\lambda$ 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)
Introduction to Genetic Epidemiology

The different faces of quality control in genetic association studies

K Van Steen
Does the same hold on a "global" (world) level?

(Paschau 2007)
4 Tests of association

What is the causal model underlying genetic association?

(Ziegler and Van Steen 2010)
4.a Single SNP

What are common association tests (dichotomous traits)?

**Observed genotype** frequencies and theoretical probabilities

<table>
<thead>
<tr>
<th></th>
<th>aa</th>
<th>aA</th>
<th>AA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$r_0$</td>
<td>$(p_{0,a})$</td>
<td>$r_1$</td>
<td>$(p_{1,a})$</td>
</tr>
<tr>
<td>Controls</td>
<td>$s_0$</td>
<td>$(p_{0,u})$</td>
<td>$s_1$</td>
<td>$(p_{1,u})$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_0$</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

**Observed allele** frequencies and theoretical probabilities

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>A</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$2r_0 + r_1$</td>
<td>$2r_2 + r_1$</td>
<td>$(p_{A,a})$</td>
</tr>
<tr>
<td>Controls</td>
<td>$2s_0 + s_1$</td>
<td>$2s_2 + s_1$</td>
<td>$(p_{A,u})$</td>
</tr>
<tr>
<td>Total</td>
<td>$2n_0 + n_1$</td>
<td>$2n_2 + n_1$</td>
<td>$2n$</td>
</tr>
</tbody>
</table>

(Ziegler and Van Steen 2010)
What are common association tests (dichotomous traits)?

**Standard allele test:**

- $\chi^2$ test of independence
- Equivalent to
  \[
  \chi^2_A = 2n \cdot \frac{[(2r_0 + r_1)(2s_2 + s_1) - (2r_2 + r_1)(2s_0 + s_1)]^2}{2r \cdot 2s \cdot (2n_0 + n_1) \cdot (2n_2 + n_1)}
  \]
- Asymptotically $\chi^2$ with 1 degree of freedom (d.f.)

**Standard genotype test:**

- $\chi^2$ test of independence
- Asymptotically $\chi^2$ with 2 d.f.

(Ziegler and Van Steen 2010)
What are common association tests (dichotomous traits)?

Penetrances for simple Mendelian inheritance patterns

<table>
<thead>
<tr>
<th>Genotype</th>
<th>General</th>
<th>Recessive</th>
<th>Dominant</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>$f_0$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ND</td>
<td>$f_1$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>$f_2$</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Trait T**: coded phenotype
- **Penetrance**: $P(T|\text{Genotype})$
- **Complete penetrance**: $P(T|DD) = 1$ (simplified definition)
What are common association tests (dichotomous traits)?

<table>
<thead>
<tr>
<th></th>
<th>Dominant</th>
<th>Heterozygote</th>
<th>Recessive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>aa</td>
<td>aA or AA</td>
<td>aa or aA</td>
</tr>
<tr>
<td>Cases</td>
<td>r_0</td>
<td>r_1 + r_2</td>
<td>r_0 + r_2</td>
</tr>
<tr>
<td>Controls</td>
<td>s_0</td>
<td>s_1 + s_2</td>
<td>s_0 + s_2</td>
</tr>
<tr>
<td>Total</td>
<td>n_0</td>
<td>n_2</td>
<td>n_0 + n_2</td>
</tr>
</tbody>
</table>

\[
\chi^2_{dom} = n \cdot \frac{(r_0(s_1 + s_2) - (r_1 + r_2)s_0)^2}{r \cdot s \cdot n_0 \cdot (n_1 + n_2)}
\]

\[
\chi^2_{het} = n \cdot \frac{(r_1(s_0 + s_2) - (r_0 + r_2)s_1)^2}{r \cdot s \cdot n_1 \cdot (n_0 + n_2)}
\]

\[
\chi^2_{rec} = n \cdot \frac{(r_0 + r_1)s_2 - r_2(s_0 + s_1)^2}{r \cdot s \cdot (n_0 + n_1) \cdot n_2}
\]
What are common association tests (dichotomous traits)?

- The Cochran-Armitage trend test measures a linear trend in proportions weighted by general measure of exposure dosage: variable x in regression model = #alleles

\[ \chi^2_{trend} = \frac{n}{rs} \cdot \frac{(2r_2s - 2r_1s + r_1s - s_1r)^2}{2n_2n + (2n_2 + n_1)(n_0 - n_2)} \]

- Max test: computes maximum over standardized tests for different genetic models, providing a global test
Which test should be used in applications?

- Trend test if no biological hypothesis
- Trend test optimal if additive genetic model
- Dom test optimal if dominant genetic model
- Rec test optimal if recessive genetic model
- Trend test identical to allele test if HWE exactly fulfilled
- Asymptotic version of Max test alternative to trend test

Hothorn & Hothorn 2009 Biom J
How are genetic effects measured?

