INTRODUCTION TO GENETIC EPIDEMIOLOGY

(EPID0754)

Prof. Dr. Dr. K. Van Steen

CHAPTER 5: POPULATION BASED ASSOCIATION STUDIES

1 Introduction

1.a Human complex diseases

1.b Genetic association studies

2 Preliminary analyses

2.a Hardy-Weinberg equilibrium

2.b Missing genotype data

2.c Haplotype and genotype data

2.d Measures of LD and estimates of recombination rates

2.e SNP tagging

3 Tests of association: single SNP

4 Tests of association: multiple SNPs

5 Dealing with population stratification

5.a Spurious associations

5.b Genomic control

5.c Structured association methods

5.d Other approaches

6 Multiple testing

6.a General setting

6.b Controlling the type I error

7 Assessing the function of genetic variants

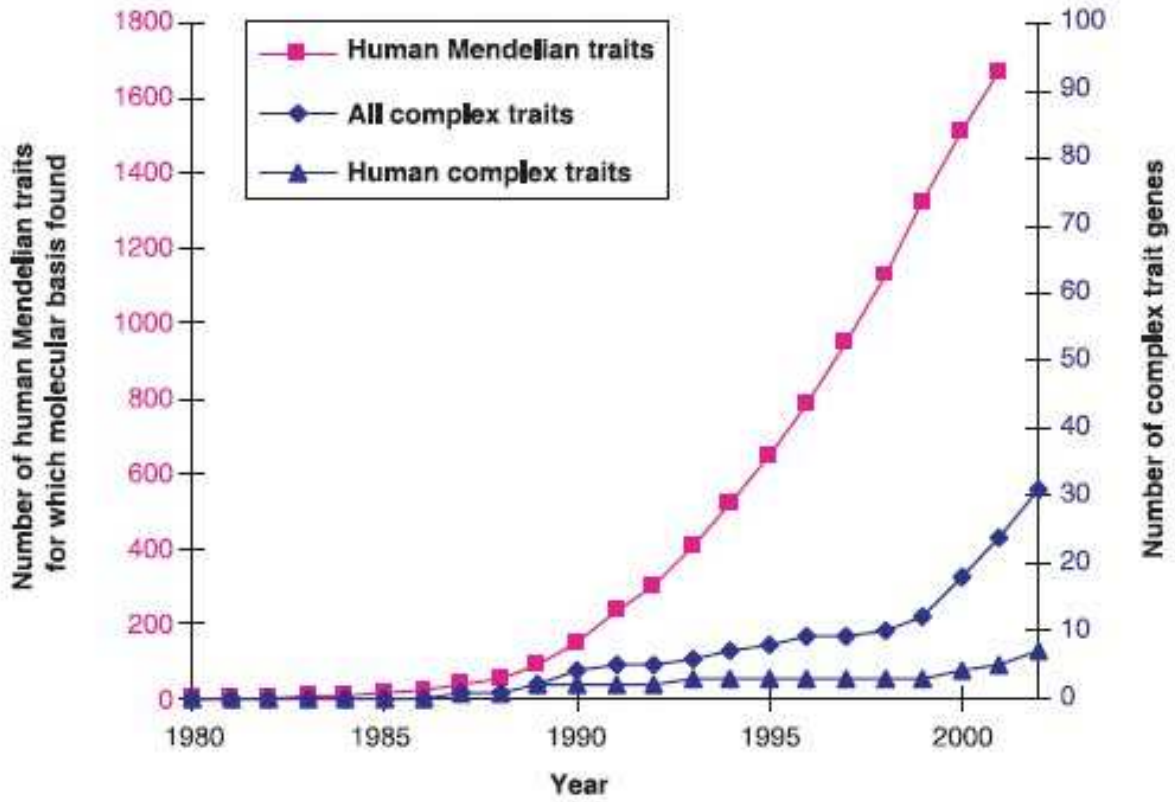
8 Proof of concept

1 Introduction

1.a Human complex diseases

Terminology

- **Complex disease**: Condition 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.



Identification of genes underlying human Mendelian traits and genetically complex traits in humans and other species. Cumulative data for human Mendelian trait genes (to 2001) include all major genes causing a Mendelian disorder in which causal variants have been identified (*58*, *59*). This reflects mutations in a total of 1336 genes. Complex trait genes were identified by the whole-genome screen approach and denote cumulative year-on-year data described in this review.

(Glazier et al 2002)

Introduction

- With the availability of the human genome sequence and those of an increasing number of other species, sequence-based gene discovery is complementing and will eventually replace map-based gene discovery.
- These and other recent developments in the field have caused a paradigm shift in biomedical research:

Structural genomics	→	Functional genomics
Genomics	→	Proteomics
Map-based gene discovery	→	Sequence-based gene discovery
Monogenic disorders	→	Multifactorial disorders
Specific DNA diagnosis	→	Monitoring of susceptibility
Analysis of one gene	→	Analysis of multiple genes in gene families, pathways, or systems
Gene action	→	Gene regulation
Etiology (specific mutation)	→	Pathogenesis (mechanism)
One species	→	Several species

Introduction

- Initial analyses of the completed chromosomal sequences suggest that the number of human genes is lower than expected.
- These findings are consistent with the idea that variations in gene regulation and the splicing of gene transcripts explain how one protein can have distinct functions in different types of tissue.
- At the beginning of the 21st century, it also seemed likely that obvious deleterious mutations in the coding sequences of genes are responsible for only a fraction of the differences in disease susceptibility between individuals, and that sequence variants affecting gene splicing and regulation must play an important part in determining disease susceptibility.

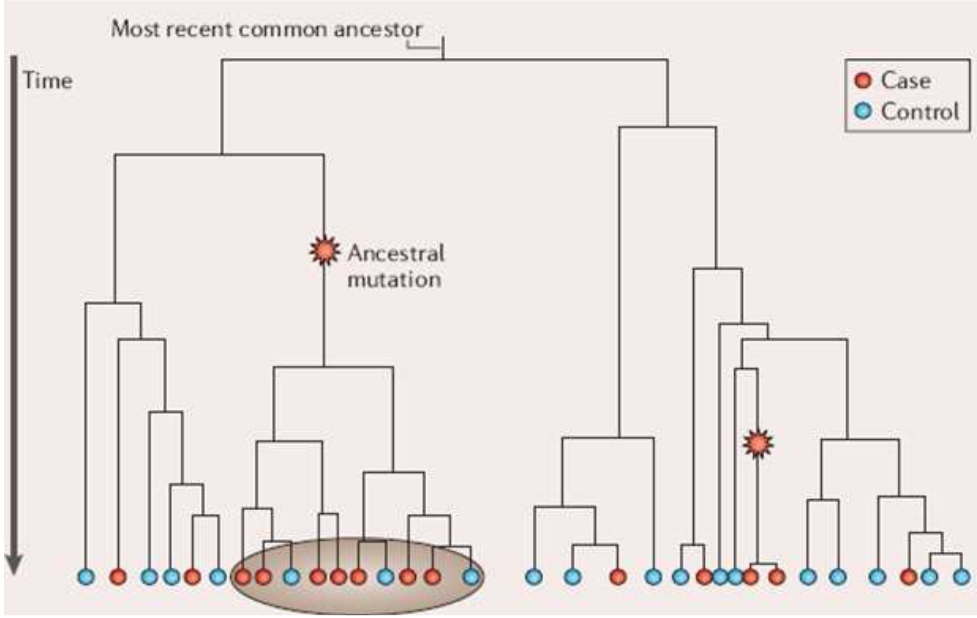
Introduction

- As only a small proportion of the millions of sequence variations in our genomes will have such functional impacts, identifying this subset of sequence variants is a challenging task.
- The success of global efforts to identify and annotate sequence variations in the human genome, which are called single-nucleotide polymorphisms (SNPs), is reflected in the abundance of SNP databases
  - www.ncbi.nlm.nih.gov/SNP,
  - http://snp.cshl.org,
  - http://hgbase.cgr.ki.se.
- However, the follow-up work of understanding how these and other genetic variations regulate the phenotypes (visual characteristics) of human cells, tissues, and organs will occupy biomedical researchers for all of the 21st century

1.b Population-based genetic association studies

Introduction

- The goal of population association studies is to identify patterns of polymorphisms that vary systematically between individuals with different disease states and could therefore represent the effects of risk-enhancing or protective alleles.



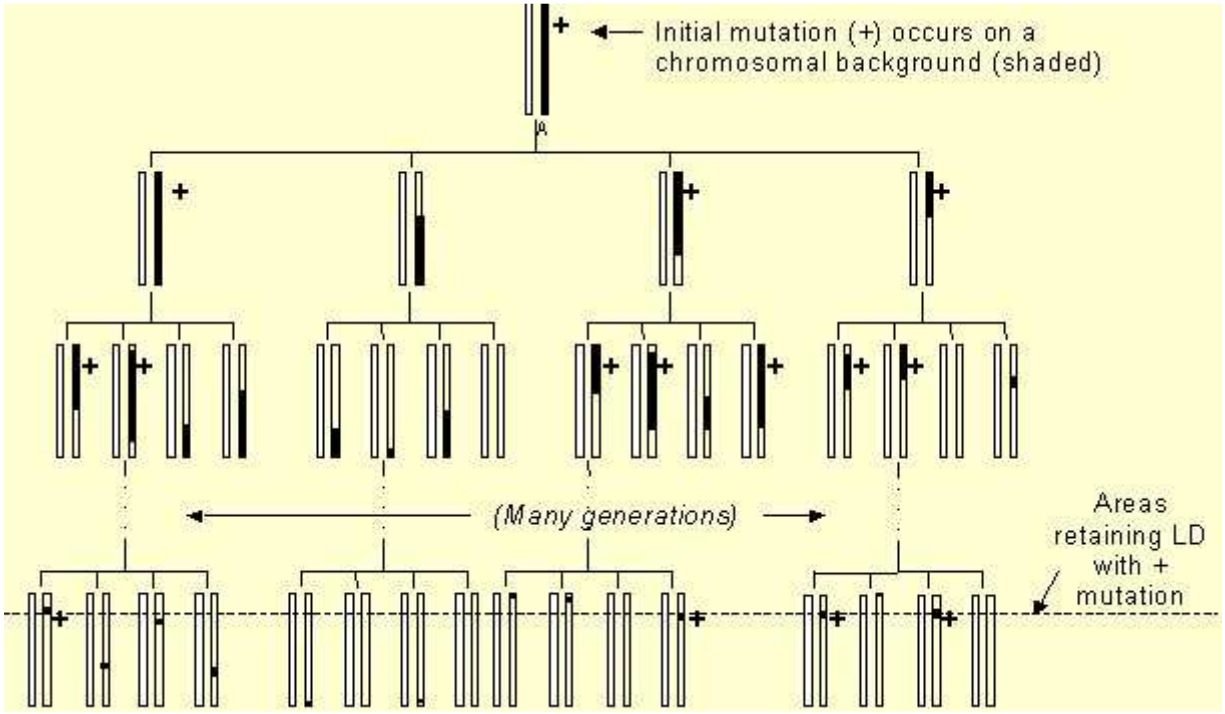
(Balding 2006)

Introduction

- When performing a genetic association study, there are a number of pitfalls one should be aware of.
- Perhaps the most crucial one is related to the realization that some patterns may arise simply by chance.
- To distinguish between true and chance effects, there are two 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 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)

Introduction

- Hence, the key concept in a (population-based) genetic association study is linkage disequilibrium.



- This gives the rational for performing genetic association studies

Types of genetic association studies

- Candidate polymorphism
  - These studies focus on an individual polymorphism that is suspected of being implicated in disease causation.
- Candidate gene
  - These studies might involve typing 5–50 SNPs within a gene (defined to include coding sequence and flanking regions, and perhaps including splice or regulatory sites).
  - The gene can be either a positional candidate that results from a prior linkage study, or a functional candidate that is based, for example, on homology with a gene of known function in a model species.

Types of genetic association studies

- Fine mapping
  - Often refers to studies that are conducted in a candidate region of perhaps 1–10 Mb and might involve several hundred SNPs.
  - The candidate region might have been identified by a linkage study and contain perhaps 5–50 genes.
- Genome-wide
  - These seek to identify common causal variants throughout the genome, and require ≥300,000 well-chosen SNPs (more are typically needed in African populations because of greater genetic diversity).
  - The typing of this many markers has become possible because of the International HapMap Project and advances in high-throughput genotyping technology

Types of population association studies

- The aforementioned classifications are not precise: some candidate-gene studies involve many hundreds of genes and are similar to genome-wide scans.
- Typically, a causal variant will not be typed in the study, possibly because it is not a SNP (it might be an insertion or deletion, inversion, or copy-number polymorphism).
- Nevertheless, a well-designed study will have a good chance of including one or more SNPs that are in strong linkage disequilibrium with a common causal variant.

Analysis of population association studies

- Statistical methods that are used in pharmacogenetics are similar to those for disease studies, but the phenotype of interest is drug response (efficacy and/or adverse side effects).
- In addition, pharmacogenetic studies might be prospective whereas disease studies are typically retrospective.
- Prospective studies are generally preferred by epidemiologists, and despite their high cost and long duration some large, prospective cohort studies are currently underway for rare diseases.
- Often a case–control analysis of genotype data is embedded within these studies, so many of the statistical analyses that are discussed in this chapter can apply both to retrospective and prospective studies.
- However, specialized statistical methods for time-to-event data might be required to analyse prospective studies.

Analysis of population association studies

- Design issues guide the analysis methods to choose from:

	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)

Analysis of population association studies

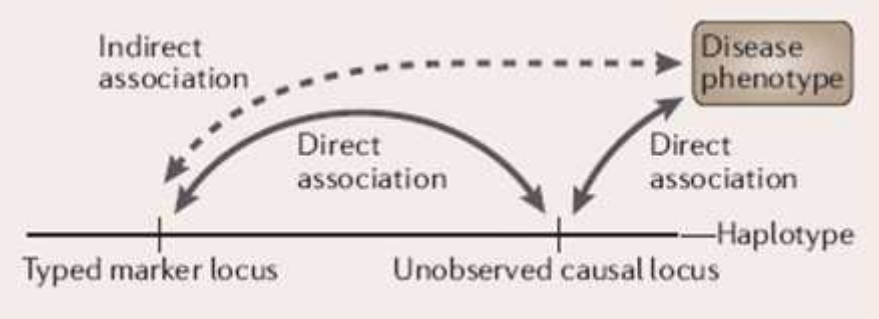
- The design of a genetic association study may refer to
  - subject design (see before)
  - marker design:
    - Which markers are most informative? Microsatellites? SNPs? CNVs?
    - Which platform is the most promising?
  - study scale:
    - Genome-wide
    - Genomic

Analysis of population association studies

- Marker design
  - Recombinations that have occurred since the most recent common ancestor of the group at the locus can break down associations of phenotype with all but the most tightly linked marker alleles.
  - This permits fine mapping if marker density is sufficiently high (say, ≥1 marker per 10 kb).
  - When the mutation entered into the population a long time ago, then a lot of recombination processes may have occurred, and hence the haplotype harboring the disease mutation may be very small.
    - This favors typing a lot of markers and generating dense maps
    - The drawback is the computational and statistical burden involved with analyzing such huge data sets.

Analysis of population association studies

- Direct versus indirect associations
  - The two direct associations that are indicated in the figure below, between a typed marker locus and the unobserved causal locus, cannot be observed, but if r^2 (a measure of allelic association) between the two loci is high then we might be able to detect the indirect association between marker locus and disease phenotype.



