Introduction to Genetic Association Analysis

David Balding
Professor of Statistical Genetics
University of Melbourne, and
University College London

1 Feb 2016

1. Preliminaries

2. Single-SNP tests of association

3. Diagnostic plots

4. Bayesian association methods

5. Multi-SNP methods

1 Preliminaries

2 Single-SNP tests of association

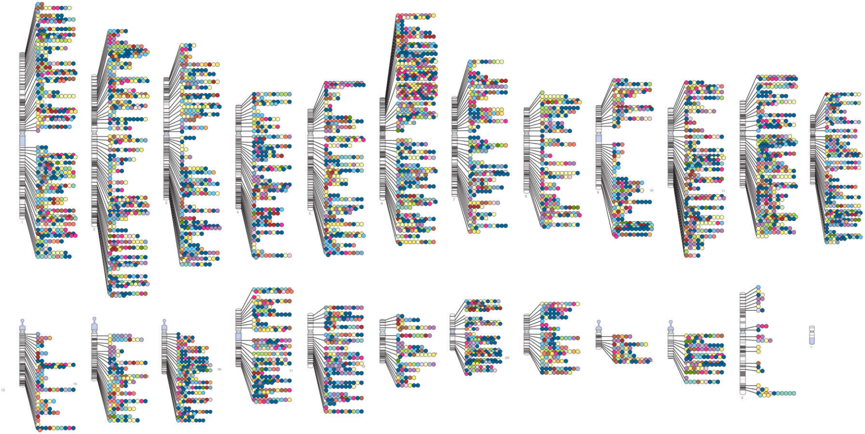
3 Diagnostic plots

4 Bayesian association methods

5 Multi-SNP methods

Single-SNP Association Analysis

Many associations have been identified through single-SNP analysis



GWAS Catalog - https://www.ebi.ac.uk/gwas/diagram

Statistical methods for association analysis

Let Y be a vector of phenotypes and X_j the corresponding vector of genotypes at the jth SNP, Each vector is typically of length 10^3 to 10^4 (number of individuals, i) and j can range up to several millions. We seek to identify population correlations between Y and one or more of the X_j.

Y is usually binary (usually case-control) or quantitative (continuous);

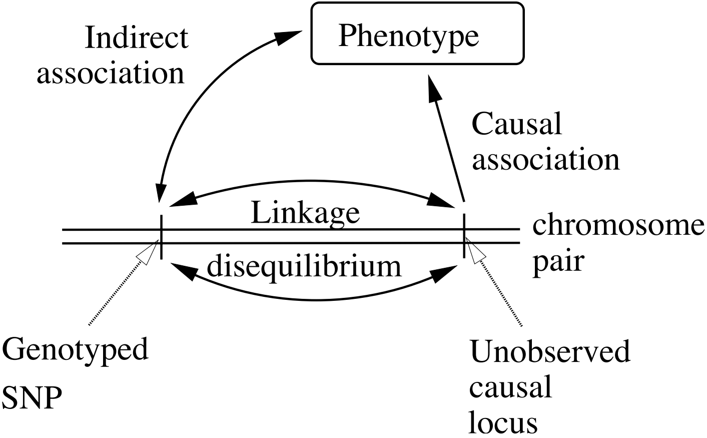
- It can also be (ordered) categorical or a count, or it can be multivariate (not considered further in this course).
- SNPs are linearly ordered along chromosomes and neighbouring SNPs may be in high linkage disequilibrium (LD) due to co-inheritance of chromosome fragments from remote ancestors;
- unlinked SNPs can also be in LD e.g. due to pop. structure.

Association analysis can be aimed at finding a:

Model selection: identify SNPs that each tag a variant causally associated with the phenotype; focus is on control of false positives, or

Prediction: identify SNPs that give optimal out-of-sample prediction of phenotype (assessed e.g. using cross-validation).

Linkage disequilibrium is our friend and our enemy



LD is essential for detecting ungenotyped causal variants using SNP data (not for sequence data):

- the higher the LD, the more likely we are to detect a causal signal,
- but the harder it is to fine-map (precisely locate) the causal variant.

1. Preliminaries
2. Single-SNP tests of association
3. Diagnostic plots
4. Bayesian association methods
5. Multi-SNP methods

Tests of association: case-control study[1]

Test independence of rows and columns in a 2 × 3 contingency table.

- Pearson (2 df) χ^2 test;
- Fisher exact test.

Genotype	0	1	2	total
Case	89	369	342	800
Control	56	250	266	572
total	145	619	608	1 372

R code:

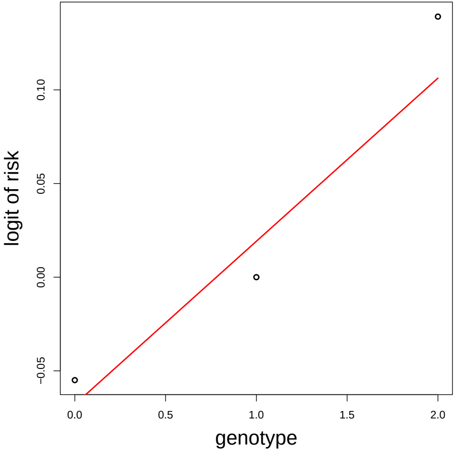
```
> cas = c(89,369,342); con = c(56,250,266)
> chisq.test(matrix(c(cas,con),3,2))
Pearson's Chi-squared test
X-squared = 2.0551, df = 2, p-value = 0.3579
> fisher.test(matrix(c(cas,con),3,2))
Fisher's Exact Test for Count Data
p-value = 0.3607
```

[1] For further details on this section see: Balding DJ, A tutorial on statistical methods for population association studies, *Nat Rev Genet* 7(10): 781-791, 2006.

