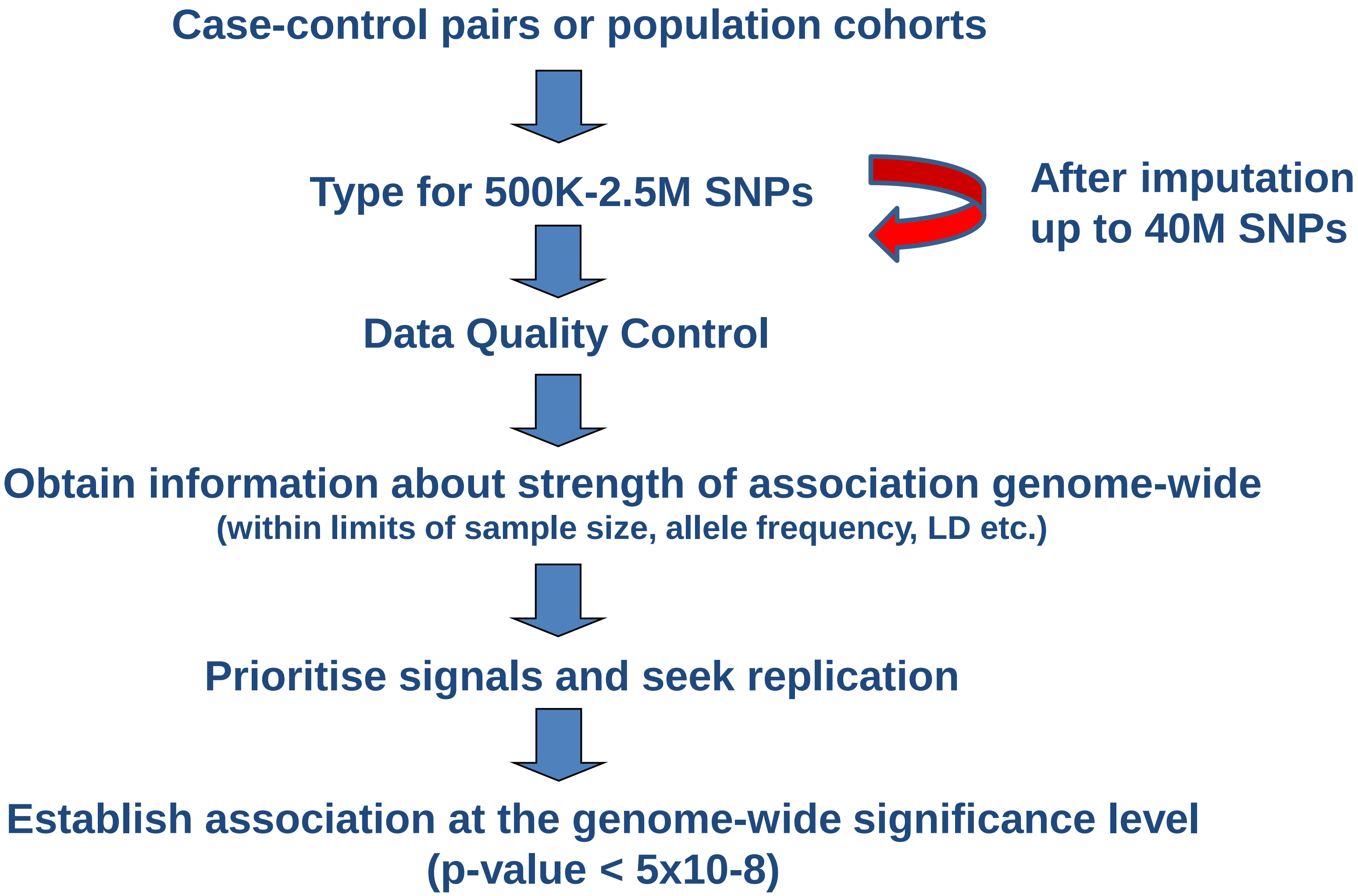
The statistics of genome-wide association studies

Ioanna Tachmazidou
Wellcome Trust Sanger Institute
it3@sanger.ac.uk

Recap: GWAS Principles



Case-control pairs or population cohorts

Type for 500K-2.5M SNPs

After imputation up to 40M SNPs

Data Quality Control

Obtain information about strength of association genome-wide

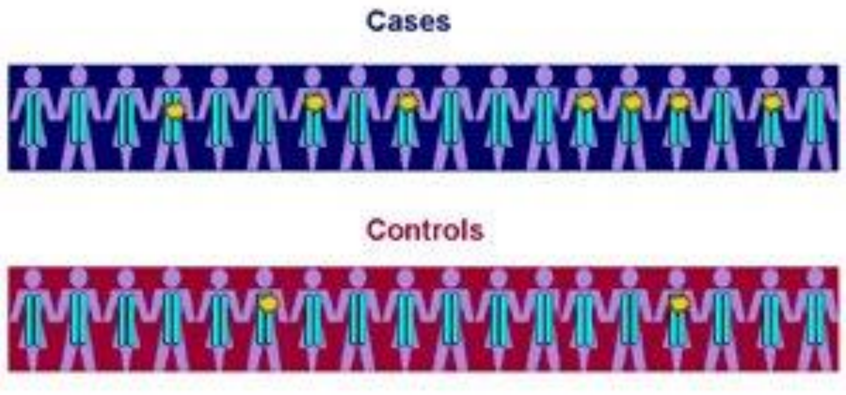
(within limits of sample size, allele frequency, LD etc.)

Prioritise signals and seek replication

Establish association at the genome-wide significance level

(p-value < 5x10-8)

Association between a genetic variant and disease



DD: minor homozygote

Dd: heterozygote

dd: major homozygote

Marker genotype	Affected	Unaffected	Total
DD	n_{2A}	n_{2U}	$n_{2.}$
Dd	n_{1A}	n_{1U}	$n_{1.}$
dd	n_{0A}	n_{0U}	$n_{0.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

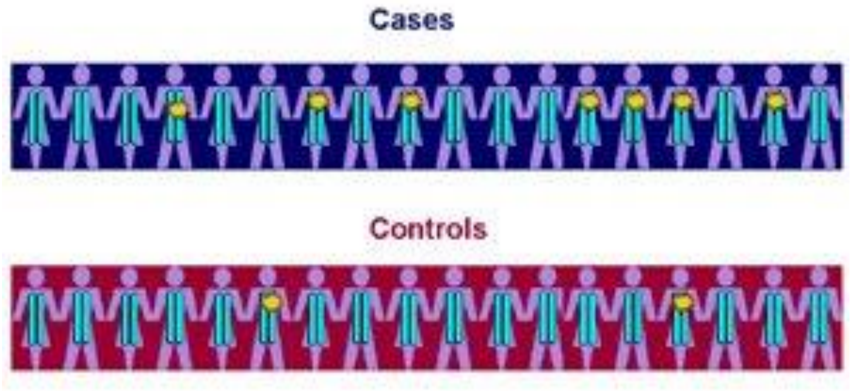
The odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. Therefore the odds ratio for genotype DD relative to dd is:

$$\mathrm{OR}(DD : dd) = \frac{\text{odds of an individual with genotype DD carrying the disease}}{\text{odds of an individual with genotype dd carrying the disease}}$$

or else:

$$\mathrm{OR}(DD : dd) = \frac{n_{2A}/n_{2U}}{n_{0A}/n_{0U}}$$

3 x 2 contingency tables



DD: minor homozygote

Dd: heterozygote

dd: major homozygote

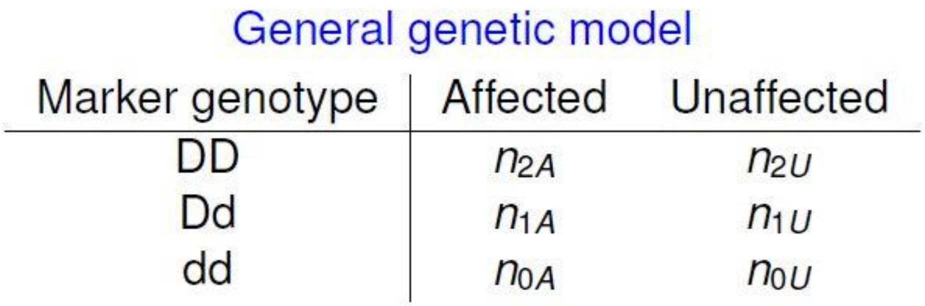
Marker genotype	Affected	Unaffected	Total
DD	n_{2A}	n_{2U}	$n_{2.}$
Dd	n_{1A}	n_{1U}	$n_{1.}$
dd	n_{0A}	n_{0U}	$n_{0.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