Analysis of population association studies

- Scale of genetic association studies

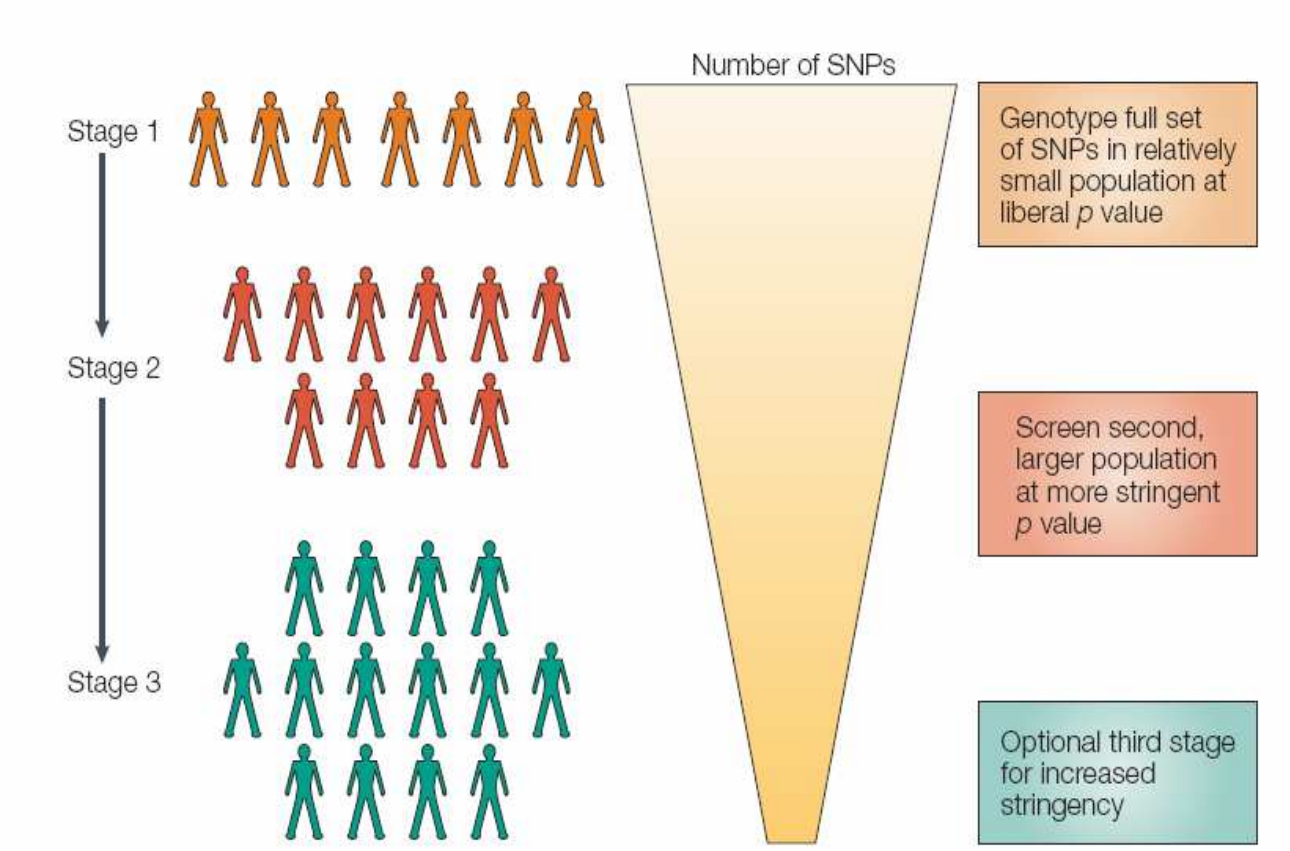
candidate gene approach

vs

genome-wide screening approach

Analysis of population association studies

- Scale of genetic association studies: multi-stage designs



Power of genetic association studies

- Broadly speaking, association studies are sufficiently powerful only for common causal variants.
- The threshold for *common* depends on sample and effect sizes as well as marker frequencies, but as a rough guide the minor-allele frequency might need to be above 5%.
- The *common disease / common variant* (CDCV) hypothesis argues that genetic variations with appreciable frequency in the population at large, but relatively low 'penetrance' (or the probability that a carrier of the relevant variants will express the disease), are the major contributors to genetic susceptibility to common diseases.

Motivation and consequence of CDCV

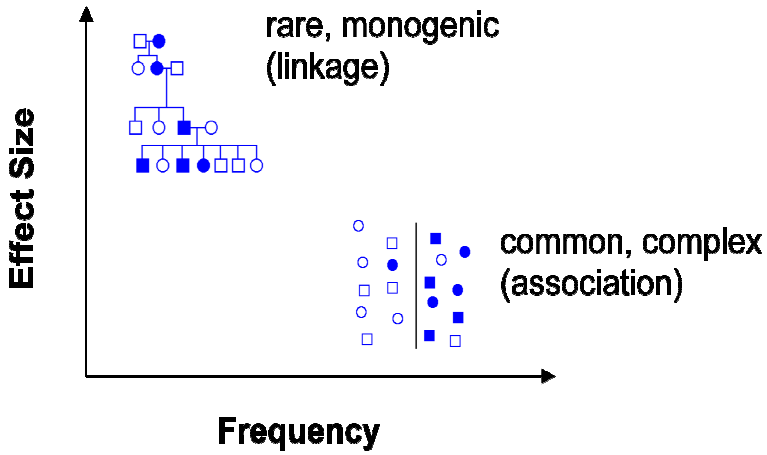
- If multiple rare genetic variants were the primary cause of common complex disease, association studies would have little power to detect them; particularly if allelic heterogeneity existed.
- The major proponents of the CDCV were the movers and shakers behind the HapMap and large-scale association studies: When this hypothesis is true, then we may be able to characterize the variation using a block like structure of common haplotyopes

Motivation and consequence of CDRV

- The competing hypothesis is cleverly 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.
- This may be the case that should expect extensive alleles or loci are interacting (Pritchard, 2001).
- Although some common variants that underlie complex diseases have been identified, and given the recent huge financial and scientific investment in GWA, there is no longer a great deal of evidence in support of the CDCV hypothesis and much of it is equivocal...
- Both CDCV and CDRV hypotheses have their place in current research efforts.

The role of genetic association studies in complex disease analysis

Which gene hunting method is most likely to give success?



- Monogenic "Mendelian" diseases
  - Rare disease
  - Rare variants
    - Highly penetrant
- Complex diseases
  - Rare/common disease
  - Rare/common variants
    - Variable penetrance

(Slide: courtesy of Matt McQueen)

Factors influencing consistency of gene-disease associations

- Variables affecting inferences from experimental studies:
  - In vitro or in vivo system studied
  - Cell type studied
  - Cultured versus fresh cells studied
  - Genetic background of the system
  - DNA constructs
  - DNA segments that are included in functional (for example, expression) constructs
  - Use of additional promoter or enhancer elements
  - Exposures
  - Use of compounds that induce or repress expression
  - Influence of diet or other exposures on animal studies

(Rebbeck et al 2004)

Factors influencing consistency of gene-disease associations

- Variables affecting epidemiological inferences:
  - Inclusion/exclusion criteria for study subject selection
  - Sample size and statistical power
  - Candidate gene choice
  - A biologically plausible candidate gene
  - Functional relevance of the candidate genetic variant
  - Frequency of allelic variant
  - Statistical analysis
  - Consideration of confounding variables, including ethincity, gender or age.
  - Whether an appropriate statistical model was applied (for example, were interactions
  - considered in addition to main effects of genes?)
  - Violation of model assumptions

(Rebbeck et al 2004)

2 Preliminary analyses

2.a Introduction

2.b Hardy-Weinberg equilibrium

2.c Missing genotype data

2.d Haplotype and genotype data

Measures of LD and estimates of recombination rates

2.e SNP tagging

2.a Introduction

- Pre-analysis techniques often performed include:
  - testing for Hardy–Weinberg equilibrium (HWE)
  - strategies to select a good subset of the available SNPs (‘tag’ SNPs)
  - inferring haplotypes from genotypes.
- Data quality is of paramount importance, and data should be checked thoroughly before other analyses are started.
- Data should be checked for
  - batch or study-centre effects,
  - for unusual patterns of missing data,
  - for genotyping errors.

Introduction

- Recall that genotype data are not raw data:
  - Genotypes have been derived from raw data using particular software tools, one being more sensitive than the other ….
- For instance, SNP quality control involves assessing
  - missing data rates,
  - Hardy-Weinberg equilibrium (HWE),
  - allele frequencies,
  - Mendelian inconsistencies (using family-data)
  - sample heterozygosity, …

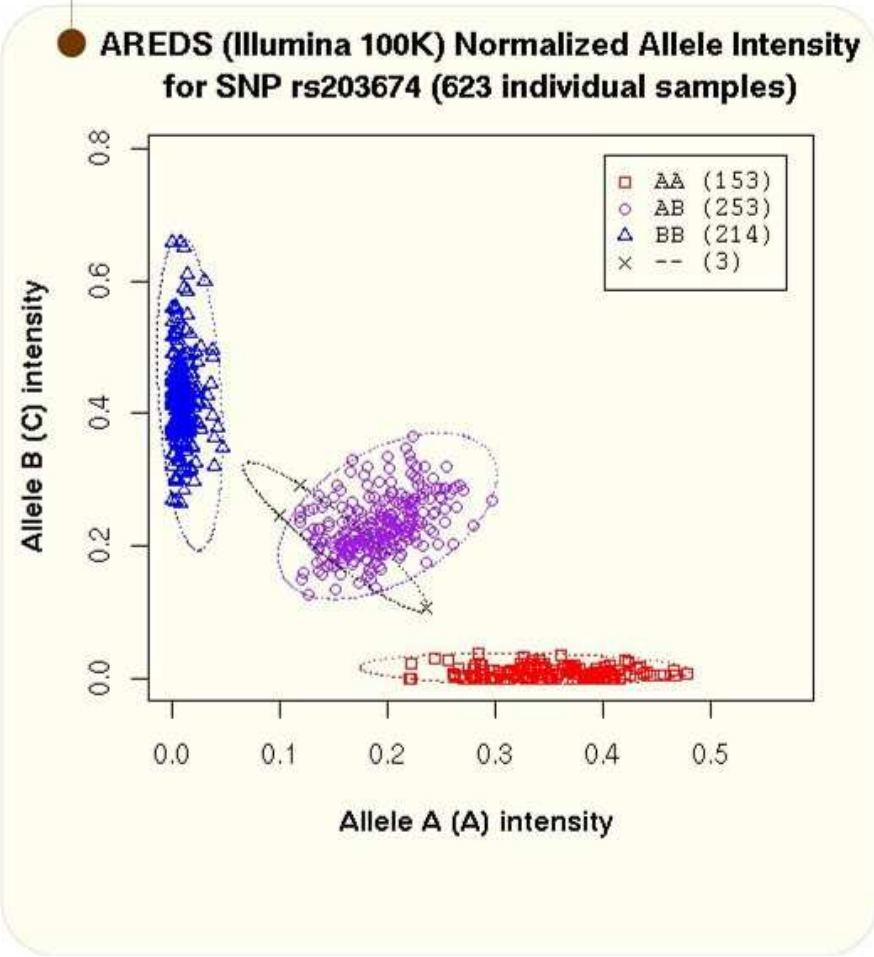
dbGaP study phs000001: rs203674 Genotype Summary

phv00000173	Genotype			Genotype Frequency
	AA	AC	CC	25% 50% 75%
Case	51	157	186	39.8% 47.2%
Control	82	88	23	42.5% 45.6%

phv00000173	Allele		Allele Frequency
	A	C	25% 50% 75%
Case	259	529	32.9% 67.1%
Control	252	134	65.3% 34.7%

phv00000173	Number of Samples		Success Rate
	Genotyped	Total	25% 50% 75%
Case	394	395	99.7%
Control	193	198	97.5%

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



(using dbGaP association browser tools)

2.b Hardy-Weinberg equilibrium

- Deviations from HWE can be due to inbreeding, population stratification or selection.
- Researchers have tested for HWE primarily as a data quality check and have discarded loci that, for example, deviate from HWE among controls at significance level $\alpha = 10^{-3}$ or 10^{-4}.
- Deviations from HWE can also be a symptom of disease association.
- So the possibility that a deviation from HWE is due to a deletion polymorphism or a segmental duplication that could be important in disease causation should certainly be considered before simply discarding loci…

Hardy-Weinberg equilibrium testing

- Testing for deviations from HWE can be carried out using a Pearson goodness-of-fit test, often known simply as 'the χ2 test' because the test statistic has approximately a χ2 null distribution.
- There are many different χ2 tests. The Pearson test is easy to compute, but the χ2 approximation can be poor when there are low genotype counts, in which case it is better to use a Fisher exact test.
- Fisher exact test does not rely on the χ2 approximation.
- The open-source data-analysis software R has an *R genetics package* that implements both Pearson and Fisher tests of HWE

Hardy-Weinberg equilibrium interpretation of test results

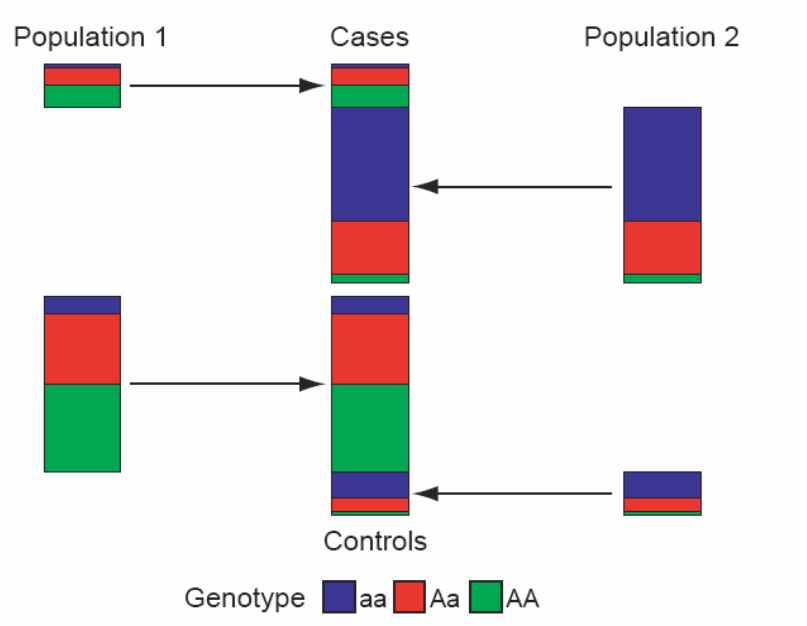
- A useful tool for interpreting the results of HWE and other tests on many SNPs is the log quantile–quantile (QQ) *p*-value plot:
  - the negative logarithm of the *i*-th smallest *p*-value is plotted against $-\log(i / (L + 1))$, where L is the number of SNPs.
- By a quantile, we mean the fraction (or percent) of points below the given value. That is, 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. 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.

Hardy-Weinberg equilibrium interpretation of test results

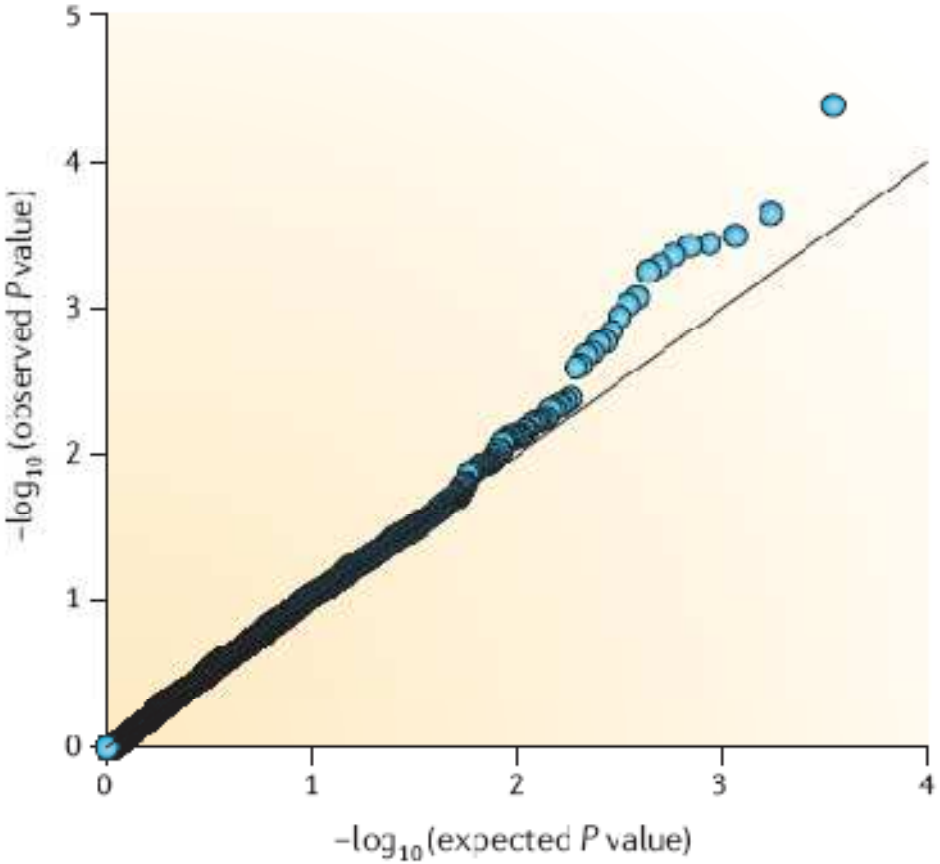
- Advantages of QQ plots include:
  - The sample sizes do not need to be equal.
  - Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
- Applied to genetic association studies and genetic association testing
  - Deviations from the $y = x$ line correspond to loci that deviate from the null hypothesis.
  - The close adherence of p-values to the black line over most of the range is encouraging as it implies that there are few systematic sources of spurious association.