(Cochran)-Armitage Trend Test (ATT)

Pearson and Fisher tests both allow the genotypic relative risks to take any value – general genetic model. But some patterns of association (genetic models) are more plausible than others.

The ATT is a single-SNP test for a linear trend of relative risk with genotype (coded as 0,1,2). `R` example:



```
> prop.trend.test(cas,cas+con)
Chi-squared Test for Trend in Proportions
X-squared = 1.9853, df = 1, p-value = 0.1588
```

Linear regression for continuous (quantitative) traits

```
plink --linear --bfile <data> --out <output> --pheno <phenfile>
```

For each j, we fit

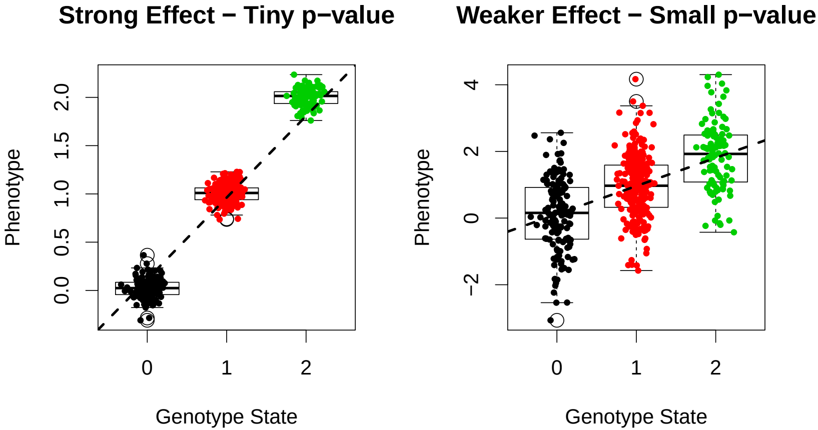
$$Y = \mu + \beta_j X_j + \epsilon$$

where ϵ is a vector of independent and identically-distributed (iid) noise variables with variance σ^2.

Solve using least squares regression: find μ and β_j that minimise $\sum_{i=1}^{n}(Y_i - \mu - X_{ij}\beta_j)^2$ where i indexes individuals.

To test whether $H_0 : \beta_j = 0$ can be rejected at each j, compute a p-value using Likelihood Ratio Test (LRT), or the Wald, Score or F Tests.

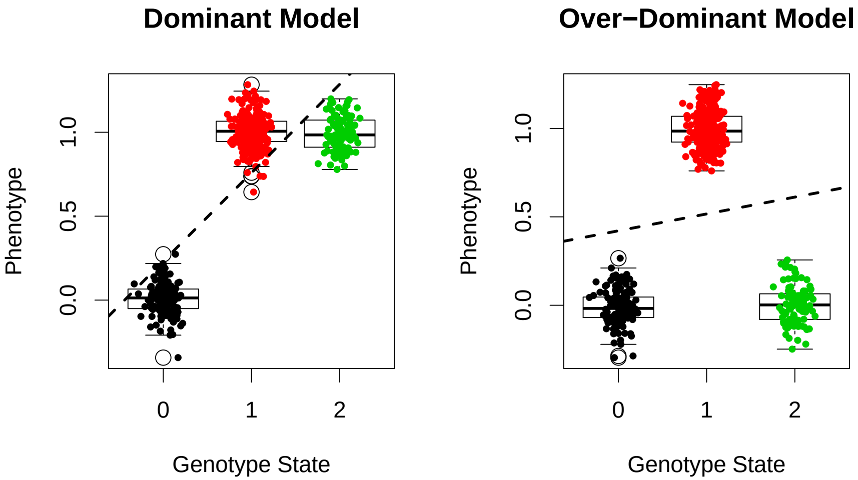
- Typically, LRT is most powerful.
- PLINK uses by default Wald, but LRT can be implemented using --assoc



The above tests are sensitive to additive differences in means:

- i.e. they have most power when mean heterozygote phenotype is midway between the two homozygotes, as illustrated above;
- this is a consequence of the 0/1/2 coding of genotypes, could test e.g. recessive or dominant models by changing the genotype coding.

1df non-additive models



Linear regression has low/no power to detect such non-linear genetic effects.