Under the null hypothesis of no disease-marker association, the rows and columns of the contingency table are independent:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

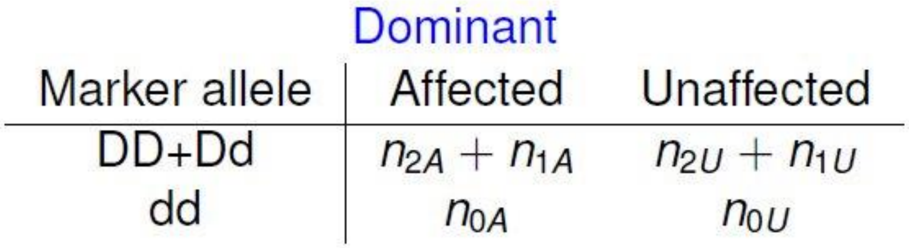
$$X^2 = \sum_{i=0,1,2} \sum_{j=A,U} \frac{(n_{ij} - \mathrm{E}(n_{ij}))^2}{\mathrm{E}(n_{ij})}, \text{ where } \mathrm{E}(n_{ij}) = \frac{n_{i.} n_{.j}}{n_{..}} \sim \chi^2_2$$

Models of inheritance



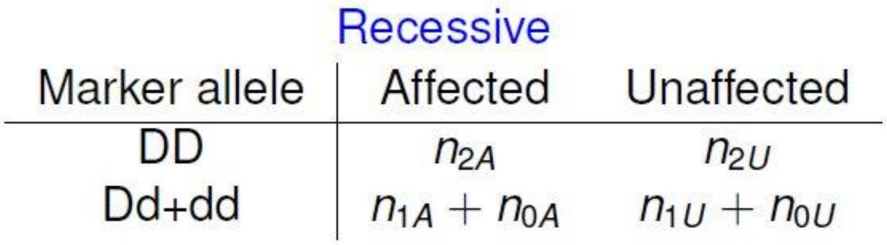
General genetic model

Marker genotype	Affected	Unaffected
DD	n_{2A}	n_{2U}
Dd	n_{1A}	n_{1U}
dd	n_{0A}	n_{0U}



Dominant

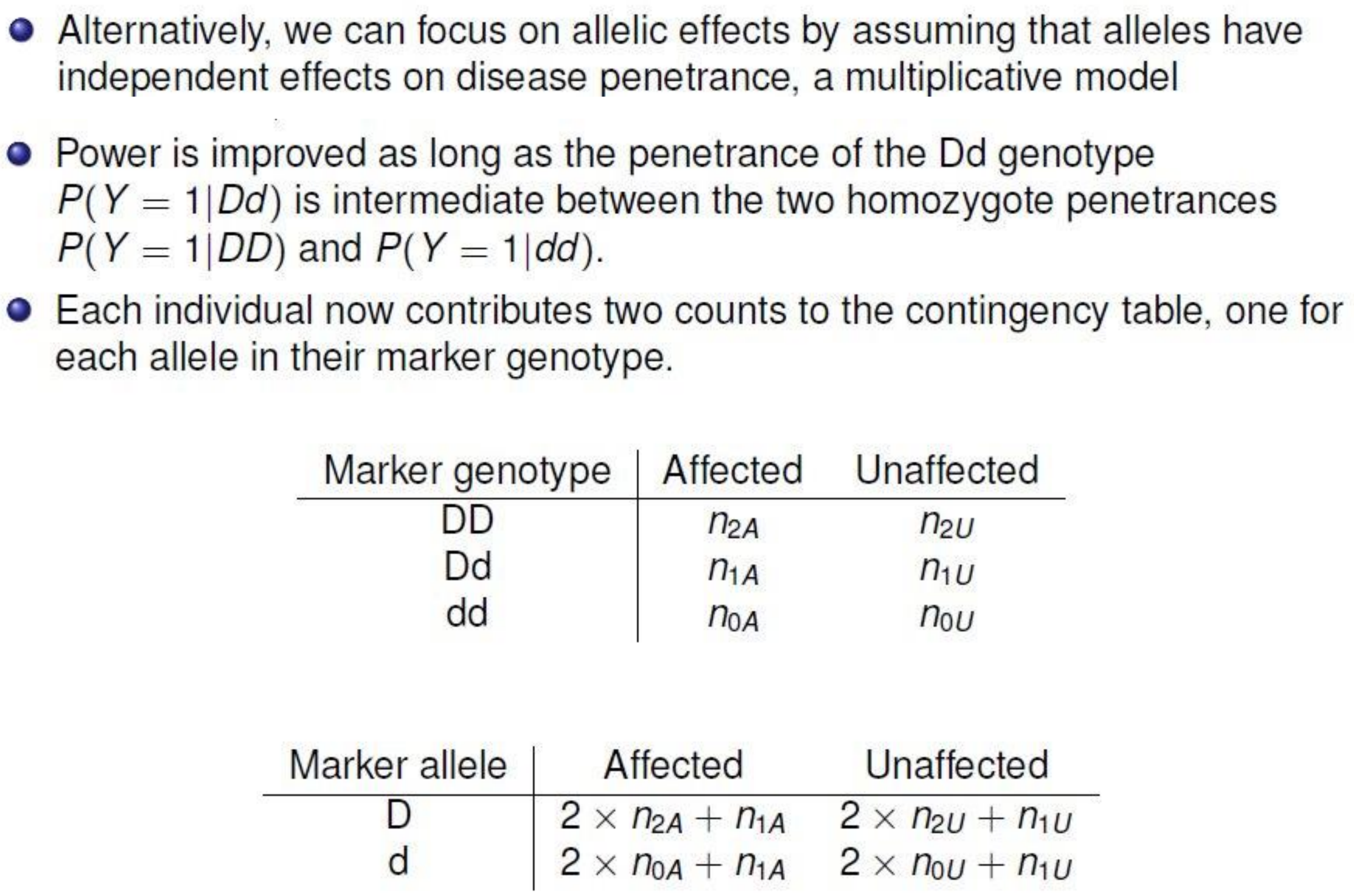
Marker allele	Affected	Unaffected
DD+Dd	$n_{2A} + n_{1A}$	$n_{2U} + n_{1U}$
dd	n_{0A}	n_{0U}



Recessive

Marker allele	Affected	Unaffected
DD	n_{2A}	n_{2U}
Dd+dd	$n_{1A} + n_{0A}$	$n_{1U} + n_{0U}$

2 x 2 contingency tables

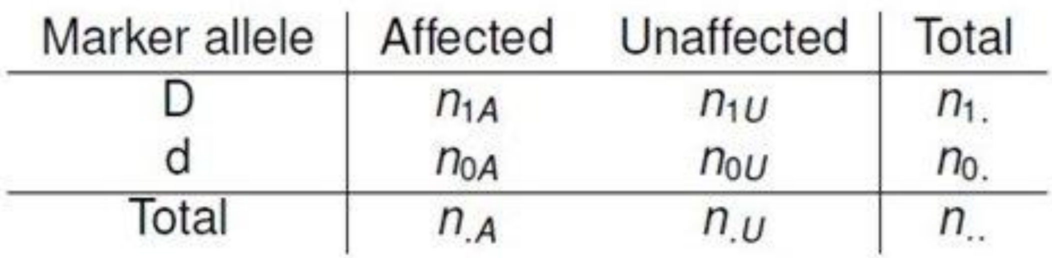


- Alternatively, we can focus on allelic effects by assuming that alleles have independent effects on disease penetrance, a multiplicative model
- Power is improved as long as the penetrance of the Dd genotype $P(Y=1|Dd)$ is intermediate between the two homozygote penetrances $P(Y=1|DD)$ and $P(Y=1|dd)$.
- Each individual now contributes two counts to the contingency table, one for each allele in their marker genotype.