Hardy-Weinberg equilibrium interpretation of test results

- In fact, spurious association is caused by two factors in population stratification (see also later).
  - A difference in proportion of individual from two (or more) subpopulation in case and controls
  - Subpopulations have different allele frequencies at the locus.



Hardy-Weinberg equilibrium interpretation of test results



(Balding 2006)

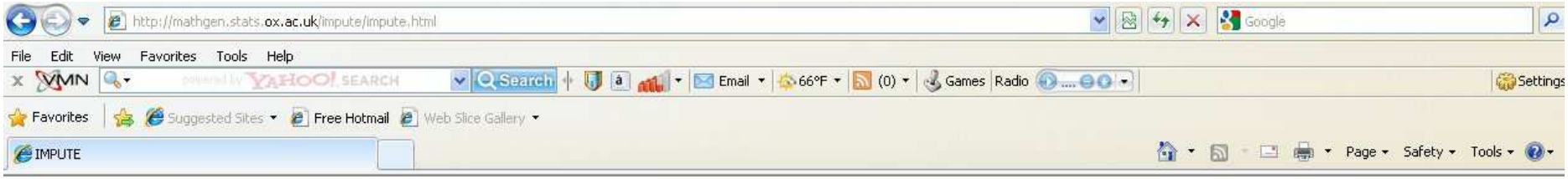
2.c Missing genotype data

Introduction

- For single-SNP analyses, if a few genotypes are missing there is not much problem.
- For multipoint SNP analyses, missing data can be more problematic because many individuals might have one or more missing genotypes.
- One convenient solution is data imputation
  - Data imputation involves replacing missing genotypes with predicted values that are based on the observed genotypes at neighbouring SNPs.
- For tightly linked markers data imputation can be reliable, can simplify analyses and allows better use of the observed data.
- For not tightly linked markers?

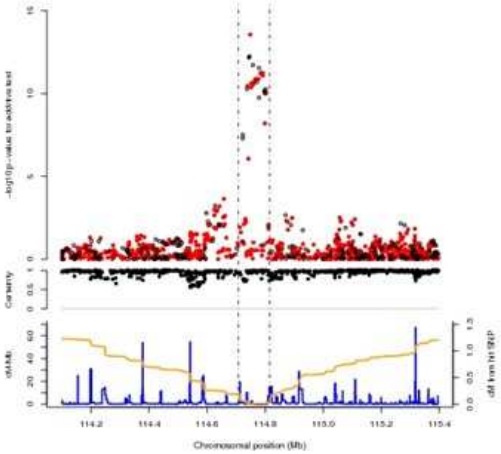
Introduction

- Imputation methods either seek a best prediction of a missing genotype, such as a
  - maximum-likelihood estimate (single imputation), or
  - randomly select it from a probability distribution (multiple imputations).
- The advantage of the latter approach is that repetitions of the random selection can allow averaging of results or investigation of the effects of the imputation on resulting analyses.
- Beware of settings in which cases are collected differently from controls. These can lead to differential rates of missingness even if genotyping is carried out blind to case-control status.
  - One way to check differential missingness rates is to code all observed genotypes as 1 and unobserved genotypes as 0 and to test for association of this variable with case-control status …



IMPUTE

IMPUTE is a program for estimating ("imputing") unobserved genotypes in SNP association studies. The program is designed to work seamlessly with the output of the genotype calling program CHIAMO and the population genetic simulator HAPGEN, and it produces output that can be analyzed using the program SNPTEST. There are currently three different versions of the **IMPUTE** software available for download: version 0.5 implements the methodology described in Marchini et al. (2007); version 1 is essentially the same as version 0.5, with a couple of added features; and version 2 implements a major extension that was introduced in Howie et al. (2009). The situations in which each version of the program can be applied are discussed below.

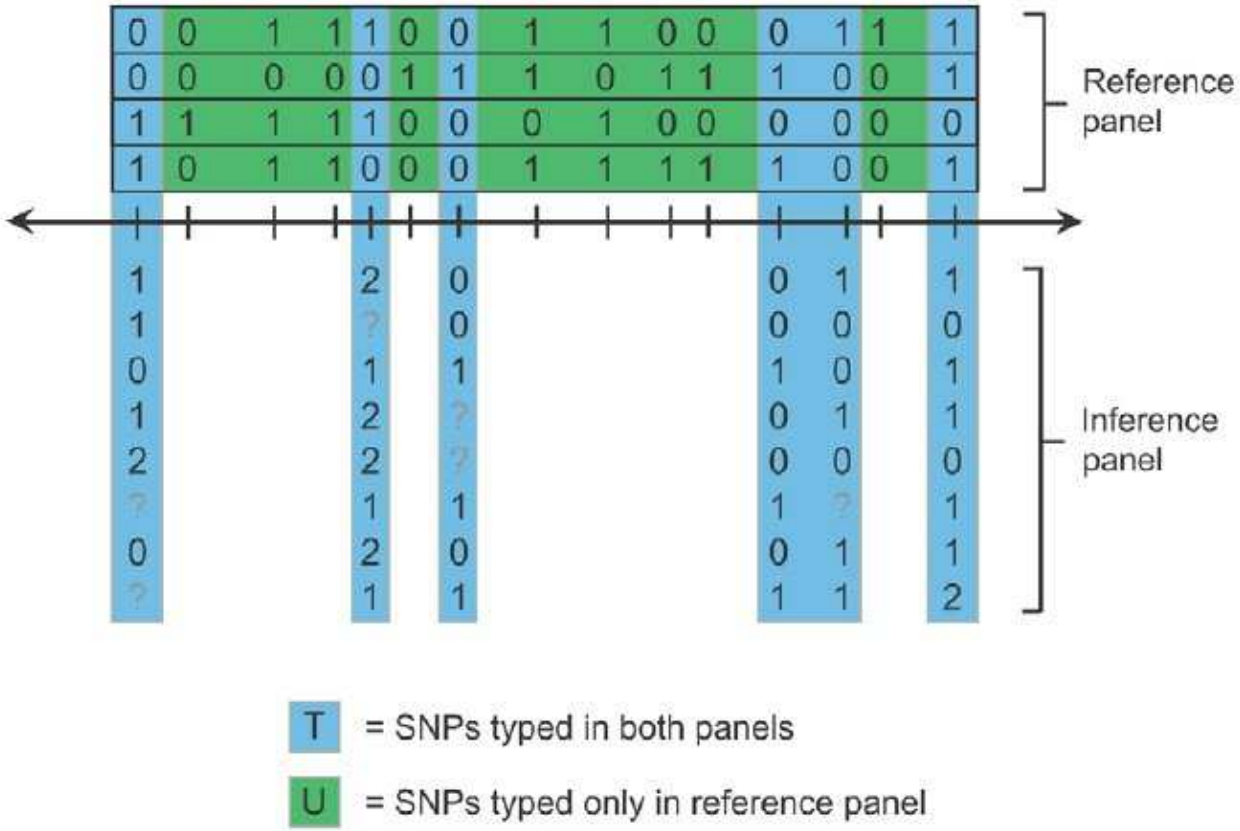


Version 0.5 (top)

IMPUTE v0.5 has now been superseded by IMPUTE v1.0, although we are keeping the website and software available for posterity. The description of **IMPUTE v1** below is equally applicable to **IMPUTE v0.5**.

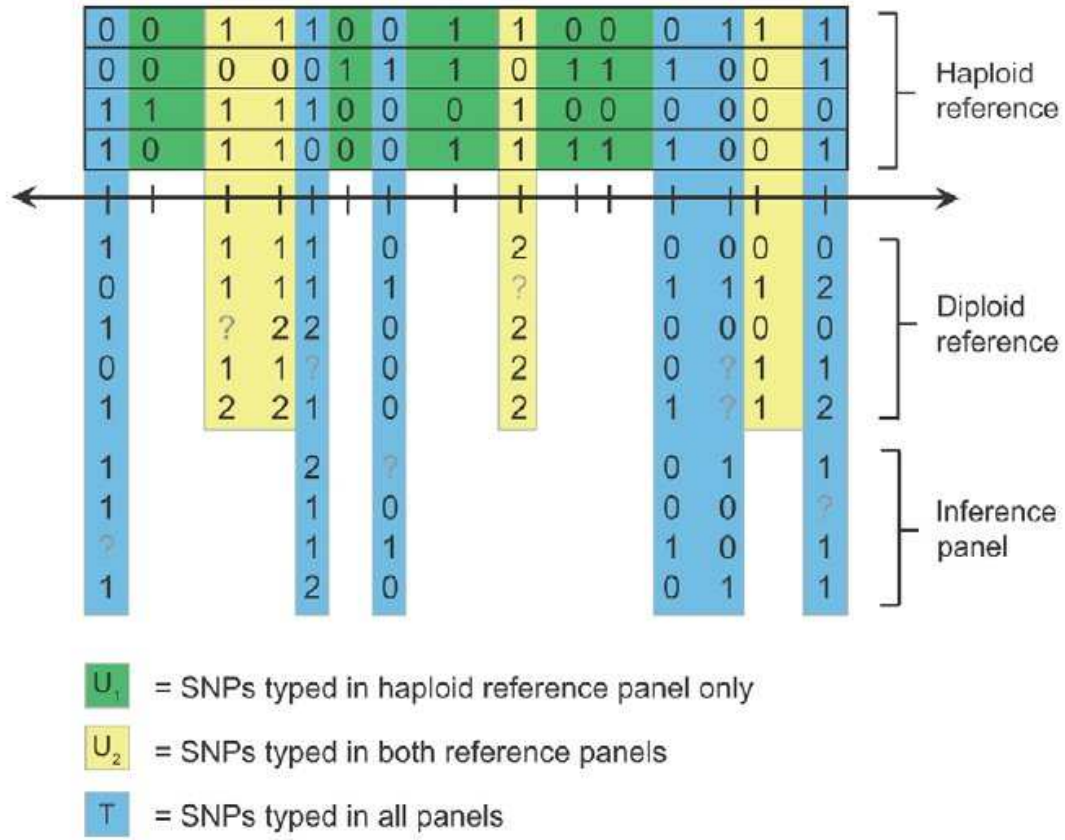
Read more about **IMPUTE v0.5**

Download **IMPUTE v0.5**



Schematic drawing of imputation Scenario A. In this drawing, haplotypes are represented as horizontal boxes containing 0's and 1's (for alternate SNP alleles), and unphased genotypes are represented as rows of 0's, 1's, 2's, and ?'s (where '1' is the heterozygous state and '?' denotes a missing genotype). The SNPs (columns) in the dataset can be partitioned into two disjoint sets: a set *T* (blue) that is genotyped in all individuals and a set *U* (green) that is genotyped only in the haploid reference panel. The goal of imputation in this scenario is to estimate the genotypes of SNPs in set *U* in the study sample.
doi:10.1371/journal.pgen.1000529.g001

(IMPUTE_v2: Howie et al 2009)



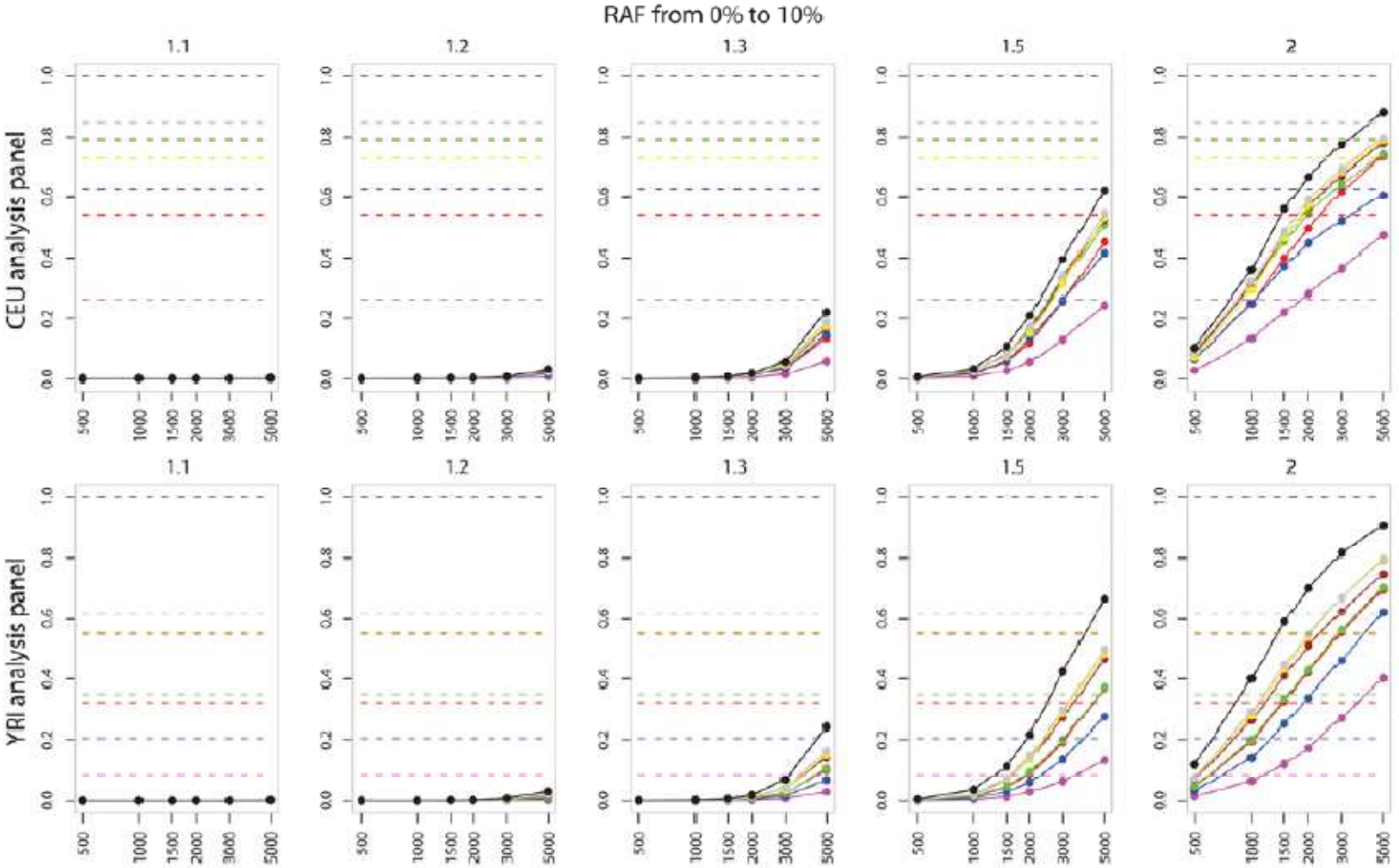
Schematic drawing of imputation Scenario B. In this drawing, haplotypes are represented as horizontal boxes containing 0's and 1's (for alternate SNP alleles), and unphased genotypes are represented as rows of 0's, 1's, 2's, and ?'s (where '1' is the heterozygous state and '?' denotes a missing genotype). The SNPs (columns) in the dataset can be partitioned into three disjoint sets: a set T (blue) that is genotyped in all individuals, a set U_2 (yellow) that is genotyped in both the haploid and diploid reference panels but not the study sample, and a set U_1 (green) that is genotyped only in the haploid reference panel. The goal of imputation in this scenario is to estimate the genotypes of SNPs in set U_2 in the study sample and SNPs in the set U_1 in both the study sample and, if desired, the diploid reference panel.
doi:10.1371/journal.pgen.1000529.g002