A more general model (2 parameters/degrees of freedom)

To allow for more general deviations away from additivity (= linearity), we can define the mean effect at each genotype in terms of μ (overall mean), a (additive effect) and d (dominant effect):

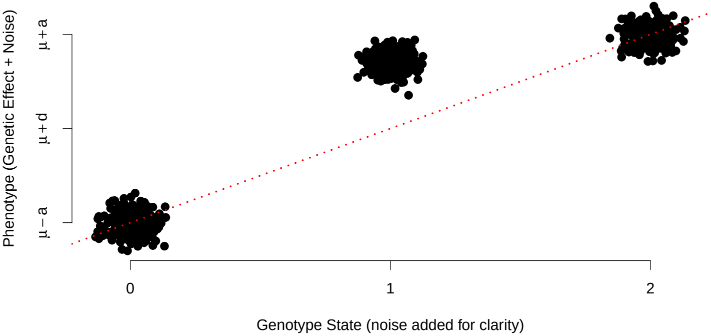
Genotype (X)	$\mathbb{E}[Y\|X]$
0	$\mu - a$
1	$\mu + d$
2	$\mu + a$

We can write this in a formula as

$$Y = \mu + (X_j - 1)a + X_j(2 - X_j)d + \epsilon = \mu - a + X_j(2d + a) - X_j^2 d + \epsilon$$

It is common to still assume the same variance for each phenotype (as for the linear model), but there is also interest in looking for changes in variance across phenotypes, instead of or in addition to changes in mean.

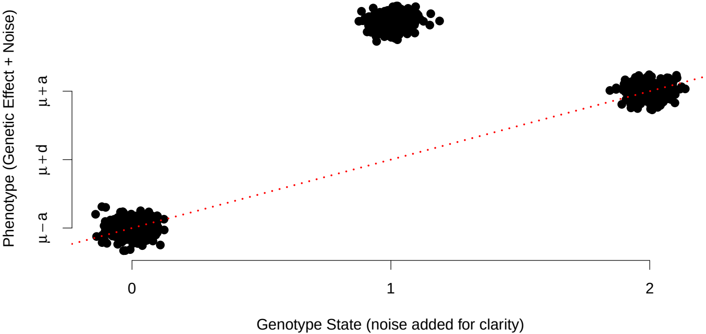
Dominant model: $d > 0$



Full dominance occurs when $d = a$.

Recessive model has $d < 0$; recessiveness is the same as dominance but with a switch of reference allele.

Over-dominant model: $d > a$



Testing for non-additive effects

In most GWAS testing for non-additivity ($d \neq 0$) is not performed:

- Most association signals found to date have been close to linear;
  - this is in part a circular argument - they have generally only looked for additive effects.
- Although dominance is common in Mendelian genetics, LD causes non-additive component of association to decay rapidly
  - testing for effects that have a low prior probability to be real inflates the false positive rate.
  - this justification for ignoring $d \neq 0$ is less convincing for studies with very dense genotype or sequence data.
- Apparent non-additive effects might be due to genotype error
  - additive signals of association are considered to be more reliable;

NB counting alleles instead of genotypes to create a 1 df test sensitive to additive genetic effects is **not** recommended.

Linear regression wth covariates

Can include covariates in any regression model (implemented in PLINK by adding --covar). This extends the linear model to

$$Y = \theta_1 Z_1 + \theta_2 Z_2 + \ldots + \beta_j X_j = \theta Z + \beta_j X_j$$

Z_1 is (automatically) set to a vector of ones, and θ_1 is the global mean (previously denoted μ)

The estimates of regression coefficients for covariates can be suppressed in the PLINK output with

```
plink --linear hide-covar --bfile <data> --out <output> \
 --pheno <phenfile> --covar <covarfile>
```

PLINK by default includes sex as a covariate when testing the X chromosome. Can be turned off using --no-sex

Logistic regression for binary (case/control) phenotypes

```
plink --logistic --bfile <data> --out <output> --pheno <phenfile>
```

Suppose Y is a vector of phenotypes coded as 1 or 2 e.g. to indicate case/control status. For each SNP in turn, consider

$$\pi = f(\theta Z + \beta_j X_j)$$

where $\pi = \mathrm{P}(Y{=}1 \,|\, X_j)$ and f is a function that maps the real line into the unit interval (0,1); f (or more correctly its inverse f^{-1}) is called the link function. Most common is the logistic link function

$$f^{-1}(\pi) = \log\left(\frac{\pi}{1-\pi}\right), \qquad \text{the inverse of} \qquad f(x) = \frac{\exp(x)}{1+\exp(x)}.$$

With this link function (called logistic regression) β_j is the log odds ratio

comparing genotypes that differ by 1 at the jth SNP:

$$\beta_j = \log\left(\frac{\pi_2/(1-\pi_2)}{\pi_1(1-\pi_1)}\right) = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_0(1-\pi_0)}\right)$$

where $\pi_k = \mathrm{P}(Y{=}1|X_j{=}k)$ for $k \in \{0, 1, 2\}$.

Logistic regression versus Pearson and ATT tests

ATT can be derived as score tests of $\beta = 0$ under the above logistic regression model

- This assumes genotype is coded as 0, 1 and 2, so that β is then scalar and the genetic model is additive on the logistic scale.
- A general genetic model on the logistic scale can be encoded (in various ways) using two variables, so that β is a vector of length two.