<table>
<thead>
<tr>
<th></th>
<th>G = 1</th>
<th>G = 0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$x_1$</td>
<td>$y_1$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>Controls</td>
<td>$x_0$</td>
<td>$y_0$</td>
<td>$n_0$</td>
</tr>
</tbody>
</table>

Case control study:

- Odds ratio: $\hat{OR}_G = \frac{x_1y_0}{y_1x_0}$

- Attributable risk in variant carriers (... in the exposed):

$$AR_G = \frac{P(\text{aff}|G = 1) - P(\text{aff}|G = 0)}{P(\text{aff}|G = 1)} = \frac{p_1 - p_0}{p_1} = \frac{RR - 1}{RR} \approx \frac{RR - 1}{RR}$$

**RR being**

$$\frac{P(\text{aff}|G = 1)}{P(\text{aff}|G = 0)} = \frac{p_1}{p_0}$$
Which odds ratios (measures of effect) can we expect?

(A and B) Histograms of susceptibility allele frequency and MAF, respectively, at confirmed susceptibility loci.  

(Iles 2008)
Which odds ratios (measures of effect) can we expect?

(C) Histogram of estimated ORs (estimate of genetic effect size) at confirmed susceptibility loci.  
(Iles 2008)
4.b Repeated single SNP tests

The regression framework

- Regression analysis is used for explaining or modeling the relationship between a single variable $Y$, called the response, output or dependent variable, and one or more predictor, input, independent or explanatory variables, $X_1, \ldots, X_m$.
- When $m=1$ it is called simple regression but when $m > 1$ it is called multiple regression or sometimes multivariate regression.
- When there is more than one $Y$, then it is called multivariate multiple regression.
- The basic syntax for doing regression in R is `lm(Y~model)` to fit linear models and `glm()` to fit generalized linear models (e.g. logistic regression models in the “dichotomous trait” setting before). Next slide: syntax!
### Syntax vs. Model

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

Some data; cholesterol levels plotted by genotype (single SNP)
Use of `lm()` in genetics

Additive model (the most commonly used)
Use of `lm()` in genetics

Dominant model (best fit to this data)
Use of `lm()` in genetics

Recessive model (least stable for rare aa)
Use of `lm()` in genetics

2 parameter model (robust but can be overkill)
### Use of `glm()` in genetics

**Logistic regression** is the ‘default’ analysis for **binary outcomes**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Type</th>
<th>Regression</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol</td>
<td>Continuous</td>
<td>Linear</td>
<td>Difference in Outcome</td>
</tr>
<tr>
<td>Blood Pressure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death</td>
<td>Binary</td>
<td>Logistic</td>
<td>Ratio of odds</td>
</tr>
<tr>
<td>Stroke</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI&gt;30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Use of `glm()` in genetics

Odds are a [gambling-friendly] measure of chance;

![Diagram showing different survival probabilities for AA, Aa, and aa genotypes.](image)
Use of `glm()` in genetics

Odds are a [gambling-friendly] measure of chance;

![Graph showing survival and death probabilities](graph.png)
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

(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)**
- 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 $j$
  \[ H_{0j} : \text{no association between SNP } j \text{ and the trait} \]

• 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).
- 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]
Introduction to Genetic Epidemiology

The different faces of quality control in genetic association studies

4. c Replication

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

J. Hirschhorn

Perspective

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

Benjamin M. Neale1 and Pak C. Sham1,2

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
- SNPs most likely to replicate:
  - Showing modest to strong statistical significance
  - Having common minor allele frequency
  - Exhibiting modest to strong genetic effect size

- Note: Multi-stage design analysis results should not be seen as “evidence for replication” ...
**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 Interpretation and follow-up

What have GWA studies learnt us about functionality? (Manolio 2010)
What have GWA studies learnt us about functionality?  
(Rebbeck et al 2004)
Are there criteria for assessing the functional significance of a variant?

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Strong Support</th>
<th>Moderate Support</th>
<th>Neutral Information</th>
<th>Evidence Against</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide Sequence</td>
<td>Variant disrupts a known functional motif</td>
<td>missense change, disrupts putative functional motif</td>
<td>-</td>
<td>Non-functional change</td>
</tr>
<tr>
<td>Evolutionary Conservation</td>
<td>Strong conservation across species, multigene family</td>
<td>Some conservation across species or multigene family</td>
<td>Not known</td>
<td>No conservation</td>
</tr>
<tr>
<td>Population Genetics</td>
<td>Strong deviations from expected frequencies</td>
<td>Some deviations from expected frequencies</td>
<td>Not known</td>
<td>No deviations from expected frequencies</td>
</tr>
<tr>
<td>Experimental</td>
<td>Consistent evidence in human target tissue</td>
<td>Some evidence</td>
<td>No data available</td>
<td>No functional effect</td>
</tr>
<tr>
<td>Exposures</td>
<td>Variant affects relevant metabolism in target tissue</td>
<td>Variant affects metabolism</td>
<td>No data available</td>
<td>Variant does not affect metabolism</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>Consistent and reproducible reports</td>
<td>Reports without replication</td>
<td>No data available</td>
<td>No association</td>
</tr>
</tbody>
</table>
“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”
- Rebbeck et al 2004. Assessing the function of genetic variants in candidate gene association studies 5: 589-