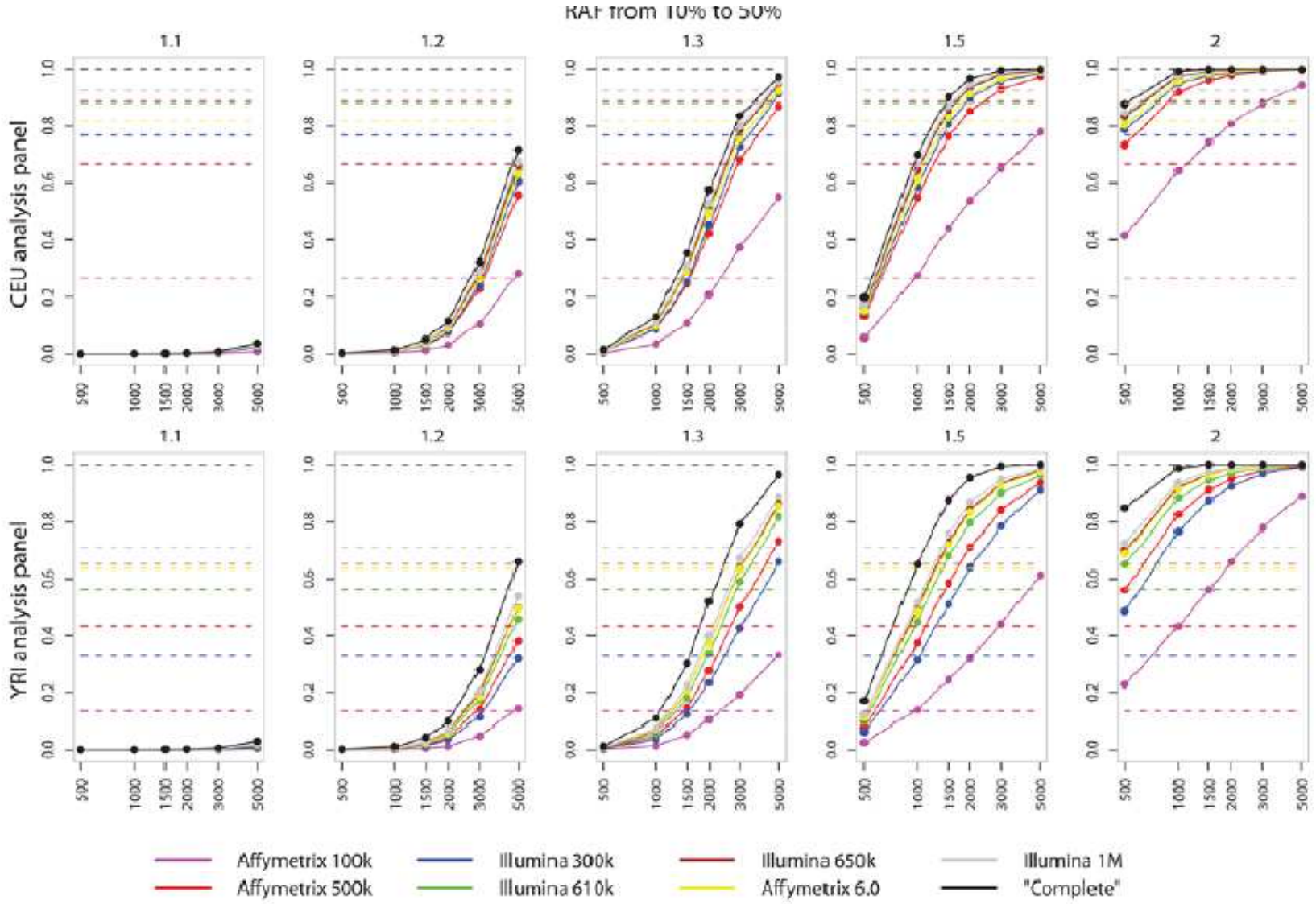
(IMPUTE_v2: Howie et al 2009)

The power of imputation

- Power for Common versus Rare alleles: Plots of power (solid lines) and coverage (dotted line) for increasing sample sizes of cases and controls (x-axis).
  - From left to right plots are given for increasing effect sizes (relative risk per allele). Both power and coverage range from 0 to 1 and are given on the y-axis. Results are for single-marker test of association.
  - The first plot show the power for rare risk alleles (RAF,0.1) and the second plot show the power for common risk alleles (RAF.0.1). doi:10.1371/journal.pgen.1000477.g003 – see next 2 slides
- The power of imputing potential benefits of increasing SNP density on the chips or from using imputation are greatest for low frequency SNPs.

(Spencer et al 2009)

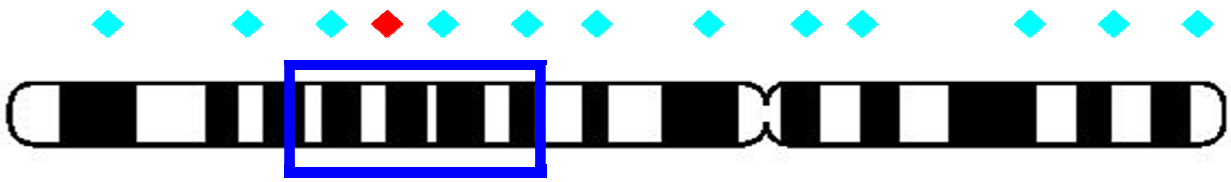

RAF from 0% to 10%
CEU analysis panel
YRI analysis panel
1.1
1.2
1.3
1.5
2
0.0
0.2
0.4
0.6
0.8
1.0
500
1000
1500
2000
3000
5000


RAF from 10% to 50%
1.1
1.2
1.3
1.5
2
CEU analysis panel
YRI analysis panel
Affymetrix 100k
Affymetrix 500k
Illumina 300k
Illumina 610k
Illumina 650k
Affymetrix 6.0
Illumina 1M
"Complete"

2.d Haplotype and genotype data

Introduction

- If we don't observe that causative locus directly, we need to combine the information of several markers in linkage disequilibrium (LD).
- Two approaches
  - Haplotype based method
  - tagSNPs based method



(Jung 2007)

Introduction

- Underlying an individual's genotypes at multiple tightly linked SNPs are the two haplotypes, each containing alleles from one parent.
- 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 DSL located at a marker, haplotype testing can be *less powerful*

Inferring haplotypes

- Direct, laboratory-based haplotyping or typing further family members to infer the unknown phase are expensive ways to obtain haplotypes. Fortunately, there are statistical methods for inferring haplotypes and population haplotype frequencies from the genotypes of unrelated individuals.
- These methods, and the software that implements them, rely on the fact that in regions of low recombination relatively few of the possible haplotypes will actually be observed in any population.
- These programs generally perform well, given high SNP density and not too much missing data.

Inferring haplotypes

- Software:
  - **SNPHAP** is simple and fast, whereas **PHASE** tends to be more accurate but comes at greater computational cost.
  - **FASTPHASE** is nearly as accurate as PHASE but much faster.
- Whatever software is used, remember that true haplotypes are more informative than genotypes.
- Inferred haplotypes are typically less informative because of uncertain phasing.
  - The information loss that arises from phasing is small when linkage disequilibrium (LD) is strong.

Measures of LD

- LD will remain crucial to the design of association studies until whole-genome resequencing becomes routinely available. Currently, few of the more than 10 million common human polymorphisms are typed in any given study.
- If a causal polymorphism is not genotyped, we can still hope to detect its effects through LD with polymorphisms that are typed (key principle behind doing genetic association analysis …).
- LD is a non-quantitative phenomenon: there is no natural scale for measuring it.
- 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).

Measures of LD

- 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’ 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

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

Properties for D'

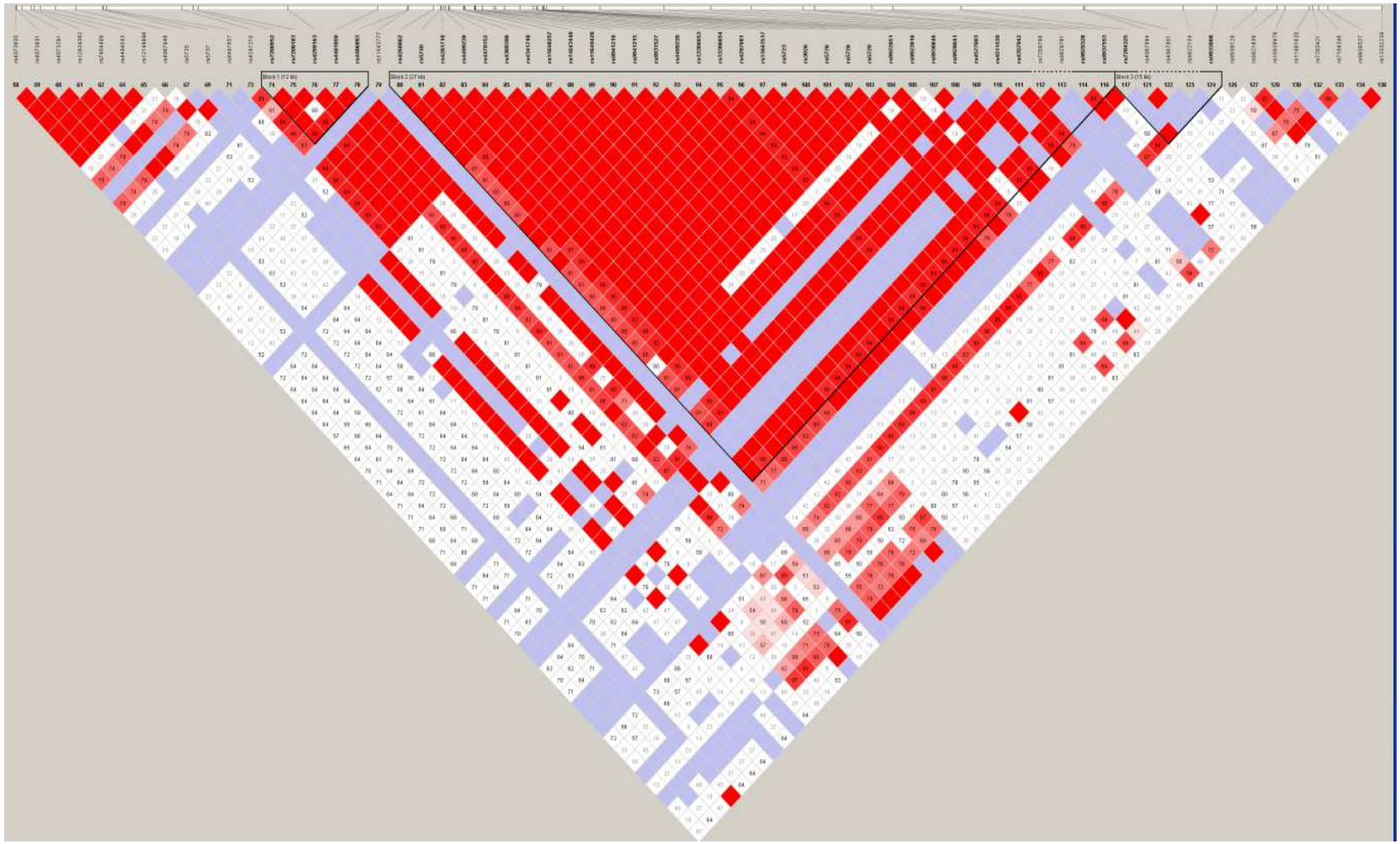
- *D*' is sensitive to even a few recombinations between the loci
- A disadvantage of *D*' is that it can be large (indicating high LD) even when one allele is very rare, which is usually of little practical interest.
- *D*' is inflated in small samples; the degree of bias will be greater for SNPs with rare alleles.
- So, the interpretation of values of D' < 1 is problematic, and values are difficult to compare between different samples because of the dependence on sample size.

Properties for r^2

- In contrast to D', r^2 is highly dependent upon allele frequency, and can be difficult to interpret when loci differ in their allele frequencies
- However, r^2 has desirable sampling properties, is directly related to the amount of information provided by one locus about the other, and is particularly useful in evolutionary and population genetics applications.
- Specifically, 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)

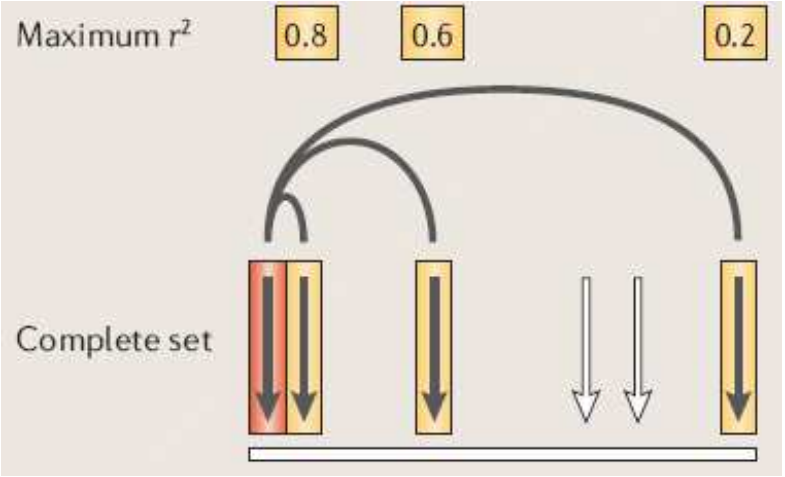
Haploview



1.e SNP tagging

Introduction

- Tagging refers to methods to select a minimal number of SNPs that retain as much as possible of the genetic variation of the full SNP set.
- Simple pairwise methods discard one (preferably that with most missing values) of every pair of SNPs with, say, $r^2 > 0.9$.
- More sophisticated methods can be more efficient, but the most efficient tagging strategy will depend on the statistical analysis to be used afterwards.
- In practice, tagging is only effective in capturing common variants.

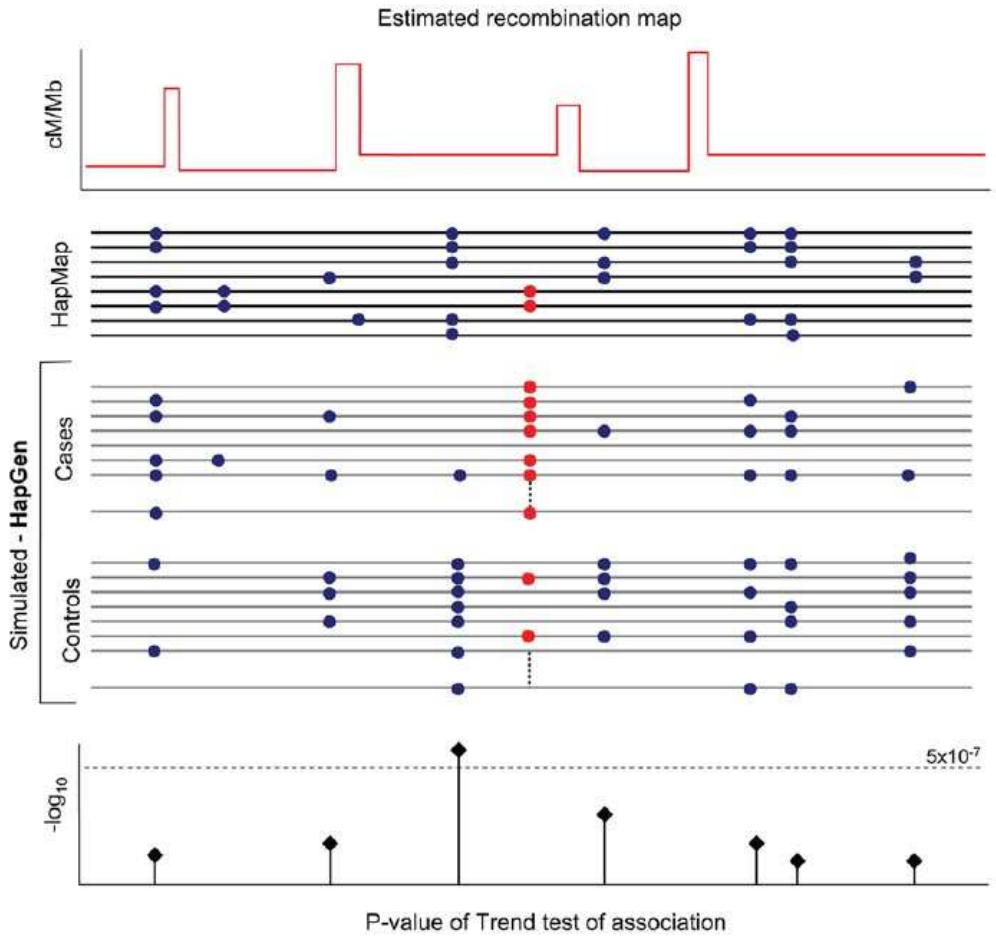


Two good reasons for tagging

- The first principal use for tagging is to select a 'good' subset of SNPs to be typed in all the study individuals from an extensive SNP set that has been typed in just a few individuals.
  - Until recently, this was frequently a laborious step in study design, but the International HapMap Project and related projects now allow selection of tag SNPs on the basis of publicly available data.
  - However, the population that underlies a particular study will typically differ from the populations for which public data are available, and a set of tag SNPs that have been selected in one population might perform poorly in another.
  - Nevertheless, recent studies indicate that tag SNPs often transfer well across populations

Two good reasons for tagging

- The second use for tagging is to select for analysis a subset of SNPs that have already been typed in all the study individuals.
- Although it is undesirable to discard available information, the amount of information lost might be small (at least, that is what is aimed for when applying SNP tagging algorithms).
- Reducing the SNP set can simplify analyses and lead to more statistical power by reducing the degrees of freedom (df) of a test.