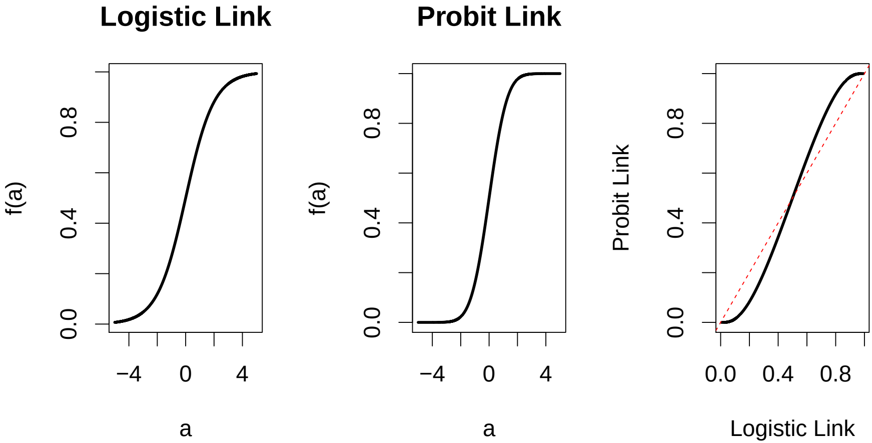
Above regression models correspond to "prospective" ascertainment: we condition on genotype and treat phenotype as the outcome. Case-control studies are usually retrospective: individuals are ascertained according to phenotype and then genotype is observed.

- Fortunately, there is theory to show that analysis based on prospective model is usually OK for retrospective data[2]
- In some settings there can be advantages to conditioning on phenotype ("reversing the regression").

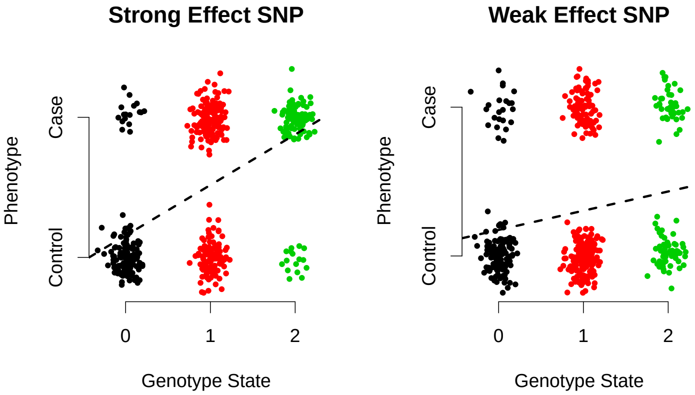
[2] Prentice R, Pyke R, Logistic disease incidence models and case-control studies. *Biometrika* 66: 403-411, 1979.

Probit regression

The probit link function is the inverse of the Gaussian (Normal) cumulative distribution function.



Linear regression for binary data



The dashed lines result from fitting $Y = \mu + \beta_j X_j + \epsilon$. NB Y is restricted to two values and so ϵ is far from Gaussian and the fitted values $\hat{\mu} + \hat{\beta}_j X_j$ are not constrained to lie in (0,1). However, in practice this works well provided that π is not too close to 0 or 1. It is particularly useful for mixed model association analysis (MMAA, more later).

1 Preliminaries

2 Single-SNP tests of association

3 Diagnostic plots

4 Bayesian association methods

5 Multi-SNP methods

Manhattan plots and (log) P-P Plots

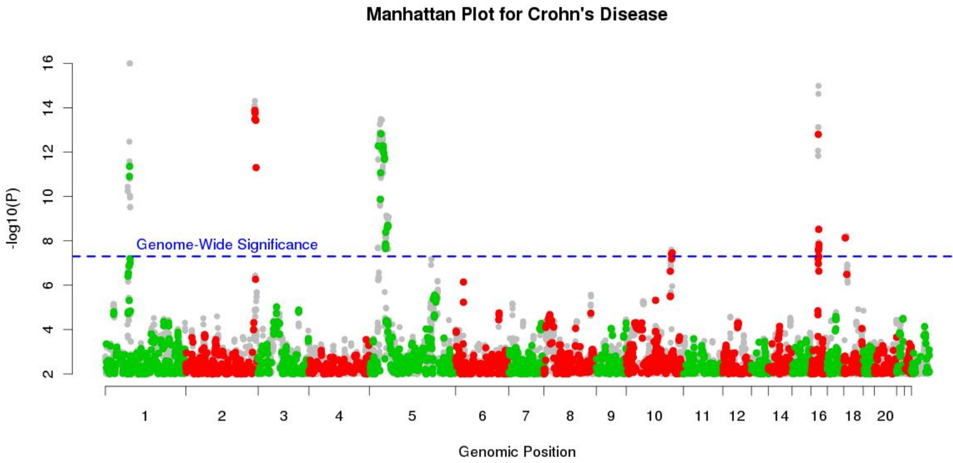
Manhattan plots show $-\log_{10}(p\text{-value})$ of SNPs ordered along the genome:

- displays genomically-local patterns of association (useful to zoom into interesting regions)
- a lonely significant SNP can suggest genotyping error.

P-P plots show $-\log_{10}(p\text{-value})$ of SNPs ordered by p-value.