Marker genotype	Affected	Unaffected
DD	n_{2A}	n_{2U}
Dd	n_{1A}	n_{1U}
dd	n_{0A}	n_{0U}

Marker allele	Affected	Unaffected
D	$2 \times n_{2A} + n_{1A}$	$2 \times n_{2U} + n_{1U}$
d	$2 \times n_{0A} + n_{1A}$	$2 \times n_{0U} + n_{1U}$

2 x 2 contingency tables



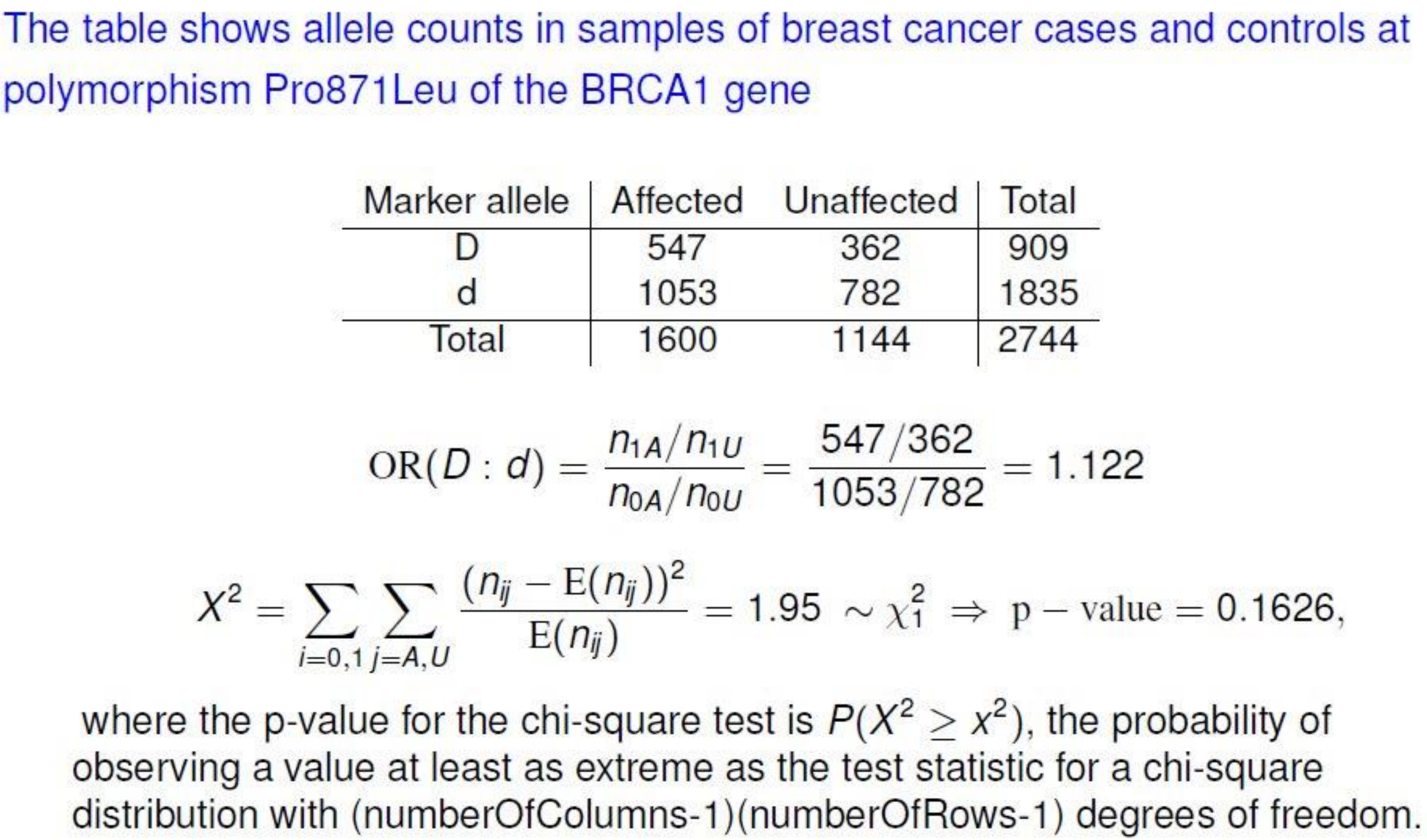
Marker allele	Affected	Unaffected	Total
D	n_{1A}	n_{1U}	$n_{1.}$
d	n_{0A}	n_{0U}	$n_{0.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

$$\mathrm{OR}(D:d) = \frac{n_{1A}/n_{1U}}{n_{0A}/n_{0U}}$$

Under the null hypothesis of no disease-marker association:

$$X^2 = \sum_{i=0,1}\sum_{j=A,U} \frac{(n_{ij} - \mathrm{E}(n_{ij}))^2}{\mathrm{E}(n_{ij})}, \text{where } \mathrm{E}(n_{ij}) = \frac{n_{i.}n_{.j}}{n_{..}} \sim \chi^2_1$$

Example



The table shows allele counts in samples of breast cancer cases and controls at polymorphism Pro871Leu of the BRCA1 gene

Marker allele	Affected	Unaffected	Total
D	547	362	909
d	1053	782	1835
Total	1600	1144	2744

$$\mathrm{OR}(D : d) = \frac{n_{1A}/n_{1U}}{n_{0A}/n_{0U}} = \frac{547/362}{1053/782} = 1.122$$

$$X^2 = \sum_{i=0,1} \sum_{j=A,U} \frac{(n_{ij} - \mathrm{E}(n_{ij}))^2}{\mathrm{E}(n_{ij})} = 1.95 \sim \chi^2_1 \Rightarrow \mathrm{p} - \mathrm{value} = 0.1626,$$

where the p-value for the chi-square test is $P(X^2 \geq x^2)$, the probability of observing a value at least as extreme as the test statistic for a chi-square distribution with (numberOfColumns-1)(numberOfRows-1) degrees of freedom.

We conclude that although the data suggest that the D allele contributes a small amount to breast cancer risk, the data are insufficient to reject the null hypothesis of no effect.

Hardy–Weinberg equilibrium (HWE)

- It is possible to treat the two alleles arising from a genotype as independent observations only under HWE. If population frequency of d is p, the expected genotype frequencies under HWE are:

Genotype	Expected frequencies
DD	$(1-p)^2$
Dd	$2p(1-p)$
dd	p^2

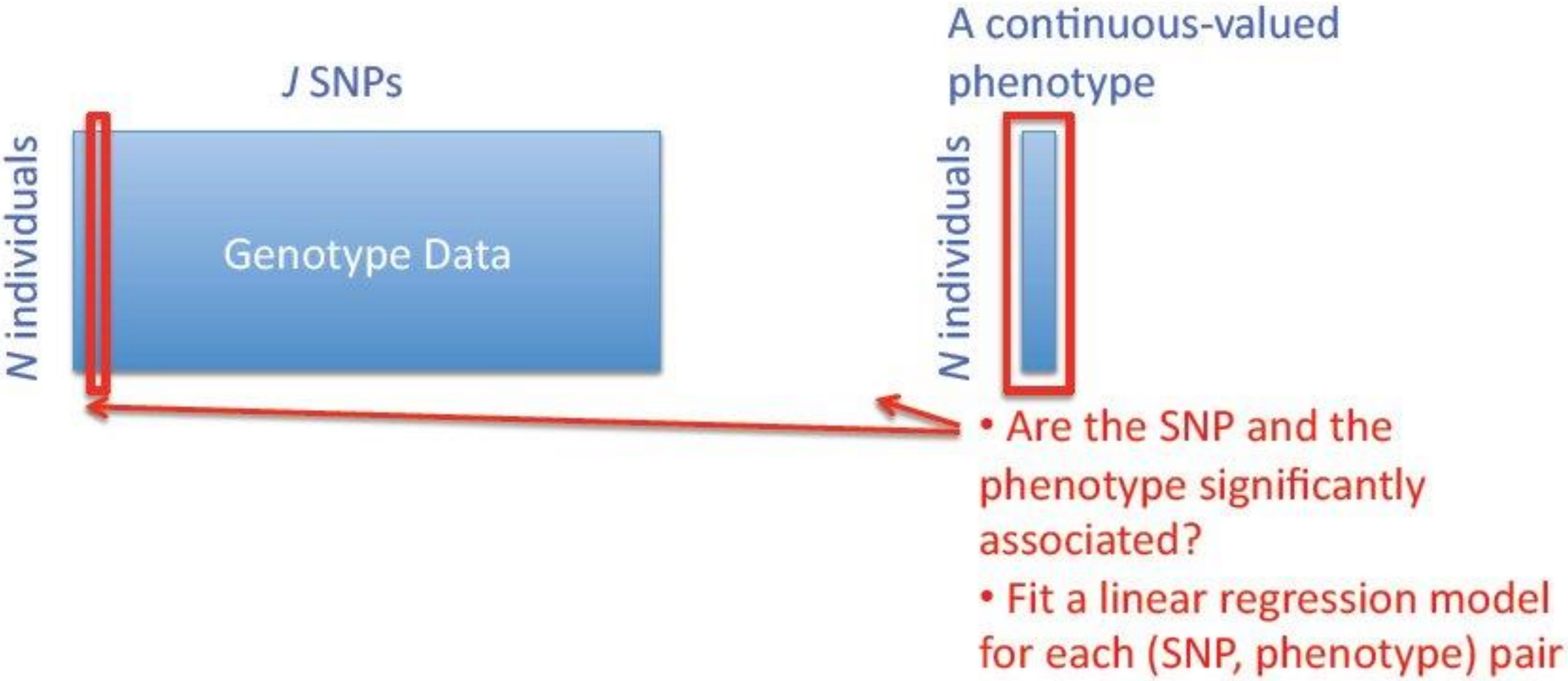
$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

- Control genotypes should be in HWE, provided the population they are selected from is random mating and is large in size.

- Deviations from HWE in controls can arise from:
  - random chance: one of every 20 markers tested will give a p-value <0.05 by chance.
  - genotyping problems: genotypes are consistently miscalled, or specific genotypes give missing values.
  - heterogeneous population: the controls may be a mix of different populations with different allele frequencies.
  - inbreeding or selection.

- Deviations from HWE in cases can be a symptom of disease association.

Association between genotypes and a quantitative trait



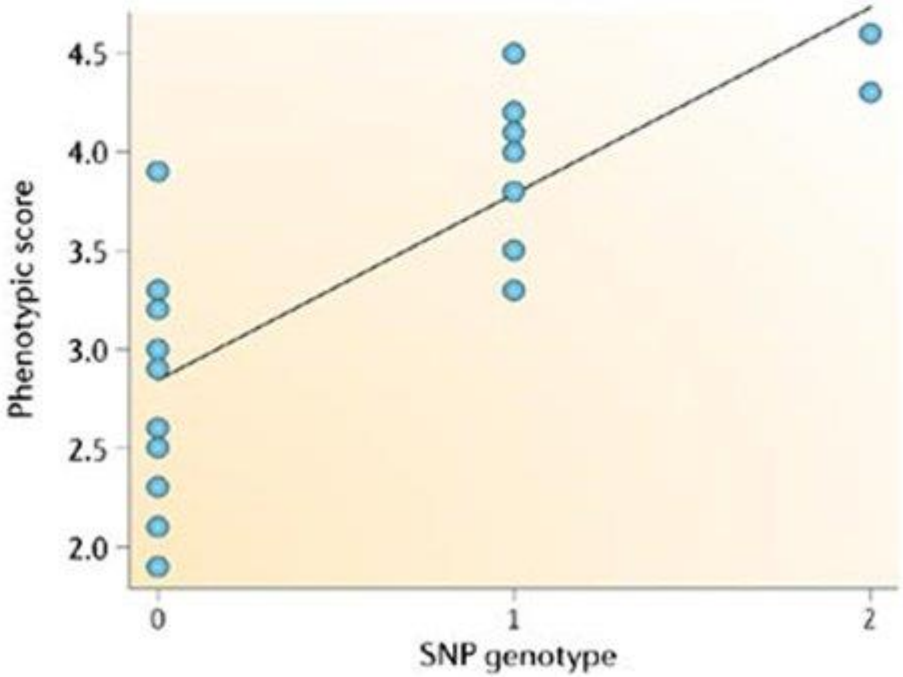
- Are the SNP and the phenotype significantly associated?
- Fit a linear regression model for each (SNP, phenotype) pair

Association between a genetic variant and a quantitative trait

Continuous traits
– Also called quantitative traits
– Cholesterol level, blood pressure etc.

One cannot create a contingency matrix as in case/control studies.

For each locus, fit a linear regression using the number of minor alleles at the given locus of the individual as covariate.



Linear regression model

A linear regression model is defined as

$$y = x\beta_1 + \beta_0 + \varepsilon$$

Data:

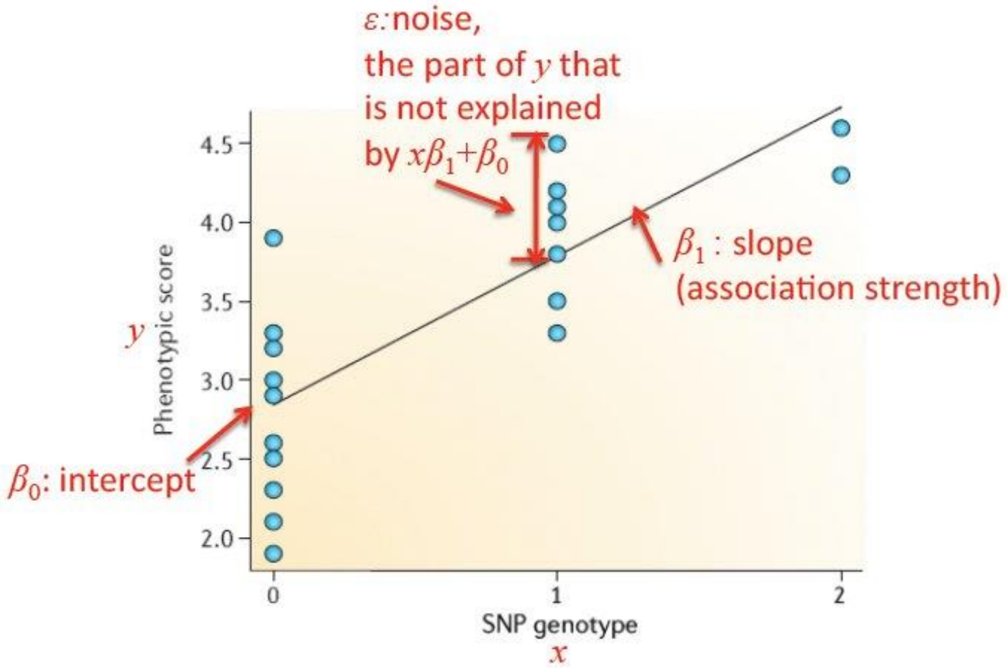
- y: a continuous trait
- x: SNP genotype at a given locus

Parameters:

- β_1: regression coefficient, represents the strength of association between x and y
- β_0: intercept term
- ε: noise or the part of y that is not explained by x (e.g., environmental effect)

Assumptions:

- The individuals in the study are not related
- The phenotype y has a normal distribution



Testing the significance of association using linear models

- We want to test whether to reject the null hypothesis that there is no linear relationship between the given SNP and phenotype:
  - Null hypothesis H_o: $\beta_1=0$
  - Alternative hypothesis H_1: $\beta_1 \neq 0$

- We compute the test statistic:

$$T = \frac{\hat{\beta}_1}{se.(\hat{\beta}_1)}$$

- The test statistic T asymptotically follows the standard normal distribution N(0,1)
- Use this to:
  - Construct a 95% confidence interval for β_1 : $\hat{\beta}_1 \pm 1.96 \times se(\hat{\beta}_1)$
  - Assess the significance of association by calculating a p-value = $2 \times \big(1 - \Phi(|T|)\big)$, where Φ is the cumulative distribution function of N(0,1).