Schematic of how power is estimated. At the top of the figure is the recombination map and haplotypes estimated from the HapMap project [1]. Using this population genetic information we simulate a case-control sample (grey lines) where the red dots indicate the disease locus and blue dots indicate linked genetic variation. By performing a test of association at each SNP on the genotyping chip we can estimate power by counting the number of simulation for which a test statistic exceed a significance threshold (dotted line). We compare genotyping chips by changing the set of SNP at which we carry out a test. See Methods.
doi:10.1371/journal.pgen.1000477.g001

(Spencer et al 2009)

3 Tests of association: single SNP

Introduction

- Population association studies compare unrelated individuals, but ‘unrelated’ actually means that relationships are unknown and presumed to be distant.
- Therefore, we cannot trace transmissions of phenotype over generations and must rely on correlations of current phenotype with current marker alleles.
- Such a correlation might be generated (but is not necessarily generated) by one or more groups of cases that share a relatively recent common ancestor at a causal locus.

A toy example

	AA	*AB*	*BB*	total
(A) Genotype counts				
Case	a = 10	b = 190	c = 800	a+b+c = 1000
Control	d = 3	e = 100	f = 900	d+e+f = 1003

	A	*B*	total
(B) Allele counts			
Case	x_{11} = 2a+b = 210	x_{12} = b+2c = 1790	2(a+b+c) = 2000
Control	x_{21} = 2d+e = 106	x_{22} = d+2f = 1900	2(d+e+f) = 2006

	AA+*AB*	*BB*	*AA*	*AB*+*BB*	total
(C) Two ways of grouping heterozygotes with homozygotes					
Case	a+b = 200	c = 800	a = 10	b+c = 990	a+b+c = 1000
Control	d+e = 103	f = 900	c = 3	d+e = 1000	d+e+f = 1003

There are 1000 case samples and 1003 control samples, whose genotype distribution is shown in the table (A); the number of A and B allele counts is in (B). The genotype counts in (C) are converted from (A) by combining AB with either AA or BB. Note that the total counts in (B) doubles the counts in (A), and the two tables in (C) correspond to the dominant and recessive models if allele *A* is considered as the risk allele.

(Li 2007)

A toy example

- A Pearson's test is a summary of discrepancy between the observed (O) and expected (E) genotype/allele count:

$$\chi^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i}$$

- For any χ^2 distributed test statistic with df degrees of freedom, one can decompose it to two χ^2 distributed test statistics with df 1 and df 2 degrees of freedom and their sum df 1 þ df 2 is equal to df.
- For example, the test statistic in the genotype based test (GBT) can be decomposed to two χ^2distributed values each with one degree of freedom.
- One of them is the test statistic in a commonly used test called Cochran–Armitage test (CAT).

A toy example

- CAT tests whether log(r), where r is the (number of cases)/(number of cases + number of controls) ratio, changes linearly with the AA, AB, BB genotype with a non-zero slope.
- Note that since AB is positioned between AA and BB genotype, the genotype is not just a categorical variable, but an ordered categorical variable.
- Also note that although CAT is genotype based, its value is closer to the allele-based ABT test statistic.

A toy example: testing

```
gc <- c(10, 190, 800, 3, 100, 900)
ac <- c(2*gc[1]+gc[2], gc[2]+2*gc[3], 2*gc[4]+gc[5], gc[5]+2*gc[6])
gc1 <- c(gc[1]+gc[2], gc[3], gc[4]+gc[5], gc[6])
gc2 <- c(gc[1],gc[2]+gc[3], gc[4], gc[5]+gc[6])
pvg <- chisq.test( matrix(gc, ncol=3, byrow=T), corr=FALSE)$p.value
pva <- chisq.test( matrix(ac, ncol=2, byrow=T), corr=FALSE)$p.value
pvg1 <-  chisq.test( matrix(gc1, ncol=2, byrow=T), corr=FALSE)$p.value
pvg2 <-  chisq.test( matrix(gc2, ncol=2, byrow=T), corr=FALSE)$p.value
pvb <- min(pvg1, pvg2)

print( c(pvg, pva, pvb))  # 6.918239e-09 9.150309e-10 1.224003e-09
pvg.f <- fisher.test( matrix(gc, ncol=3, byrow=T))$p.value
pva.f <- fisher.test( matrix(ac, ncol=2, byrow=T))$p.value
pvg1.f <-  fisher.test( matrix(gc1, ncol=2, byrow=T))$p.value
pvg2.f <-  fisher.test( matrix(gc2, ncol=2, byrow=T))$p.value
pvb.f <- min(pvg1.f, pvg2.f)
print( c(pvg.f, pva.f, pvb.f))  # 2.412721e-09 8.047005e-10 1.132535e-09

pvcat <- prop.trend.test(gc[1:3], gc[1:3]+gc[4:6], score=c(0, 0.5, 1))$p.value
print( c(pvcat) ) # 9.820062e-10

gc <- gc*2
...                             # repeat the tests
print( c(pvg, pva, pvb))        # 4.786203e-17 4.716312e-18 8.379499e-18
print( c(pvg.f, pva.f, pvb.f))  # 1.231881e-17 3.485271e-18 6.810263e-18
print( c(pvcat) ) # 5.422705e-18
```

A toy example: testing

- What is the effect of choosing a different genetic model?
- What is the effect of choosing a genotype test versus an allelic test?
- Are allelic tests always applicable?
- When do you expect the largest differences between Pearson's chi-square and Fisher's exact test?
- What is the effect of doubling the sample size on these tests?
- How can you protect yourself against uncertain disease models?

A toy example: estimation

```
ci.or <- function(counts, alpha){    # alpha=0.05 corresponds to 95%CI
 f <- qnorm(1- alpha/2)              # if alpha=0.05, f=1.96
 or <- counts[1]*counts[4]/(counts[2]*counts[3])
 sq <- sqrt(1/counts[1]+1/counts[2]+1/counts[3]+1/counts[4])
 upper <- exp( log(or) + f*sq)
 lower <- exp( log(or) - f*sq)
 res <- c(lower, or, upper)
 res
}

print( ci.or(ac, 0.05))             # 1.650411 2.102878 2.679390
print( ci.or(ac, 0.01))             # 1.529428 2.102878 2.891339

ac <- ac*2                          # double the sample size
print( ci.or(ac, 0.05))             # 1.771784 2.102878 2.495842
print( ci.or(ac, 0.01))             # 1.678927 2.102878 2.633882
```

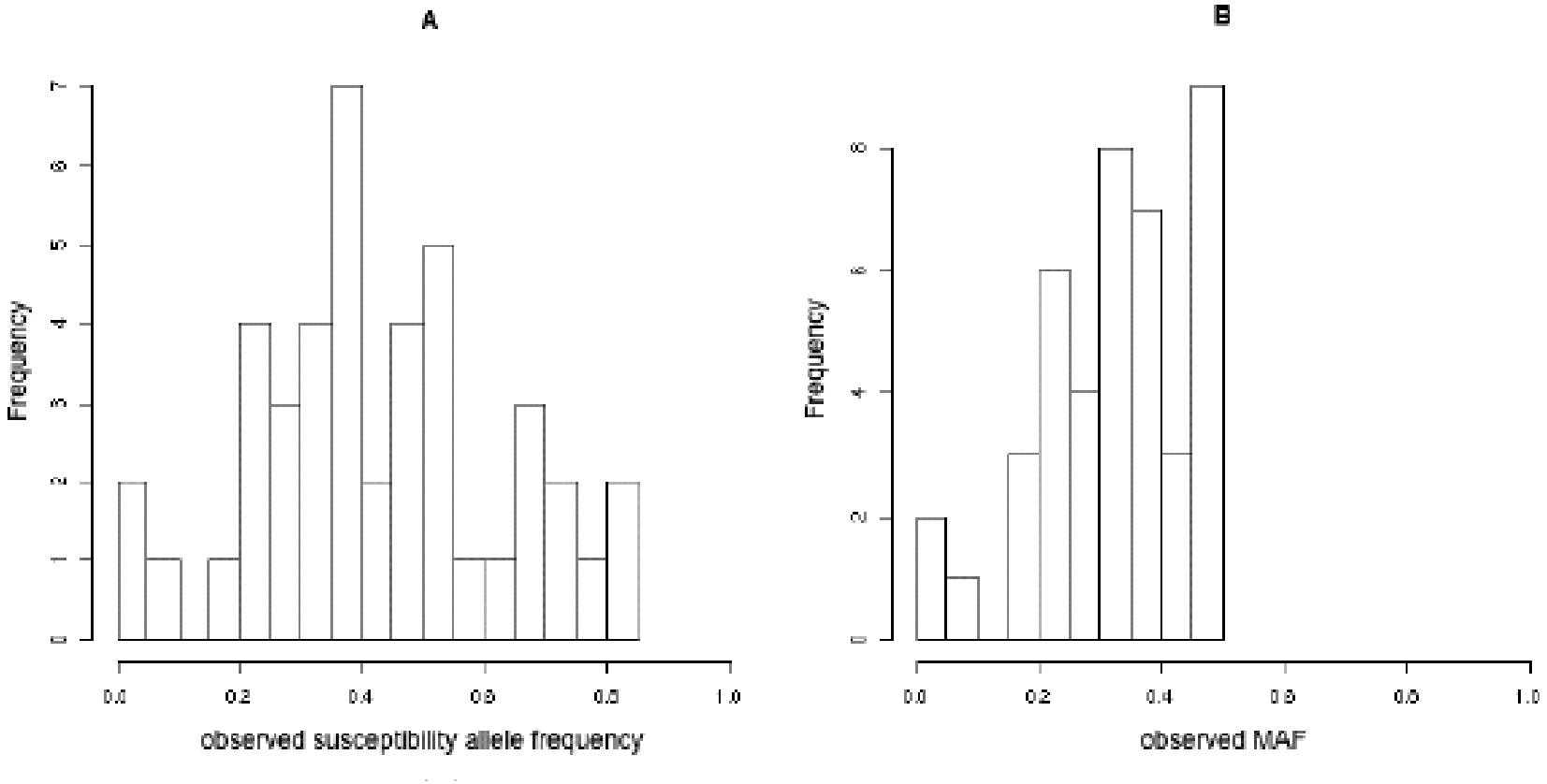
A toy example: estimation

- Will all packages give you the same output when estimating odds ratios with confidence intervals, assuming the data and the significance level are the same?
- What is the effect of decreasing the significance level?
- What is the effect of doubling the sample size?

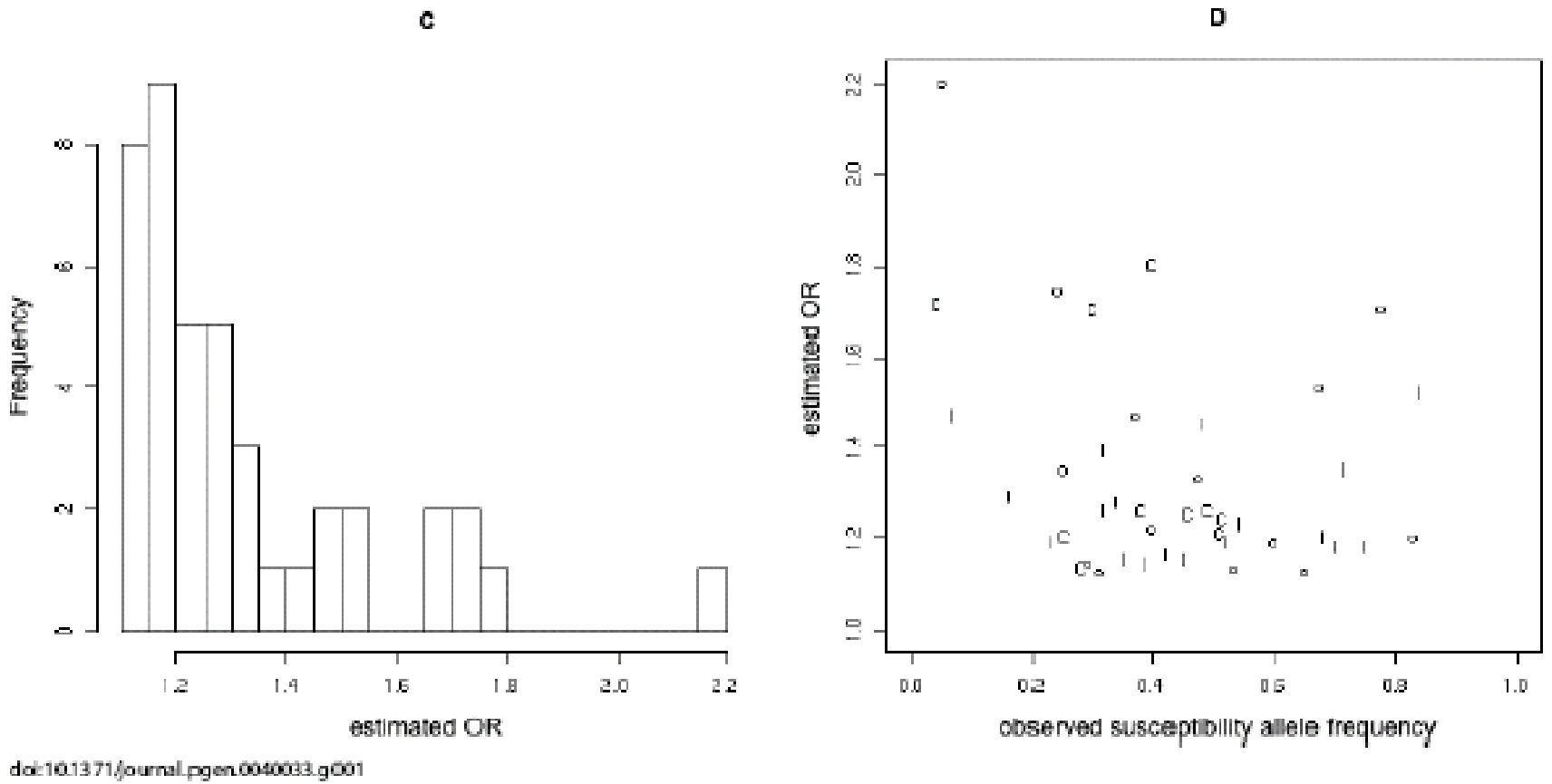
Which odds ratios can we expect?

- Many genome scientists are turning back to study rare disorders that are traceable to defects in single genes, and whose causes have remained a mystery.
- The change is partly a result of frustration with the disappointing results of genome-wide association studies (GWAS).
- Rather than sequencing whole genomes, GWAS studies examine a subset of DNA variants in thousands of unrelated people with common diseases. Now, however, sequencing costs are dropping, and whole genome sequences can quickly provide in-depth information about individuals, enabling scientists to locate genetic mutations that underlie rare diseases by sequencing a handful of people.