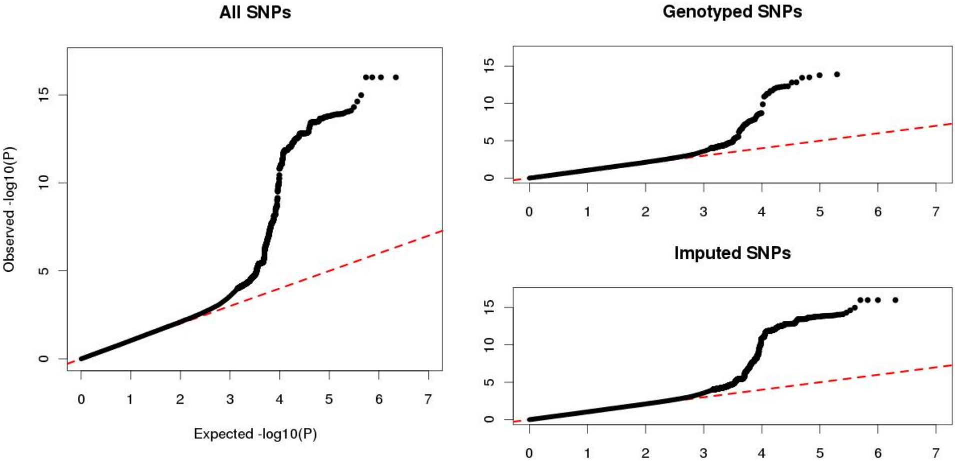
- A large number of points above the diagonal suggests genome-wide inflation of test statistics due e.g. to population structure.
- Can investigate pattern of association for different SNP categories, e.g. genotyped vs imputed.
- Q-Q plot is similar but shows test statistic values rather than $-\log_{10}(p\text{-value})$.

Manhattan plot of 1 *p*-value per SNP, genome-wide



Only SNPs with $P < 0.01$ are shown (> 2 on the $-\log_{10}$ scale used on the x-axis). Green and red colouring is used to distinguish chromosomes. Grey points represent imputed SNPs.

(log) P-P Plots



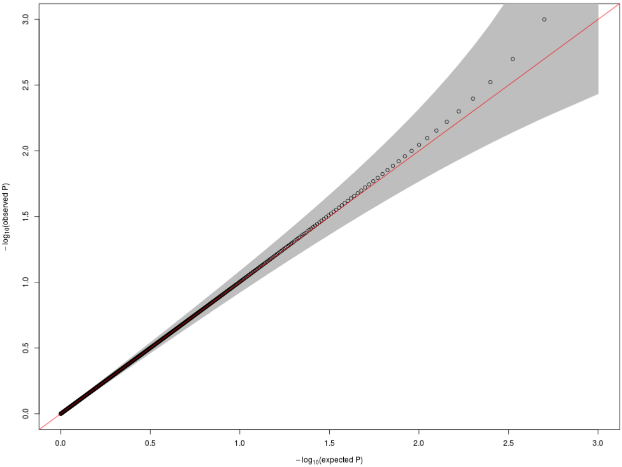
All SNPs

Genotyped SNPs

Imputed SNPs

It may be preferable to thin SNPs prior to creating plot.

(log) P-P Plots



Can add approximate (point-wise) or simultaneous 95% confidence intervals using (e.g.) R package qqplot

1. Preliminaries

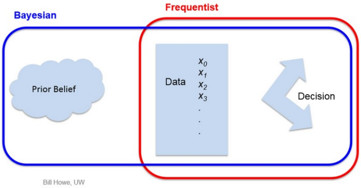
2. Single-SNP tests of association

3. Diagnostic plots

4. Bayesian association methods

5. Multi-SNP methods

Bayesian vs Frequentist



Bill Howe, UW

Frequentist methods focus on performance of methods under repeats of the analysis with imagined new datasets drawn from the assumed sampling distribution.

Bayesian methods condition on the observed data and seek a probability distribution for the unknown of interest. To obtain this *posterior distribution* they start with a *prior distribution* based on other available information, and then update it based on the data using Bayes Theorem.

Bayesian Statistics

Bayesian methods are criticised because conclusions depend on the choice of prior distribution, and encoding background information always has a subjective element. They can also be computationally demanding.

However,

- Bayesian methods generally answer the right question;
- the impact of the prior can be examined through sensitivity analysis;
- frequentist approaches are often equivalent to Bayesian analyses with unexamined and sometimes unreasonable priors;
- subjectivity is unavoidable and better to make it explicit;
- computational issues are being reduced.

Computing the posterior probability of association (PPA)

Wellcome Trust Case Control Consortium (2007)[3] was the first major GWAS to report Bayes Factors (BF):

$$\mathrm{BF} = \frac{P(\mathrm{data}|H_1)}{P(\mathrm{data}|H_0)}$$

under both strictly additive model and a general model that gives most weight to near-additive models. Then, to compute the PPA:

$$\mathrm{PPA} = \frac{\pi \mathrm{BF}}{1{-}\pi + \pi \mathrm{BF}} \qquad \text{where} \quad \pi = \frac{P(H_1)}{P(H_0)}.$$

π may vary across SNPs, depending on MAF, proximity to genes of interest, conservation across species,.... Typically $\pi \approx 10^{-4}$ (so *a priori* about 0.3 Mb of the genome has some true association).

The Bayesian solution to the problem of choosing the genetic model is to average the BF, weighted according to the plausibilities of different models.