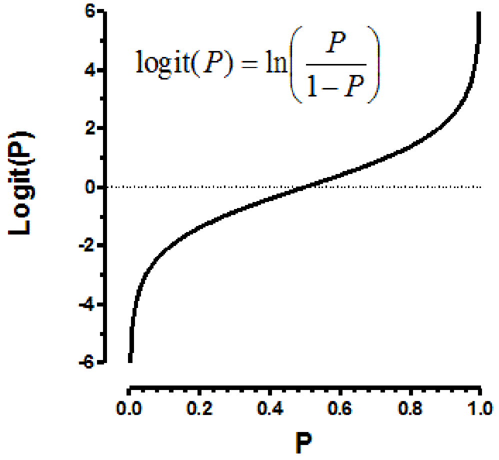
Association between a genetic variant and a binary trait

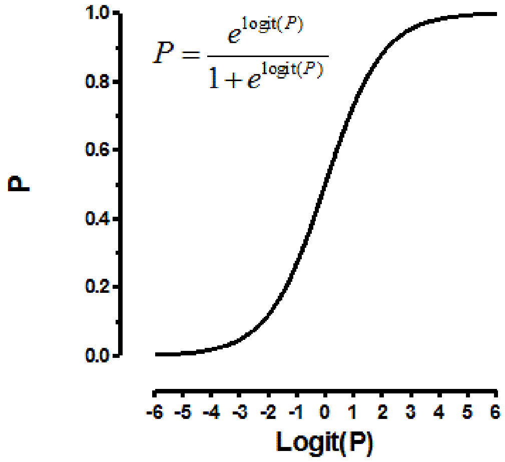
- Logistic regression is an extension of the linear regression model that allows for case/control data to be modelled.

- Suppose N independent random variables Y_1, Y_2, …, Y_N with value 0/1 corresponding to controls/cases.

- We model the probability of individual i being a case given he carries x_i = 0, 1, 2 copies of the minor allele at genotype x_i being a case π_i as:

$$\text{logit}\,(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i$$





Linkage disequilibrium (LD)

- LD is a measure of the degree of correlation between alleles at different loci.
- It refers to the tendency for alleles at closely linked loci on the same chromosome to be associated in the general population.

Marker A Marker B	0	1	
0	p_{00}	p_{01}	$p_{0.}$
1	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	

p_{ij} population proportion of ij haplotypes

$p_{i.}$ and $p_{.j}$ marginal allele proportions

Under linkage equilibrium (LE):

$$p_{ij} = p_{i.} \times p_{.j}$$

A natural measure of LD:

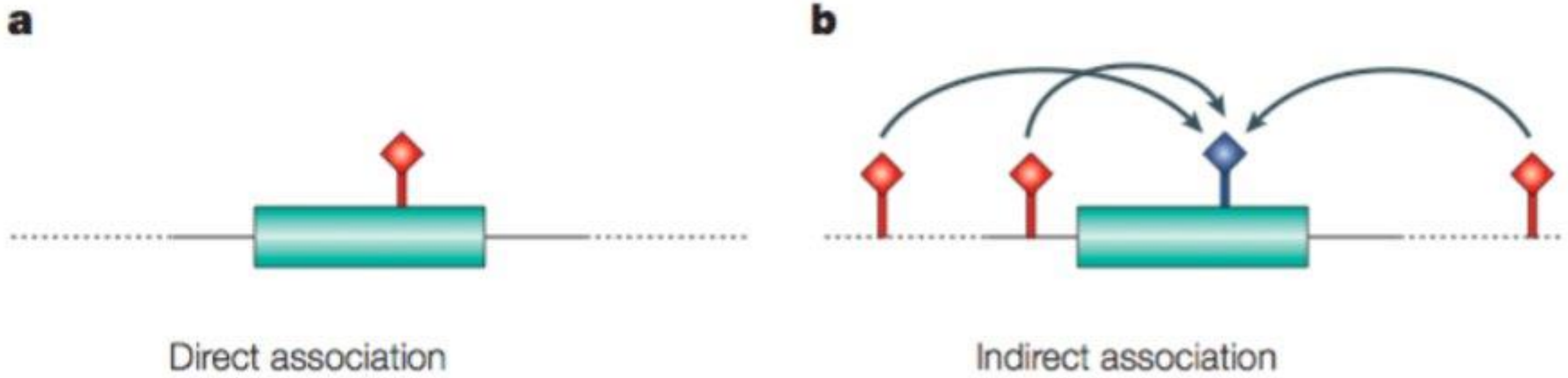
$$D_{ij} = p_{ij} - p_{i.} \times p_{.j}$$

$$r^2 = \frac{\left(p_{ij} - p_{i.} \times p_{.j}\right)^2}{p_{i.} \times p_{.j} \times (1 - p_{i.}) \times (1 - p_{.j})}$$

- Squared correlation coefficient r^2 ranges between 0 and 1:
  - 1 when the two markers provide identical information.
  - 0 when they are in perfect equilibrium.
- Measures loss in efficiency when marker A is replaced with marker B in an association study.

Association does not imply causation

Hirschhorn and Daly: Nature Reviews Genetics 6: 95 (2005)

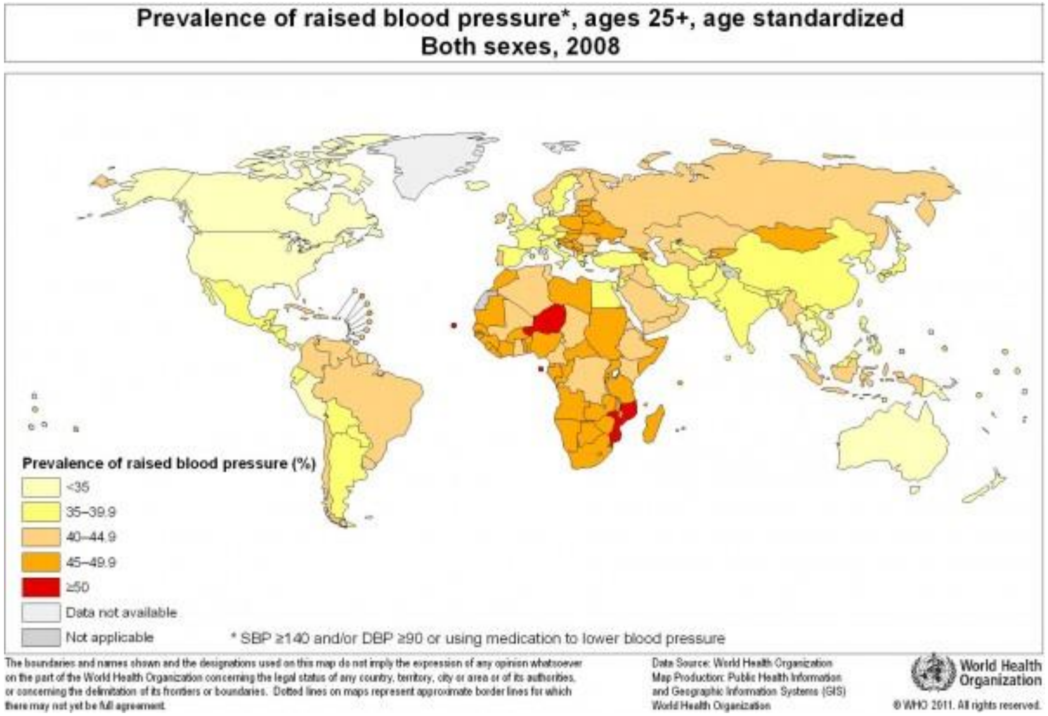


Population stratification

- The problem arises if our underlying population is actually a mix of ancestrally distinct populations with different values of disease prevalence and SNP allele frequency.