(Nature News: Published online 22 September 2009 | **461**, 459 (2009) | doi:10.1038/461458a)



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



doi:10.1371/journal.pgen.0040033.g001

(C) Histogram of estimated ORs (estimate of genetic effect size) at confirmed susceptibility loci. (D) Plot of estimated OR against susceptibility allele frequency at confirmed susceptibility loci. (Iles 2008)

The use of regression analysis

- Regression-type problems were first considered in the 18th century concerning navigation using astronomy.
- Legendre developed the method of least squares in 1805. Gauss claimed to have developed the method a few years earlier and showed that the least squares was the optimal solution when the errors are normally distributed in 1809.
- The methodology was used almost exclusively in the physical sciences until later in the 19th century. Francis Galton coined the term regression to mediocrity in 1875 in reference to the simple regression equation in the form

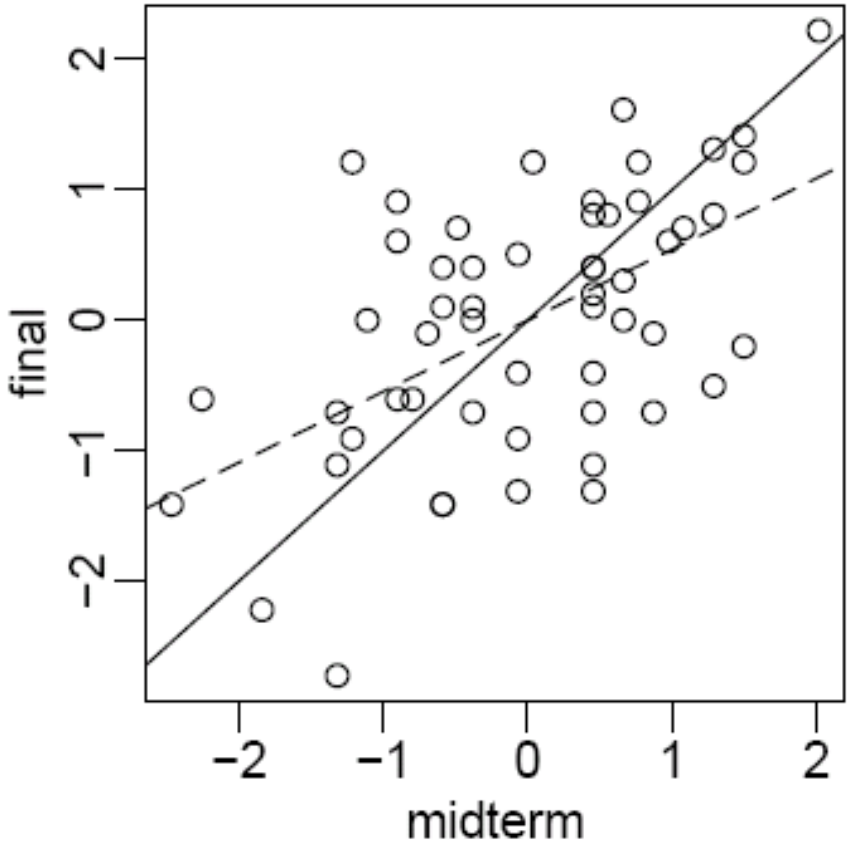
$$\frac{y-\bar{y}}{SD_y} = r\frac{(x-\bar{x})}{SD_x}.$$

The use of regression analysis

- Galton used this equation to explain the phenomenon that sons of tall fathers tend to be tall but not as tall as their fathers while sons of short fathers tend to be short but not as short as their fathers.
- This effect is called the regression effect.
- We can illustrate this effect with some data on scores from a course
  - When we scale each variable to have mean 0 and SD 1 so that we are not distracted by the relative difficulty of each exam and the total number of points possible.

How does this simplify the regression equation?

The use of regression analysis



(Faraway 2002)

The use of regression analysis

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

The use of regression analysis

- The basic syntax for doing regression in R is lm(Y~model) to fit linear models and glm() to fit generalized linear models.
- Linear regression and logistic regression are special type of models you can fit using lm() and glm() respectively.
- General syntax rules in R model fitting are given on the next slide.

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

The use of regression analysis

- Quantitative models always rest on assumptions about the way the world works, and regression models are no exception.
- There are four principal assumptions which justify the use of linear regression models for purposes of prediction:
  - linearity of the relationship between dependent and independent variables
  - independence of the errors (no serial correlation)
  - homoscedasticity (constant variance) of the errors
    - versus time
    - versus the predictions (or versus any independent variable)
  - normality of the error distribution.

(http://www.duke.edu/~rnau/testing.htm)

Linear regression analysis

- If any of these assumptions is violated (i.e., if there is nonlinearity, serial correlation, heteroscedasticity, and/or non-normality), then the forecasts, confidence intervals, and insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading.
- Violations of linearity are extremely serious--if you fit a linear model to data which are nonlinearly related, your predictions are likely to be seriously in error, especially when you extrapolate beyond the range of the sample data.
- How to detect:
  - nonlinearity is usually most evident in a plot of the observed versus predicted values or a plot of residuals versus predicted values, which are a part of standard regression output. The points should be symmetrically distributed around a diagonal line in the former plot or a

horizontal line in the latter plot. Look carefully for evidence of a "bowed" pattern, indicating that the model makes systematic errors whenever it is making unusually large or small predictions.

- How to fix: consider
  - applying a nonlinear transformation to the dependent and/or independent variables--if you can think of a transformation that seems appropriate. For example, if the data are strictly positive, a log transformation may be feasible. Another possibility to consider is adding another regressor which is a nonlinear function of one of the other variables. For example, if you have regressed Y on X, and the graph of residuals versus predicted suggests a parabolic curve, then it may make sense to regress Y on both X and X^2 (i.e., X-squared). The latter transformation is possible even when X and/or Y have negative values, whereas logging may not be.

Linear regression analysis

- Violations of independence are also very serious in time series regression models: serial correlation in the residuals means that there is room for improvement in the model, and extreme serial correlation is often a symptom of a badly mis-specified model, as we saw in the auto sales example. Serial correlation is also sometimes a byproduct of a violation of the linearity assumption--as in the case of a simple (i.e., straight) trend line fitted to data which are growing exponentially over time.
- How to detect:
  - The best test for residual autocorrelation is to look at an autocorrelation plot of the residuals. (If this is not part of the standard output for your regression procedure, you can save the RESIDUALS and use another procedure to plot the autocorrelations.)

- Ideally, most of the residual autocorrelations should fall within the 95% confidence bands around zero, which are located at roughly plus-or-minus 2-over-the-square-root-of-n, where n is the sample size.
- Thus, if the sample size is 50, the autocorrelations should be between +/- 0.3. If the sample size is 100, they should be between +/- 0.2. Pay especially close attention to significant correlations at the first couple of lags and in the vicinity of the seasonal period, because these are probably not due to mere chance and are also fixable.

- How to fix:
  - Minor cases of positive serial correlation (say, lag-1 residual autocorrelation in the range 0.2 to 0.4) indicate that there is some room for fine-tuning in the model. Consider adding lags of the dependent variable and/or lags of some of the independent variables.

- Major cases of serial correlation usually indicate a fundamental structural problem in the model. You may wish to reconsider the transformations (if any) that have been applied to the dependent and independent variables. It may help to stationarize all variables through appropriate combinations of differencing, logging, and/or deflating.

Linear regression analysis

- Violations of homoscedasticity make it difficult to gauge the true standard deviation of the forecast errors, usually resulting in confidence intervals that are too wide or too narrow. In particular, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. Heteroscedasticity may also have the effect of giving too much weight to small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.
- How to detect:
  - look at plots of residuals versus time and residuals versus predicted value, and be alert for evidence of residuals that are getting larger (i.e., more spread-out) either as a function of time or as a function of the predicted value. (To be really thorough, you might also want to plot residuals versus some of the independent variables.)

- How to fix:
  - In time series models, heteroscedasticity often arises due to the effects of inflation and/or real compound growth, perhaps magnified by a multiplicative seasonal pattern. Some combination of logging and/or deflating will often stabilize the variance in this case. Stock market data may show periods of increased or decreased volatility over time--this is normal and is often modeled with so-called ARCH (auto-regressive conditional heteroscedasticity) models in which the error variance is fitted by an autoregressive model. Such models are beyond the scope of this course--however, a simple fix would be to work with shorter intervals of data in which volatility is more nearly constant. Heteroscedasticity can also be a byproduct of a significant violation of the linearity and/or independence assumptions, in which case it may also be fixed as a byproduct of fixing those problems.

Linear regression analysis

- Violations of normality compromise the estimation of coefficients and the calculation of confidence intervals. Sometimes the error distribution is "skewed" by the presence of a few large outliers. Since parameter estimation is based on the minimization of squared error, a few extreme observations can exert a disproportionate influence on parameter estimates. Calculation of confidence intervals and various signficance tests for coefficients are all based on the assumptions of normally distributed errors. If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow.
- How to detect:
  - the best test for normally distributed errors is a normal probability plot of the residuals. This is a plot of the fractiles of error distribution versus the fractiles of a normal distribution having the same mean and

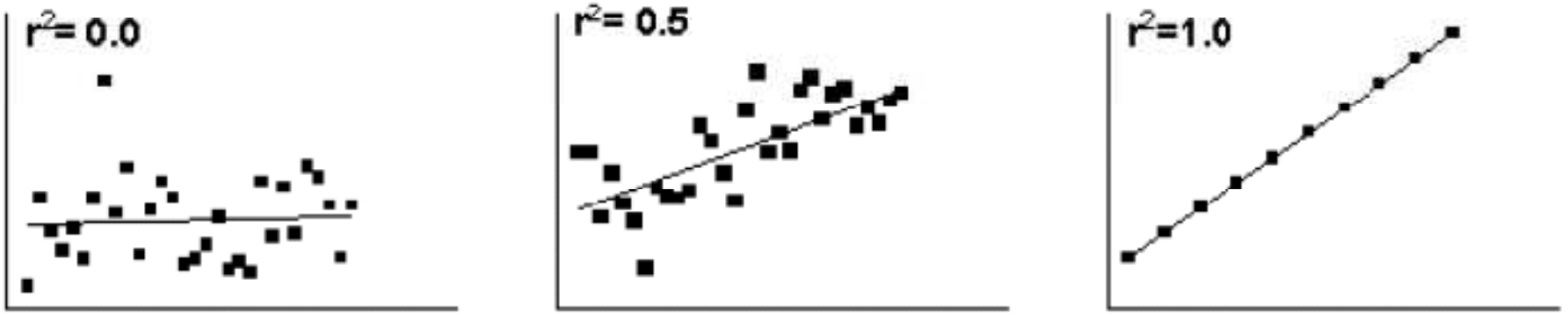
variance. If the distribution is normal, the points on this plot should fall close to the diagonal line. A bow-shaped pattern of deviations from the diagonal indicates that the residuals have excessive skewness (i.e., they are not symmetrically distributed, with too many large errors in the same direction). An S-shaped pattern of deviations indicates that the residuals have excessive kurtosis--i.e., there are either two many or two few large errors in both directions.

- How to fix:
  - violations of normality often arise either because (a) the distributions of the dependent and/or independent variables are themselves significantly non-normal, and/or (b) the linearity assumption is violated. In such cases, a nonlinear transformation of variables might cure both problems. In some cases, the problem with the residual distribution is mainly due to one or two very large errors. Such values should be

scrutinized closely: are they genuine (i.e., not the result of data entry errors), are they explainable, are similar events likely to occur again in the future, and how influential are they in your model-fitting results? (The "influence measures" report is a guide to the relative influence of extreme observations.) If they are merely errors or if they can be explained as unique events not likely to be repeated, then you may have cause to remove them. In some cases, however, it may be that the extreme values in the data provide the most useful information about values of some of the coefficients and/or provide the most realistic guide to the magnitudes of forecast errors.

Linear regression analysis

- The value r2 is a fraction between 0.0 and 1.0, and has no units. An r2 value of 0.0 means that knowing X does not help you predict Y.
- There is no linear relationship between X and Y, and the best-fit line is a horizontal line going through the mean of all Y values. When
- r2 equals 1.0, all points lie exactly on a straight line with no scatter. Knowing X lets you predict Y perfectly.



Is linear regression the correct type of analysis for you?

Question	Discussion
Can the relationship between X and Y be graphed as a straight line?	In many experiments the relationship between X and Y is curved, making linear regression inappropriate. Either transform the data, or use a program (such as GraphPad Prism) that can perform nonlinear curve fitting.
Is the scatter of data around the line Gaussian (at least approximately)?	Linear regression analysis assumes that the scatter is Gaussian.
Is the variability the same everywhere?	Linear regression assumes that scatter of points around the best-fit line has the same standard deviation all along the curve. The assumption is violated if the points with high or low X values tend to be further from the best-fit line. The assumption that the standard deviation is the same everywhere is termed *homoscedasticity*.
Do you know the X values precisely?	The linear regression model assumes that X values are exactly correct, and that experimental error or biological variability only affects the Y values. This is rarely the case, but it is sufficient to assume that any imprecision in measuring X is very small compared to the variability in Y.
Are the data points independent?	Whether one point is above or below the line is a matter of chance, and does not influence whether another point is above or below the line.
Are the X and Y values intertwined?	If the value of X is used to calculate Y (or the value of Y is used to calculate X) then linear regression calculations are invalid.

Use of `lm()` in genetics

For a continuous outcome,

```
lm(outcome ~ genetic.predictor, [...] )
```

estimates the association between `outcome` and `predictor`

The **optional** arguments [...] might be

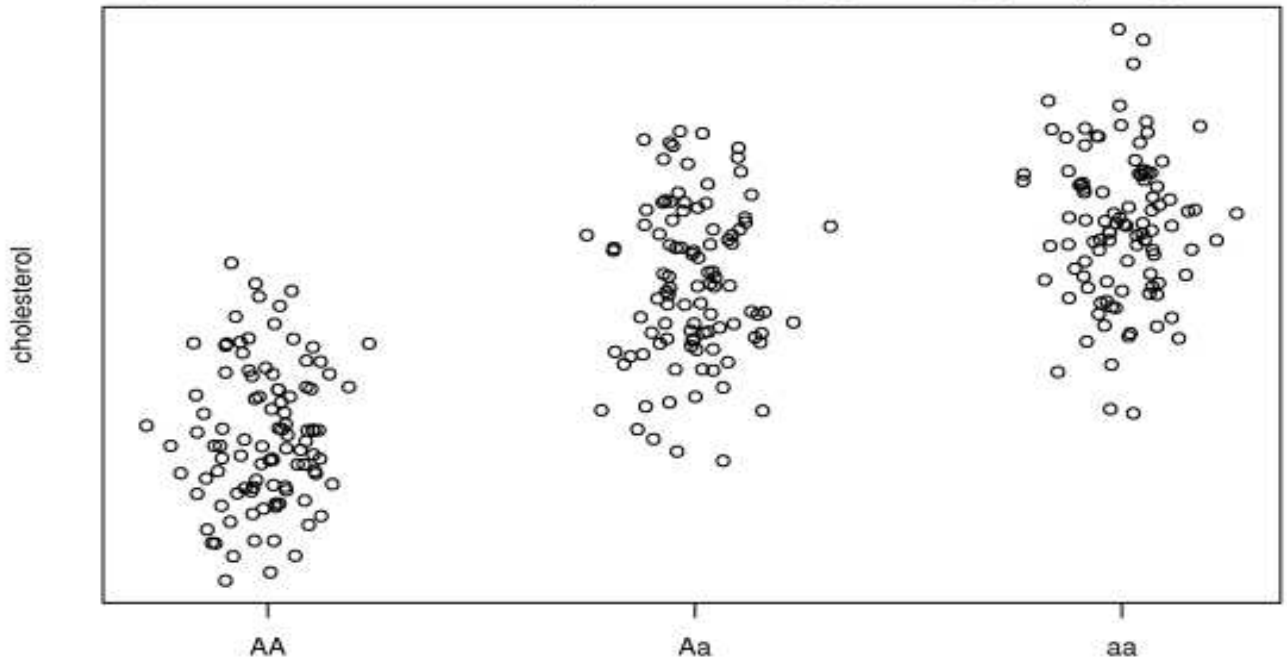
- `data=my.data` – your dataset
- `subset=race==CEPH` – use partial data
- `weights` – for advanced analyses

Model Description	`predictor`	Common name
Number of minor alleles	`(g=='Aa') + 2*(g=='aa')` or `as.numeric(g)`	Additive
Presence of minor allele	`(g=='Aa') | (g=='aa')`	Dominant
Homozygous for minor allele	`g=='aa'`	Recessive
Distinct effects for hetero/homozygous	`factor(g)`	2 parameter, or "2 df"