[3] WTCCC, Genome-wide association study of 14K cases of seven common diseases and 3K shared controls, *Nature*, 447:661-78, 2007.

Weighting additive and non-additive models in BF

		p-value		BF	PPA	
Trait	SNP	Trend	General	($\log_{10}$)	$\pi = 10^{-4}$	$\pi = 10^{-5}$
BD	rs420259	2.2×10^{-4}	6.3×10^{-8}	4.1	0.56	0.11
CD	rs9858542	7.7×10^{-7}	3.6×10^{-8}	4.7	0.83	0.33
T2D	rs9939609	5.2×10^{-8}	1.9×10^{-7}	5.3	0.95	0.67
CD	rs17221417	9.4×10^{-12}	4.0×10^{-11}	8.9	0.99999	0.99987
T1D	rs17696736	2.2×10^{-15}	1.5×10^{-14}	12.5	1.00000	1.00000

Here, BF is computed as a 4:1 weighting of additive and general models (as defined by WTCCC 2007).

- 1st row: $\log_{10}(\text{BF}) = 2.0$ (additive model); taking $\pi = 10^{-4}$, PPA $=$ 0.01; likely to be ignored.
- Under general mode, $\log_{10}(\text{BF}) = 4.8$ and PPA $= 0.86$.
- But, general model often not tested – additive tests preferred.
- With 4:1 weighting, $\log_{10}(\text{BF}) = 4.1$, and PPA$=$0.56.
- Only 20% weight given to general model, but BF captures strong non-additive signal while still emphasising additivity.
- Don't calculate PPA for many models and pick the largest!

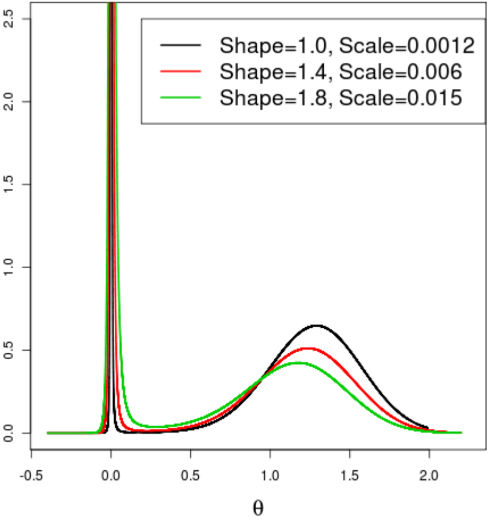
Effect sizes under an additive model

- Under the additive model, WTCCC assumed a $N(0, 0.2)$ prior on effect size (log odds).
- A drawback of this is rapid decay in the tails.

Example: effect of prior.

- the SEARCH collaborative group (08) reporting that variants in SLC01B1 are associated with statin-induced myopathy.
- most significant SNP is rs4363657, with $p = 4.1 \times 10^{-9}$.
- Using WTCCC prior, PPA ≈ 0.02
- Other Bayesian analyses with more plausible priors give (e.g. mixture of Gaussians) PPA ≈ 0.4.
- Big influence of prior, because data suggest very large effect size for a rare allele: WTCCC says this is *a priori* implausible.
- $p < 10^{-8}$ is conventionally regarded as highly significant, but Bayesian analysis says we should be far from convinced.

Estimate effect size instead of testing[4]



- Division of SNPs into null or non-null is artificial;
- reality is a distribution of effect sizes that puts much weight near zero;
- can be modelled using Normal-Exponential-Gamma (NEG) prior, and posterior density obtained numerically;
- no BF in this approach, but posterior $P(|\theta| > 0.1) =$ 0.47, 0.39 and 0.35.

[4] For further details see: Stephens M, Balding DJ, Bayesian statistical methods for genetic association studies. *Nat Rev Genet*, 10(10), 681-690, 2009.

1. Preliminaries

2. Single-SNP tests of association

3. Diagnostic plots

4. Bayesian association methods

5. Multi-SNP methods

Testing multiple SNPs and Interactions

Regression models can include multiple SNPs and covariates; interactions among SNPs (G×G), or SNPs with covariates (G×E). **Problem**: too many predictors → overfitting, discussed further below. Reasons to include interactions in models:

1. to help identify causal factors; better models can allow more convincing evidence for association;
2. to clarify mechanisms of effect, especially for E factors;
3. improved prediction.

In humans, few convincing reports of significant interactions to date:

- replication difficult;
- huge space of possible hypotheses: very strong signal required to overcome low prior/strong multiple testing penalty.
- easier to establish main effects first, then look for interactions.

Pathway-based analyses correspond to many weak G×G.
Case-only designs can be effective for G×E.