- Case-control association studies assume that any difference in SNP genotypes between cases and controls is due solely to their difference in disease status, and not to any difference in genetic background.

- False positive evidence of association will occur at genetic markers that differ in genotype frequencies between the subpopulations.



Prevalence of raised blood pressure*, ages 25+, age standardized
Both sexes, 2008

Prevalence of raised blood pressure (%)
<35
35–39.9
40–44.9
45–49.9
≥50
Data not available
Not applicable
World Health Organization

* SBP ≥140 and/or DBP ≥90 or using medication to lower blood pressure

The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement.

Data Source: World Health Organization
Map Production: Public Health Information and Geographic Information Systems (GIS)
World Health Organization

Genomic control

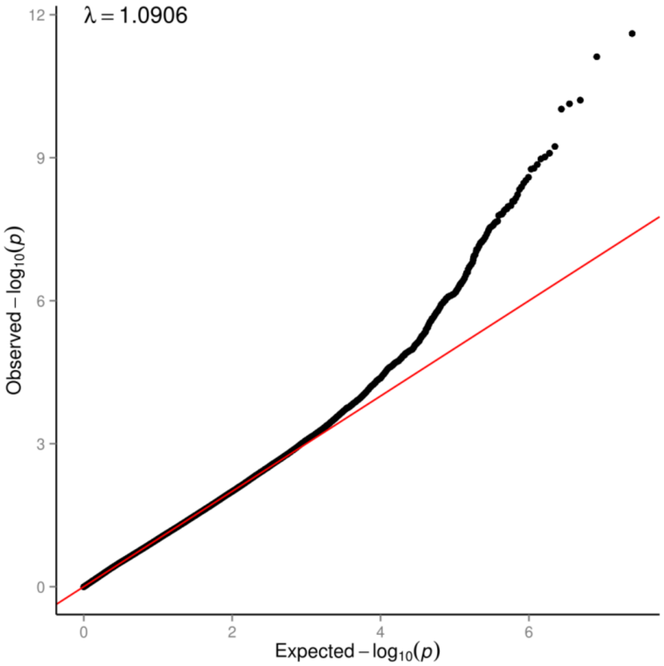
- Under genomic control, population stratification causes the χ^2 association test statistics to be inflated by a constant multiplicative factor λ.

- We can estimate λ as median(X_i^2)/0.456, where 0.456 is the median of the χ^2 test statistic under the null assumption of no association.

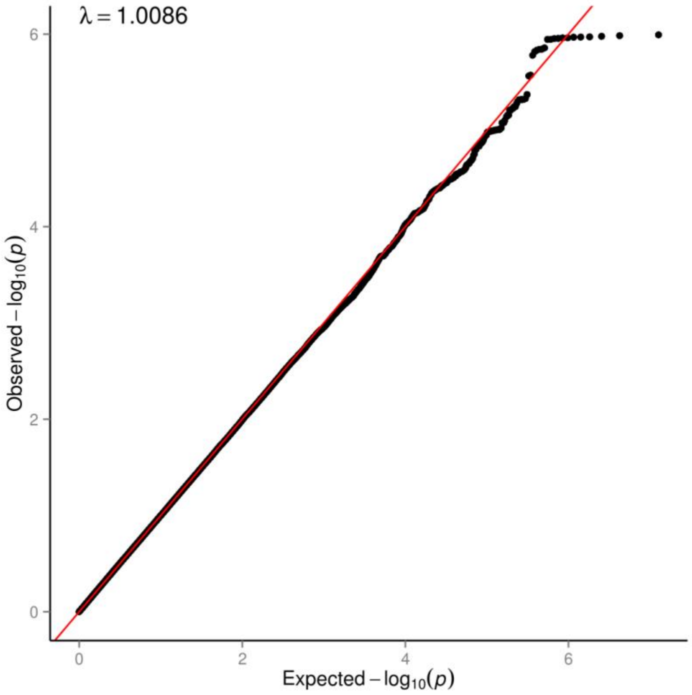
- If λ > 1, then population stratification is assumed to exist and a correction is applied by dividing the observed association test χ^2 statistic values by λ.

- In a GWA study, λ can be determined using all of the genotyped markers; the effect on the inflation factor of potential causal SNPs is assumed to be negligible.