(Rice 2008)

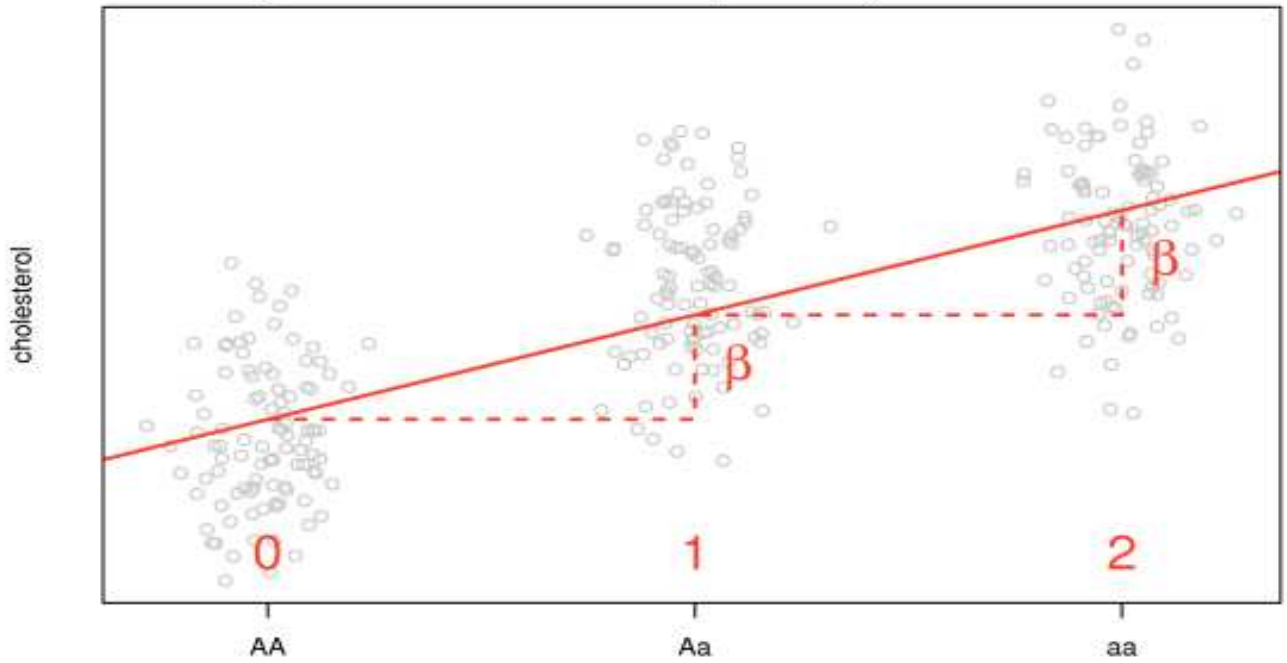
Use of `lm()` in genetics

Some data; cholesterol levels plotted by genotype (single SNP)



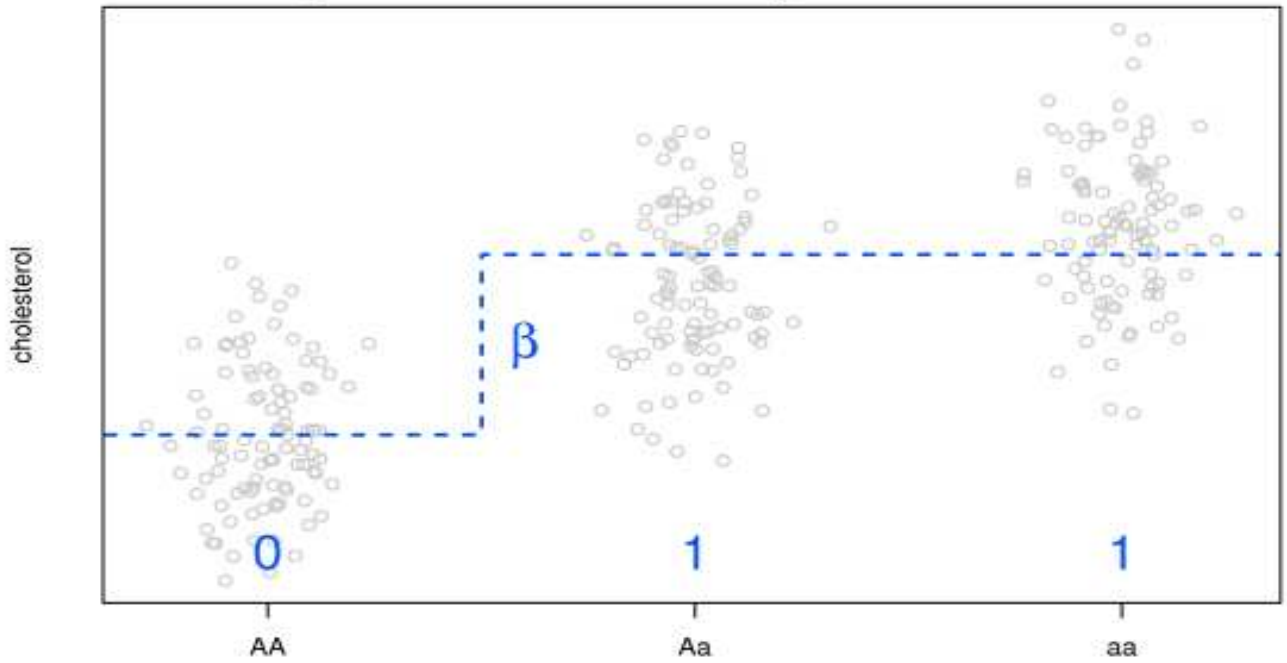
Use of `lm()` in genetics

Additive model (the most commonly used)



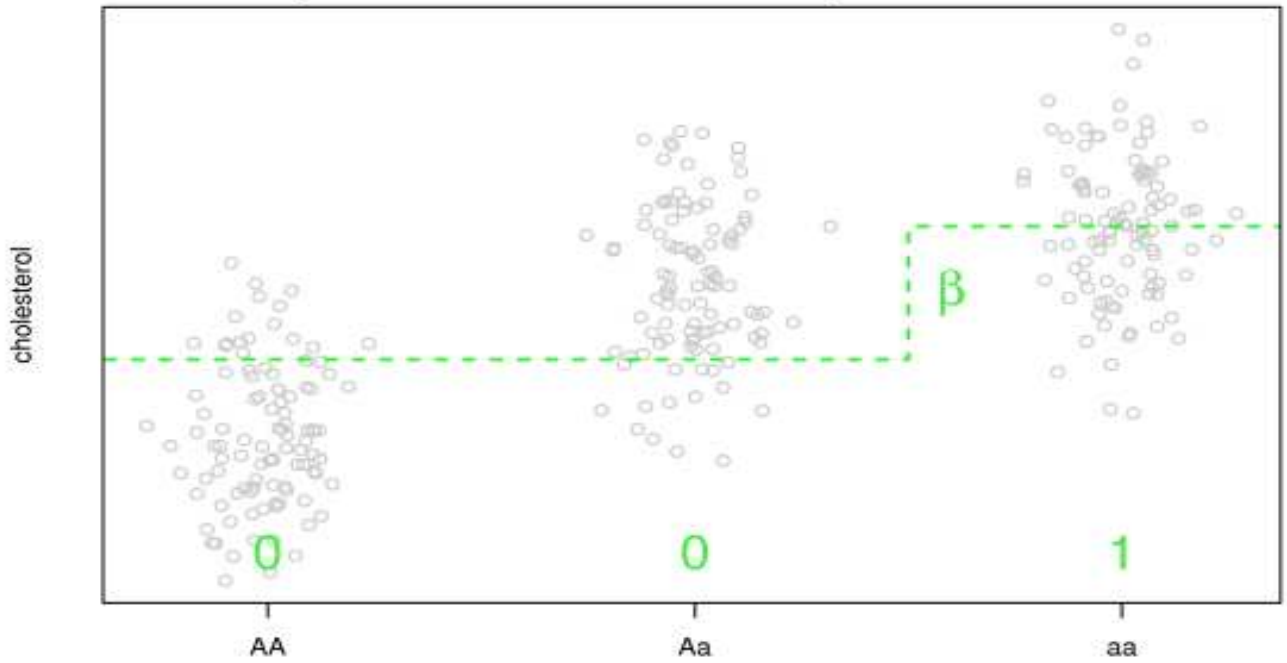
Use of `lm()` in genetics

Dominant model (best fit to this data)



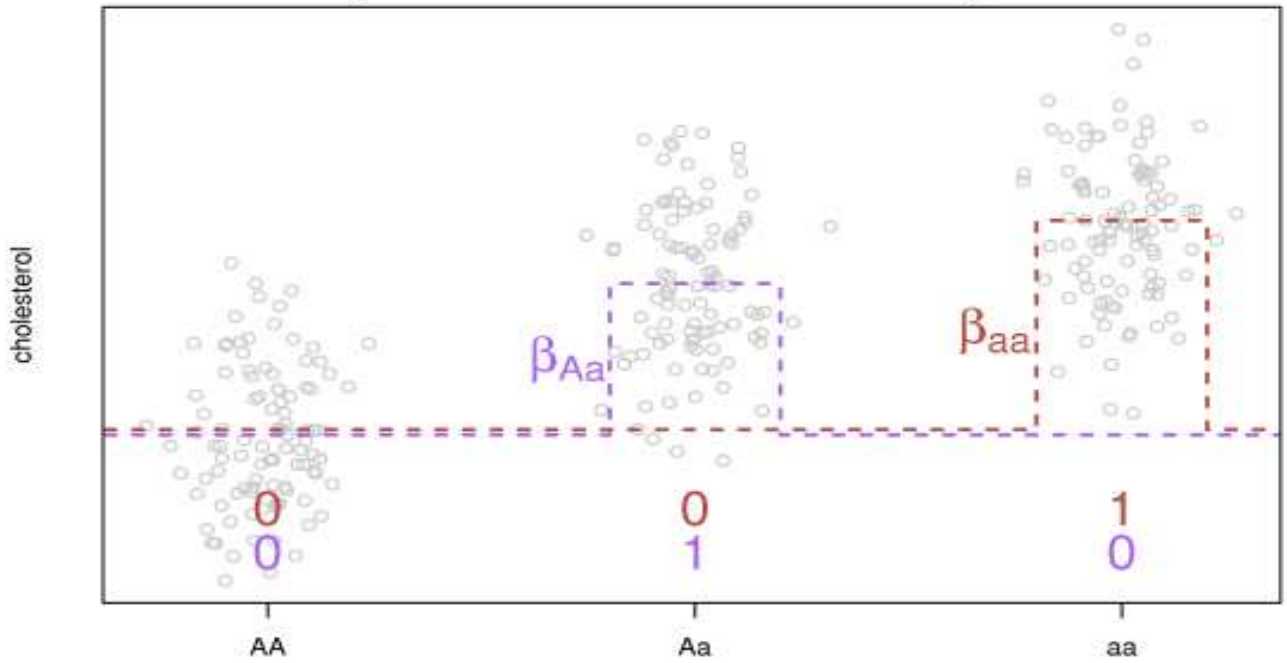
Use of `lm()` in genetics

Recessive model (least stable for rare aa)



Use of `lm()` in genetics

2 parameter model (robust but can be overkill)



lm(): Estimates, Intervals, p-values

lm() produces **point estimates** for your model;

```
> n.minor <- (g=="Aa") + 2*(g=="aa")
> my.lm <- lm( cholesterol ~ n.minor )
> my.lm
Call:
lm(formula = cholesterol ~ n.minor)
Coefficients:
(Intercept)      n.minor
     0.2104       0.9507
```

– also available via `my.lm$coefficients`.

The `coefficients` in the output tell you the **additive increase** in `outcome` associated with a **one-unit** difference in the `genetic predictor`.

The coefficient for `n.minor` is in units of `cholesterol`

lm(): Estimates, Intervals, p-values

You will also want **confidence intervals**;

```
> confint.default(my.lm)
                  2.5 %    97.5 %
(Intercept) 0.08391672 0.3368275
n.minor     0.85279147 1.0486953
```

Remember to **round these numbers** to an appropriate number of significant figures! (2 or 3 is usually enough)