Haplotype analyses

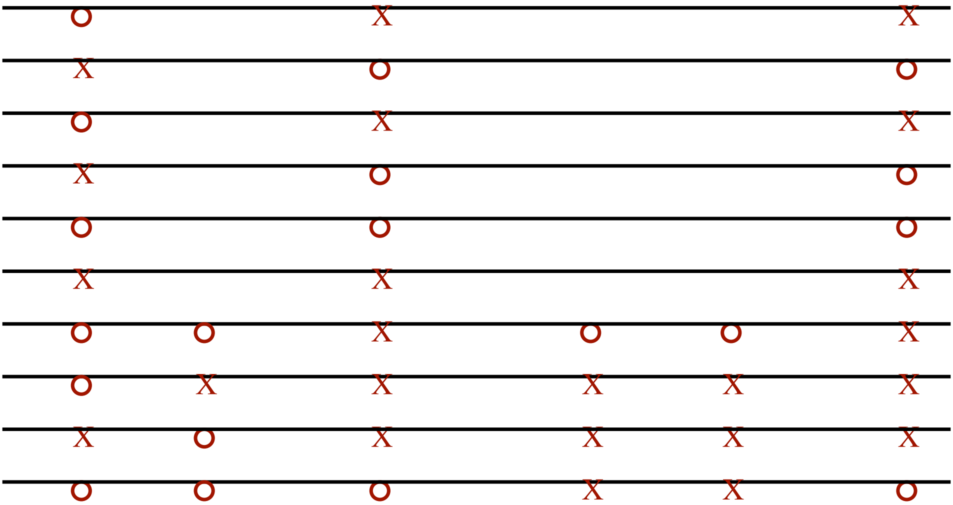
- In single-SNP tests, only one genotyped SNP at a time can "tag" a causal variant.
- The **haplotype block** model of the human and other genomes views the genomes as consisting of many high-LD blocks, usually separated by recombination hotspots.
- This suggests the use of **haplotypes** rather than SNP alleles, e.g. as predictors in regression models
  - gametic phase can be inferred from population samples; need to incorporate uncertainty in subsequent inferences;
  - problem of too many haplotypes, many of them rare, can be reduced e.g. via clustering;
  - haplotype analyses thought to be advantageous for capturing effect of multiple causal variants in *cis*;
  - many different approaches/software, e.g. UNPHASED (Dudbridge 08).

Haplotype methods enjoyed limited success and were complicated. Success of imputation methods has largely replaced interest in haplotype analyses.

Imputation

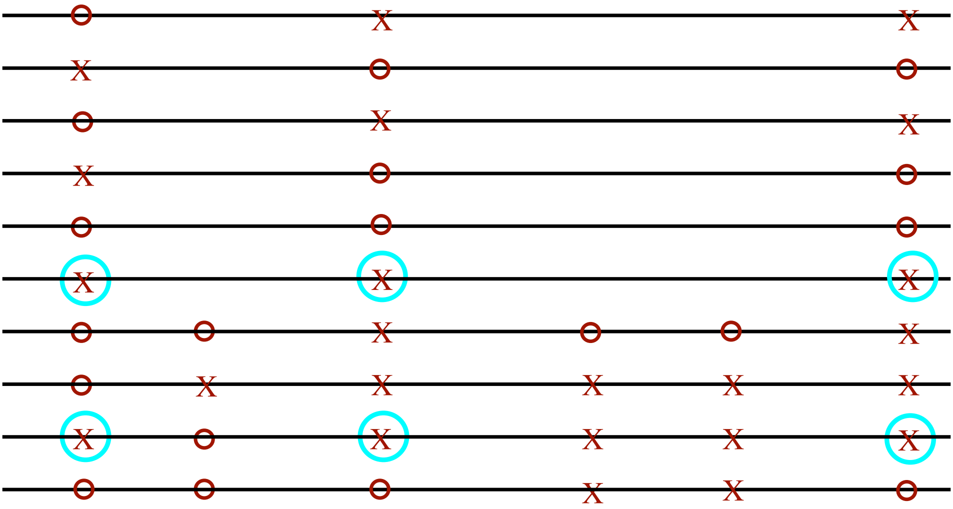
- Imputation is estimation of missing values in a dataset.
- It is widely used in the analysis of epidemiological studies
  - for many analyses, an individual's entire record might be discarded because of a few missing values
  - imputation allows the available information to be analysed.
- In genetic epidemiology, because of strong LD between tightly-linked SNPs, imputation is usually highly accurate in
  - replacing sporadic missing values
  - checking for genotyping errors.
- It is so effective that it can even be used to impute *all* the genotypes at a SNP that was not genotyped in the study.
  - The information on LD at the missing SNP comes from an external reference panel, such as HapMap or 1K Genomes, in which the SNP has been genotyped.
- Using this approach the number of SNPs available for analysis can be increased from, say, 0.5M actually genotyped SNPs up to several million genotyped + imputed SNPs.