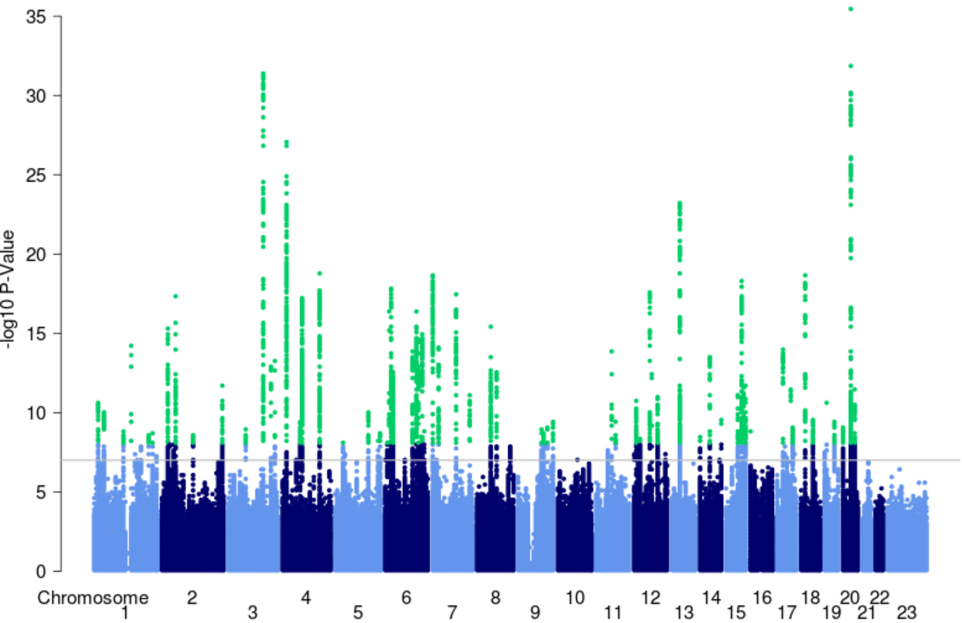
Before



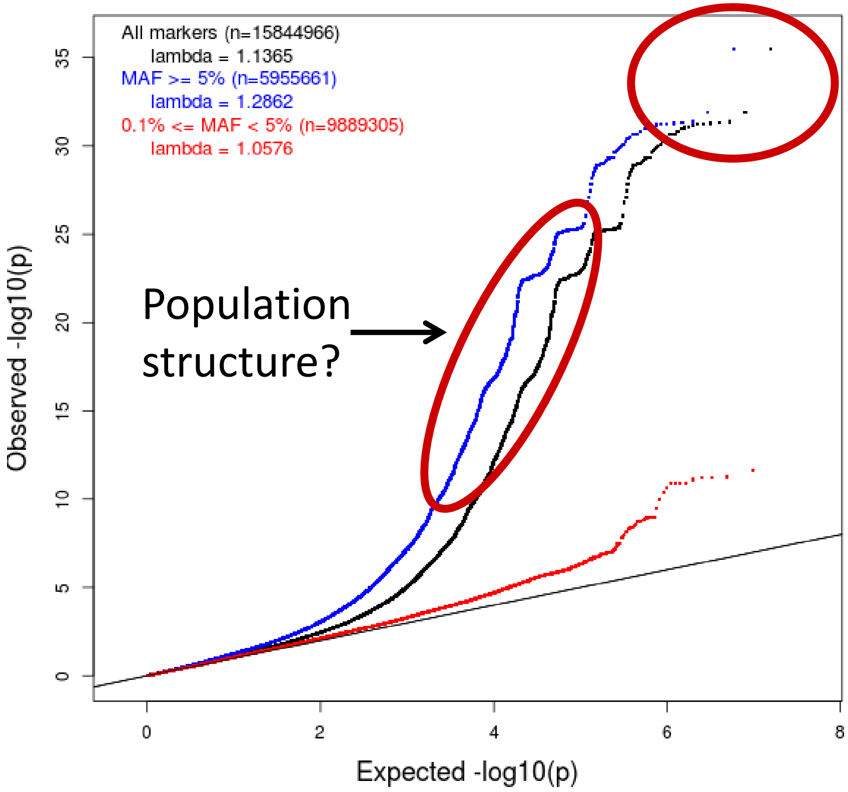
After



Height

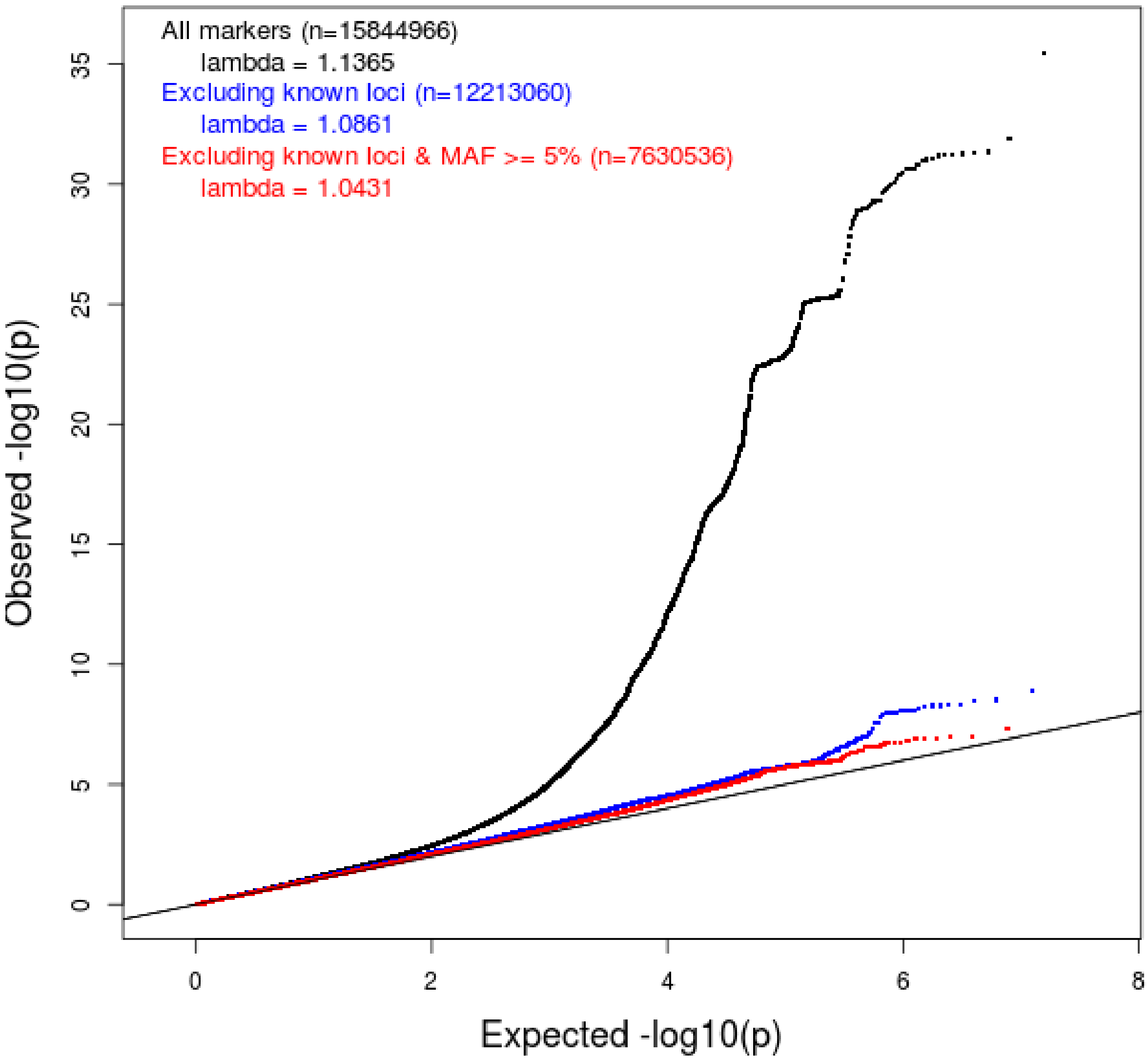


True hits?



Population structure?

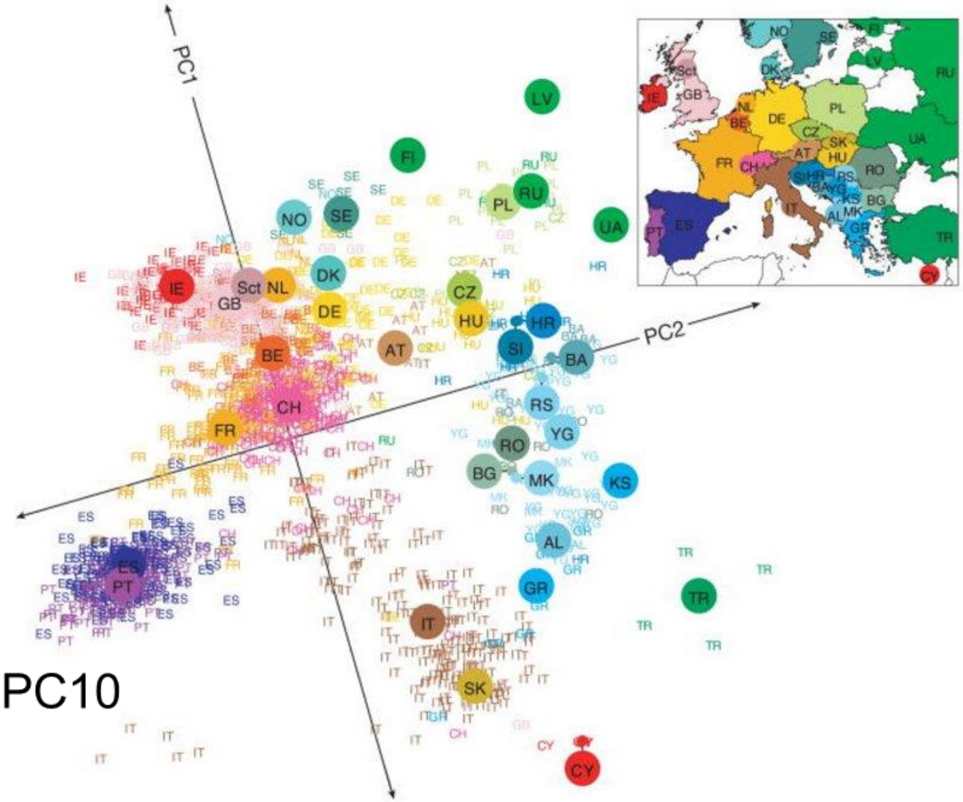
Height


All markers (n=15844966)
lambda = 1.1365
Excluding known loci (n=12213060)
lambda = 1.0861
Excluding known loci & MAF >= 5% (n=7630536)
lambda = 1.0431
Observed -log10(p)
0
5
10
15
20
25
30
35
Expected -log10(p)
0
2
4
6
8

Principal components analysis (PCA)

- PCA is the most widely used approach for identifying and adjusting for ancestry difference among sample individuals.
- It is used to calculate principal components (PCs) that reduce the data to a small number of dimensions, whilst describing as much of the variability between individuals as possible.
- The top PCs are viewed as continuous axes of variation that reflect genetic variation due to ancestry in the sample.
- An application of PCA to genetic data showed that among Europeans the first two PCs computed using 200,000 SNPs could map their country of origin quite accurately in the plane.
- To adjust for any residual population structure during association testing, PCs are included as covariates in a regression.

$$y = \beta_0 + x\beta_1 + \beta_2\text{PC1} + \beta_3\text{PC2} + \ldots + \beta_{11}\text{PC10}$$



Novembre et al Nature 2008

Single test

- When performing a test, we compute how likely it is that the data we observed would be generated under the null hypothesis that there is no effect.