We are **seldom** interested in the `Intercept`

lm(): Estimates, Intervals, p-values

Two-sided **p-values** are also available;

```
> summary(my.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.21037    0.06426   3.274  0.00119 **
n.minor      0.95074    0.04977  19.101  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this data, we have **strong evidence** of an **additive effect** of the minor allele on cholesterol

`summary(my.lm)` gives **many** other details – ignore for now

Confidence intervals are just `Estimate` $\pm$ 2$\times$`Std.Error`

Use of `glm()` in genetics

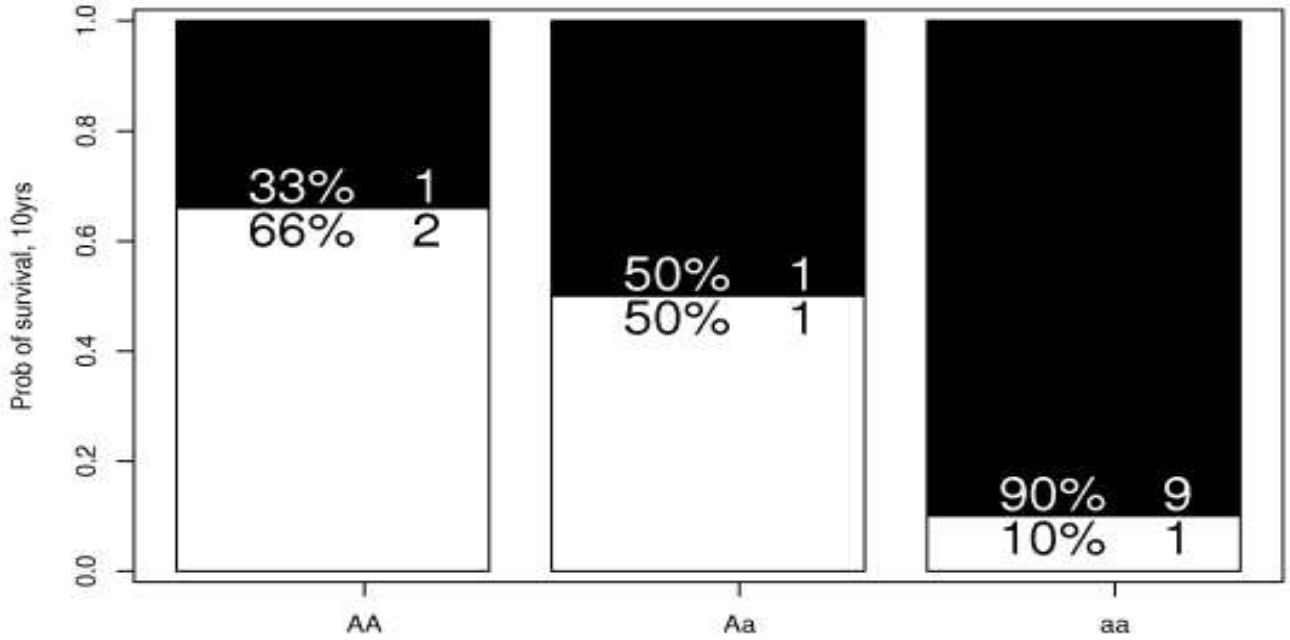
Logistic regression is the 'default' analysis for **binary outcomes**

Outcome	Type	Regression	Scale
Cholesterol Blood Pressure BMI	Continuous	Linear	Difference in Outcome
Death Stroke BMI>30	Binary	Logistic	Ratio of odds

What are **odds**? Really just **probability**...

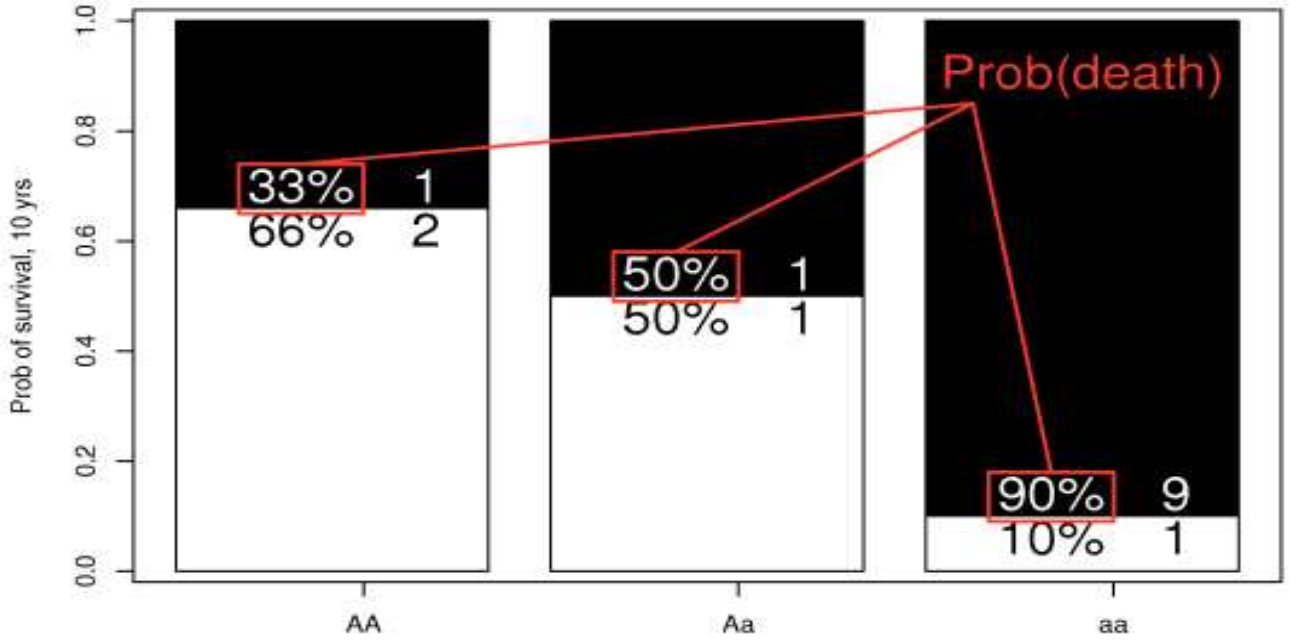
Use of `glm()` in genetics

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



Use of `glm()` in genetics

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



4 Tests of association: multiple SNPs

Introduction

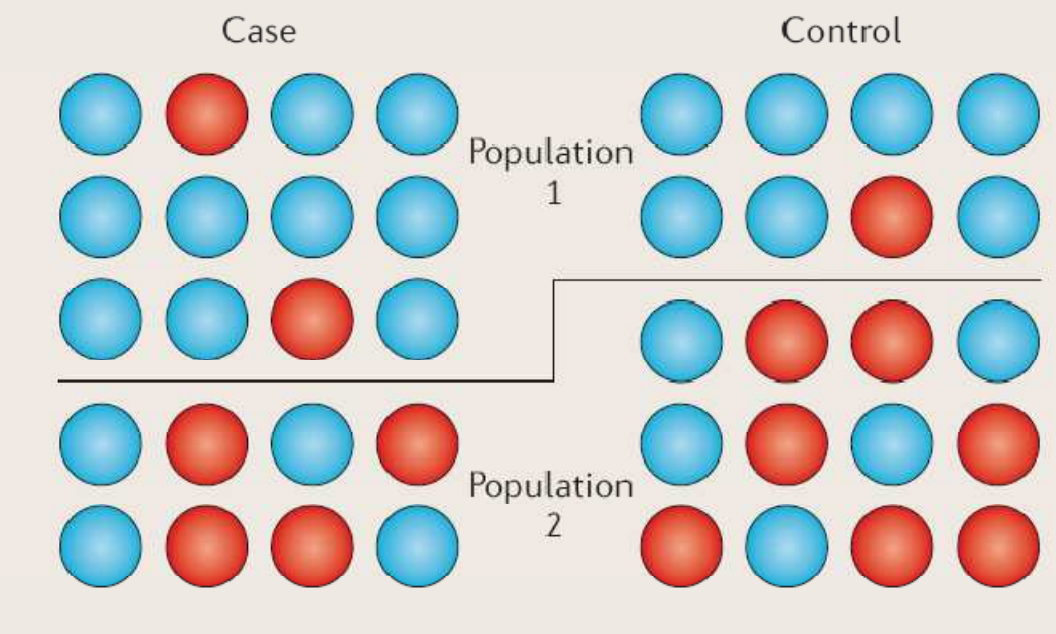
- Choices to be made:
  - Enter multiple markers in one model
    - Analyze the markers as independent contributors (see earlier example R code)
    - Analyze the markers as potentially interacting (see Chapter 9)
  - Construct haplotypes from multiple tightly linked markers and analyze accordingly
- All these analyses are easily performed in a “regression” context
  - In particular, for case / control data, logistic regression is used, where disease status is regressed on genetic predictors

Multiple marker testing is NOT the same as testing for epistasis

- Epistasis = gene-gene interaction (learn more about this later)
- Gene-Gene interaction studies come in different flavors (Marchini et al., 2005)
  - Disease associated interactions among unlinked markers.
  - Search over all pairs of loci on genome
  - Two-stage strategy
    - first stage => search loci meeting with lenient threshold;
    - second stage => test interaction between screened loci with strict threshold.
- Power to detect interaction is affected by many factors (as before), including allele frequency at the disease-associated loci and LD between the markers and disease-associated loci.

5 Dealing with population stratification

5.a Spurious associations



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

5.b Genomic Control

- In Genomic Control (GC), a 1-df association test statistic (usually, CAT) 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 $\lambda > 1$ the test statistics are divided by λ.
- There is an analogous procedure for a general (2 df) test; The method can also be applied to other testing approaches.
- 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).

5.c Structured Association methods

- 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.
- These 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….

5.d Other approaches to handle the effects of population substructure

Include extra covariates in regression models used for association modeling/testing

- Null SNPs can mitigate the effects of population structure when included as covariates in regression analyses.
- Like GC, this approach does not explicitly model the population structure and is computationally fast, but it is much more flexible than GC because epistatic and covariate effects can be included in the regression model.
- Empirically, the logistic regression approaches show greater power than GC, but their type-1 error rate must be determined through simulation.
- Simulations can be quite intensive! How many replicates are sufficient?

Principal components analysis

- When many null markers are available, principal components analysis provides a fast and effective way to diagnose population structure.
- In European data, the first 2 principal components nicely reflect the N-S and E-W axes

Unrelateds are “distantly” related

- Alternatively, a mixed-model approach that involves estimated kinship, with or without an explicit subpopulation effect, has recently been found to outperform GC in many settings.
- Given large numbers of null SNPs, it becomes possible to make precise statements about the (distant) relatedness of individuals in a study so that in theory it should be possible to provide a complete solution to the problem of population stratification.

6 Multiple testing

6.a General setting

Introduction

- Multiple testing is a thorny issue, the bane of statistical genetics.
  - The problem is not really the number of tests that are carried out: even if a researcher only tests one SNP for one phenotype, if many other researchers do the same and the nominally significant associations are reported, there will be a problem of false positives.
- The genome is large and includes many polymorphic variants and many possible disease models. Therefore, any given variant (or set of variants) is highly unlikely, *a priori*, to be causally associated with any given phenotype under the assumed model.
- So strong evidence is required to overcome the appropriate scepticism about an association.

6.b Controlling the overall type I error

Frequentist paradigm

- The frequentist paradigm of controlling the overall type-1 error rate sets a significance level α (often 5%), and all the tests that the investigator plans to conduct should together generate no more than probability α of a false positive.
- In complex study designs, which involve, for example, multiple stages and interim analyses, this can be difficult to implement, in part because it was the analysis that was planned by the investigator that matters, not only the analyses that were actually conducted.

Frequentist paradigm

- In simple settings the frequentist approach gives a practical prescription:
  - if n SNPs are tested and the tests are approximately independent, the appropriate per-SNP significance level α' should satisfy

$$\alpha = 1 - (1 - \alpha')^n,$$

    which leads to the Bonferroni correction $\alpha' \approx \alpha / n$.
- For example, to achieve $\alpha = 5\%$ over 1 million independent tests means that we must set $\alpha' = 5 \times 10^{-8}$. However, the *effective number* of independent tests in a genome-wide analysis depends on many factors, including sample size and the test that is carried out.

When markers (and hence tests) are tightly linked

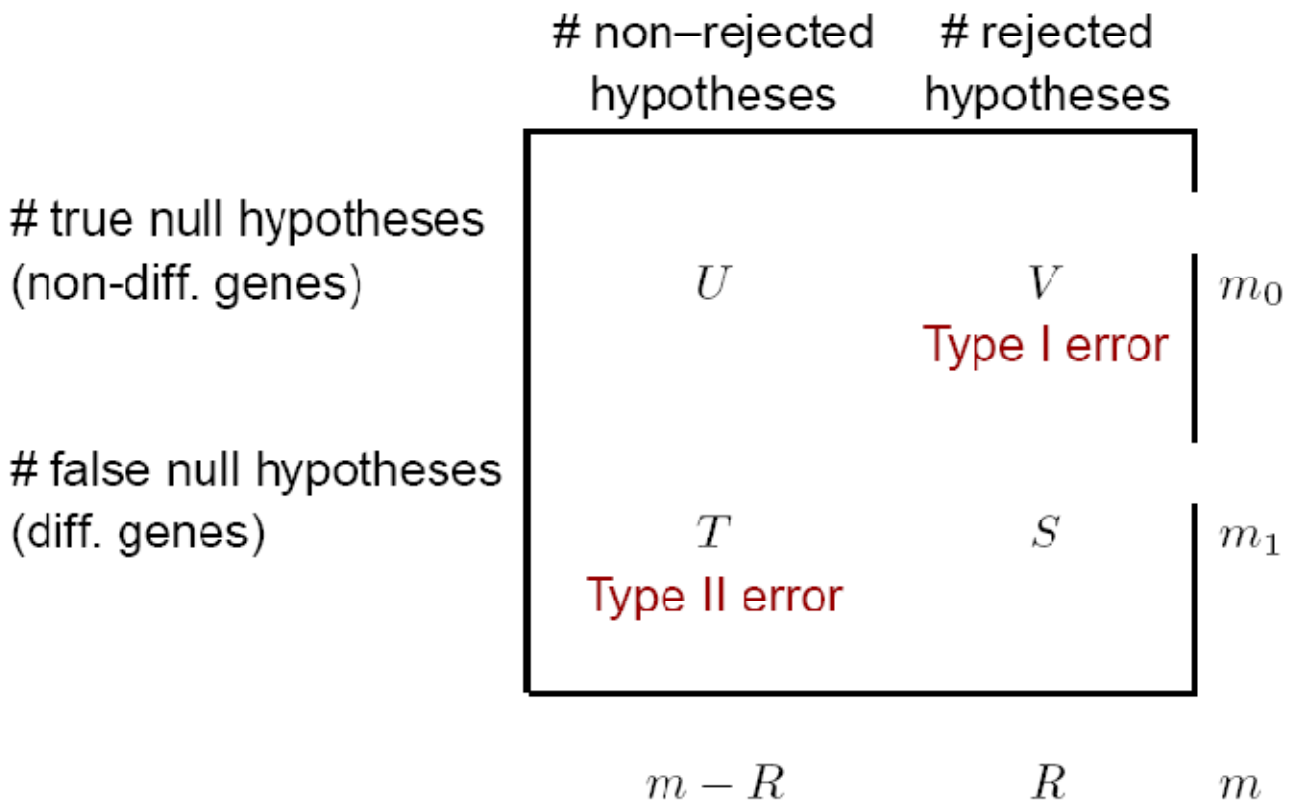
- For tightly linked SNPs, the Bonferroni correction is conservative.
- A practical alternative is to approximate the type-I error rate using a permutation procedure.
  - Here, the genotype data are retained but the phenotype labels are randomized over individuals to generate a data set that has the observed LD structure but that satisfies the null hypothesis of no association with phenotype.
  - By analysing many such data sets, the false-positive rate can be approximated.
  - The method is conceptually simple but can be computationally demanding, particularly as it is specific to a particular data set and the whole procedure has to be repeated if other data are considered.

The 5% magic percentage

- Although the 5% global error rate is widely used in science, it is inappropriately conservative for large-scale SNP-association studies:
  - Most researchers would accept a higher risk of a false positive in return for greater power.
- There is no “rule” saying that the 5% value cannot be relaxed, but another approach is to monitor the false discovery rate (FDR) instead
- The FDR refers to the *proportion of false positive test results among all positives*.

FDR control

- In particular,



(Benjamini and Hochberg 1995: FDR=E(Q); Q=V/R when R>0 and Q=0 when R=0)

FDR control

- FDR measures come in different shapes and flavor.
  - But under the null hypothesis of no association, *p*-values should be uniformly distributed between 0 and 1;
  - FDR methods typically consider the actual distribution as a mixture of outcomes under the null (uniform distribution of p-values) and alternative (P-value distribution skewed towards zero) hypotheses.
  - Assumptions about the alternative hypothesis might be required for the most powerful methods, but the simplest procedures avoid making these explicit assumptions.

Cautionary note

- The usual frequentist approach to multiple testing has a serious drawback in that researchers might be discouraged from carrying out additional analyses beyond single-SNP tests, even though these might reveal interesting associations, because all their analyses would then suffer a multiple-testing penalty.
- It is a matter of common sense that expensive and hard-won data should be investigated exhaustively for possible patterns of association.
- Although the frequentist paradigm is convenient in simple settings, strict adherence to it can be dangerous: true associations may be missed!
    - Under the Bayesian approach, there is no penalty for analysing data exhaustively because the prior probability of an association should not be affected by what tests the investigator chooses to carry out.

7 Assessing the function of genetic variants

Criteria for assessing the functional significance of a variant

Criteria	Strong support for functional significance	Moderate support for functional significance	Evidence against functional significance
Nucleotide sequence	Variant disrupts a known functional or structural motif	Variant is a missense change or disrupts a putative functional motif; changes to protein structure might occur	Variant disrupts a non-coding region with no known functional or structural motif
Evolutionary conservation	Consistent evidence from multiple approaches for conservation across species and multigene families	Evidence for conservation across species or multigene families	Nucleotide or amino-acid residue not conserved
Population genetics	In the absence of laboratory error, strong deviations from expected population frequencies in cases and/or controls in a particular ethnicity	In the absence of laboratory error, moderate to small deviations from expected population frequencies in cases and/or controls; effects are not well characterized by ethnicity	Population genetics data indicates no deviations from expected proportions
Experimental evidence	Consistent effects from multiple lines of experimental evidence; effect in human context is established; effect in target tissue is known	Some (possibly inconsistent) evidence for function from experimental data; effect in human context or target tissue is unclear	Experimental evidence consistently indicates no functional effect
Exposures (for example, genotype–environment interaction studies)	Variant is known to affect the metabolism of the exposure in the relevant target tissue	Variant might affect metabolism of the exposure or one of its components; effect in target tissue might not be known	Variant does not affect metabolism of exposure of interest
Epidemiological evidence	Consistent and reproducible reports of moderate-to-large magnitude associations	Reports of association exist; replication studies are not available	Prior studies show no effect of variant

(Rebbeck et al 2004)

8 Proof of concept

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-7}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point P values between 10^{-5} and 5×10^{-7}) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes observed at most loci identified. This study thus represents a thorough validation of the GWA approach. It has also demonstrated that careful use of a shared control group represents a safe and effective approach to GWA analyses of multiple disease phenotypes; has generated a genome-wide genotype database for future studies of common diseases in the British population; and shown that, provided individuals with non-European ancestry are excluded, the extent of population stratification in the British population is generally modest. Our findings offer new avenues for exploring the pathophysiology of these important disorders. We anticipate that our data, results and software, which will be widely available to other investigators, will provide a powerful resource for human genetics research.

References:

- 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-
- Balding D 2006. A tutorial on statistical methods for population association studies. Nature Reviews Genetics, 7, 781-791. *(also good background reading material)*

Background reading:

- Hardy et al 2009. Genomewide association studies and human disease. NEJM 360: 1786-.
- 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-
- Ensenauer et al 2003. Primer on medical genomics. Part VIII: essentials of medical genetics for the practicing physician

In-class discussion documents

- Lunetta 2008. Genetic association studies. *Circulation:* 118: 96-101
- Hardy et al. 2009. Genomewide association studies and human disease N Engl J Med 360;17
- Cordell et al 2005. Genetic Epidemiology 3: Genetic association studies. *The Lancet*; 366: 1121–31