A simplified view of genotype imputation: 1

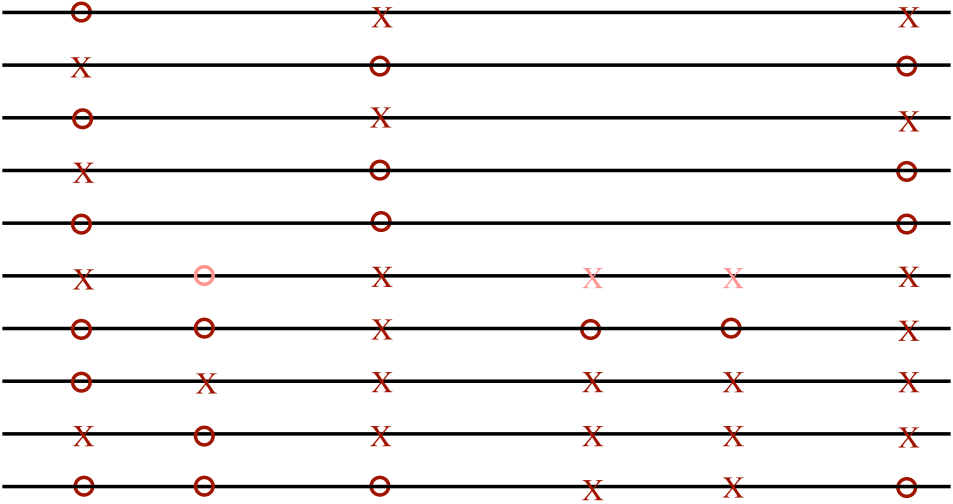


o		x			x
x		o			o
o		x			x
x		o			o
o		o			o
x		x			x
o	o	x	o	o	x
o	x	x	x	x	x
x	o	x	x	x	x
o	o	o	x	x	o

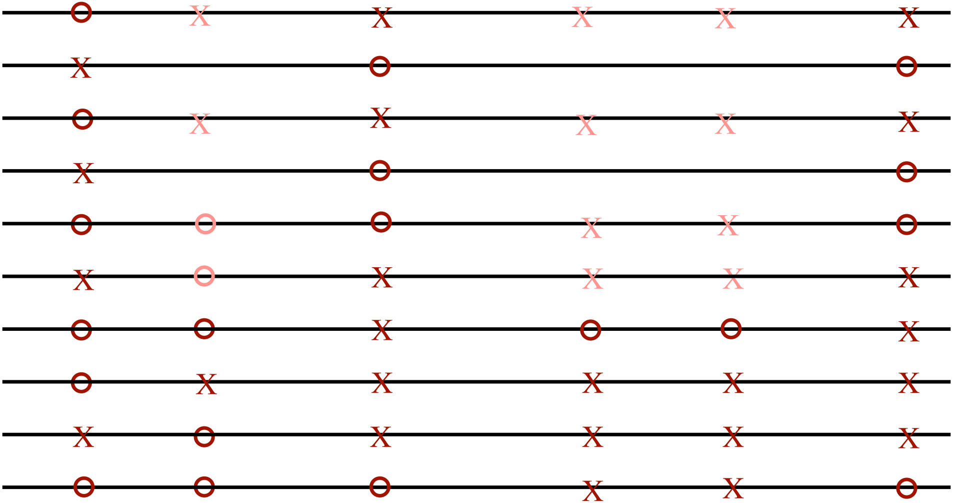
A simplified view of genotype imputation: 2



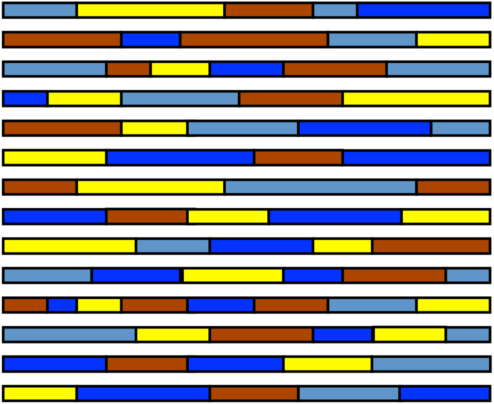
A simplified view of genotype imputation: 3



A simplified view of genotype imputation: 4



Population genetics models for imputation



- Imputation methods are based on the idea that, due to small effective population size of our ancestors, each modern chromosome can be viewed as a mosaic of copies of a small number of ancestral chromosomes
- Most popular statistical model to implement this is a **Hidden Markov Model (HMM)**.

- Emitted states are the observed alleles;
- Hidden states are the ancestral chromosomes[5]
  - change according to a Markov model usually at recombination hotspots
- Solving for the hidden states gives genotypes at missing SNPs.

[5] Li N, Stephens M (2003) *Genetics*,165(4): 2213-33.

Probabilistic genotype calls

Imputation programs give a probability distribution for the genotype at the missing SNP

- e.g. AA 98%; AC 1.5%; CC 0.5%

The variance of the probability distribution over genotypes (coded as 0,1,2) gives a measure of imputation quality

- (1/3,1/3,1/3) ⇒ high variance, poor quality
- (0.01,0.99,0.00) ⇒ low variance, high quality if model well calibrated.

Analysis can be based on

- Most likely genotype
- Dosage = expected number of minor alleles
  - e.g. $1 \times 0.015 + 2 \times 0.005 = 0.025$
- Integrating over the probability distribution.

Dosage is simplest, and works well

- but requires an additive model of association.

Software for imputation

- IMPUTE2 (Howie, Donnelly, Marchini 2009) gives very good performance; use in conjunction with SHAPEIT (Delaneau, Zagury, Marchini, 2013) for phasing.
- MACH[6]. Nearly best performance, good support, runs quickly.
- BEAGLE: Browning & Browning Very fast, slightly worse performance.

We will provide a guide to SHAPEIT and IMPUTE2 in Module 18.

[6] Fuchsberger C, Abecasis G, Hinds D. minimac2: faster genotype imputation. Bioinformatics 2014