- If this value is below some predetermined significance threshold, we reject the null hypothesis.

- In testing a single hypothesis, a p-value < 0.05 is accepted as sufficient evidence to reject the null hypothesis and make a claim of association:
  - Single test on 5% significance level
  - P(Making an error) = α = 5%
  - P(Not making an error) = $1 - \alpha$ = 95%

Multiple testing

- Suppose you test 500,000 SNPs for association with disease. Using the standard p-value cut-off of 0.05, you expect 500,000 x 0.05 = 25,000 to be deemed “significant” by chance alone.

- The significance threshold needs to be adjusted for the number of independent tests performed to minimize the chance of reporting a false positive (type I error):
  - FWER = Probability of getting at least one false discovery
  - Control the FWER at 5%

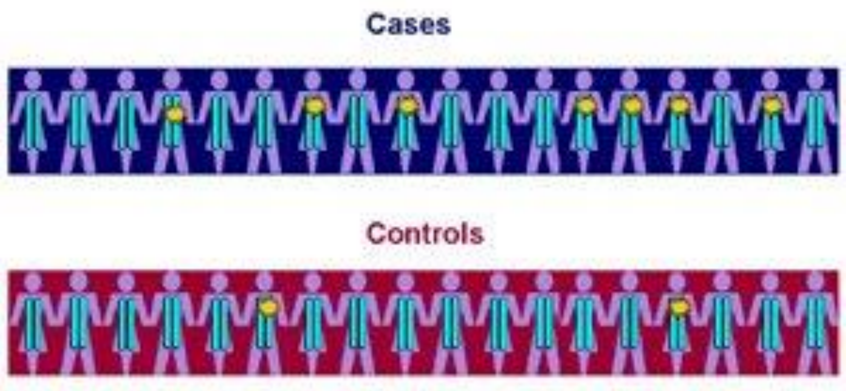
- A widely accepted method to account for multiple comparisons is the Bonferroni correction, in which the p-value threshold is determined by dividing the desired type I error by the total number of independent tests performed.

- In GWAs, the number of independent common variation in European populations is estimated to be 1 million, and thus p = $5x10^{-8}$ (0.05/1,000,000) has become the widely accepted “genome-wide significance” threshold.

- Application of genome-wide significance has helped ensure that the majority of SNP-disease associations discovered through GWAS have been robust and reproducible.

Appendix

Association between a genetic variant and disease



DD: variant homozygote

Dd: heterozygote

dd: common homozygote

Marker genotype	Affected	Unaffected	Total
DD	n_{2A}	n_{2U}	$n_{2.}$
Dd	n_{1A}	n_{1U}	$n_{1.}$
dd	n_{0A}	n_{0U}	$n_{0.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

The odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. Therefore the odds ratio for genotype DD relative to dd is:

$$\mathrm{OR}(DD:dd) = \frac{\text{odds of an individual with genotype DD carrying the disease}}{\text{odds of an individual with genotype dd carrying the disease}}$$

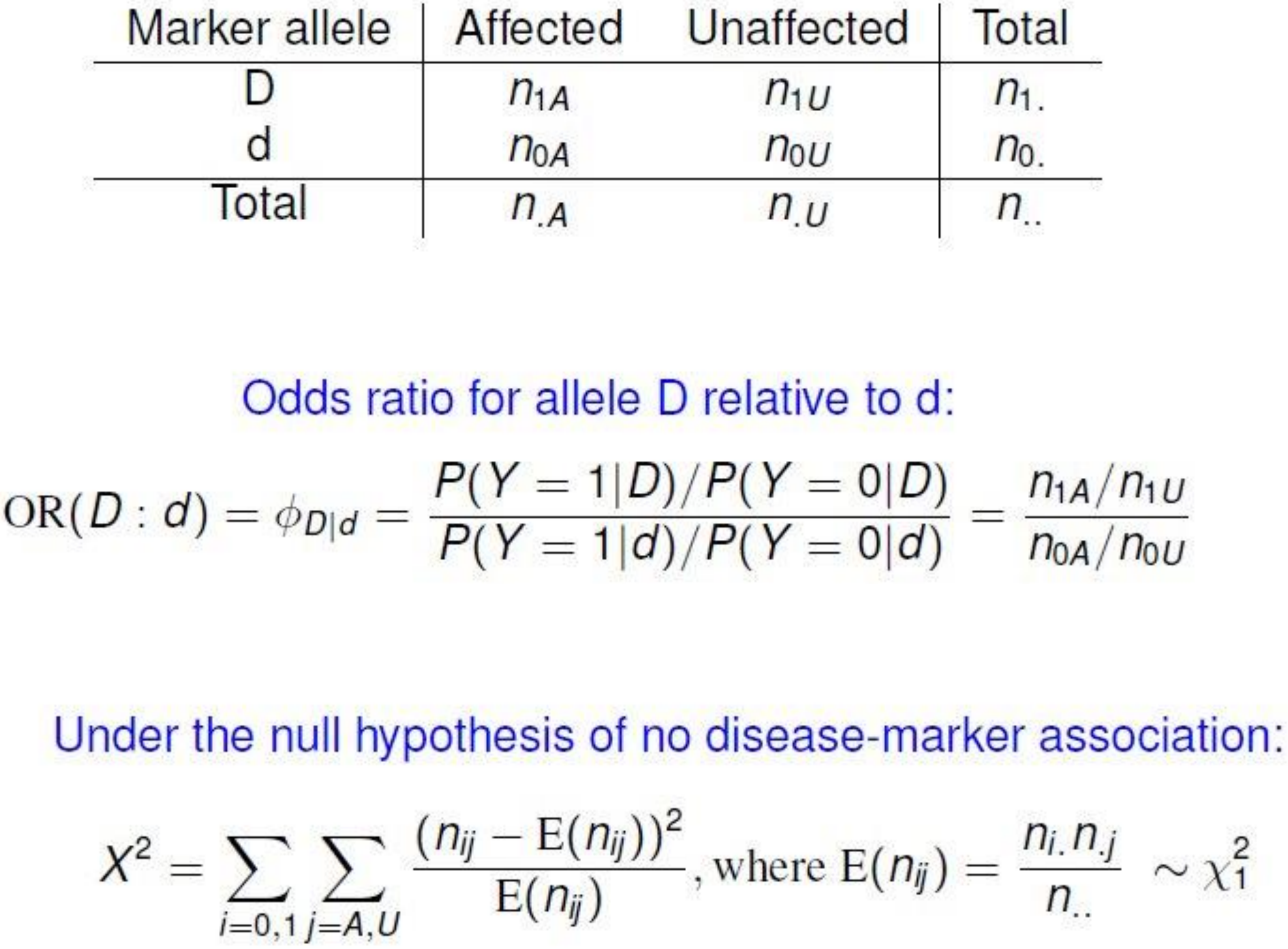
or else:

$$\mathrm{OR}(DD:dd) = \frac{P(Y=1|DD)/P(Y=0|DD)}{P(Y=1|dd)/P(Y=0|dd)}$$

or else:

$$\mathrm{OR}(DD:dd) = \phi_{DD|dd} = \frac{n_{2A}/(n_{2A}+n_{2U})}{n_{2U}/(n_{2A}+n_{2U})} \Big/ \frac{n_{0A}/(n_{0A}+n_{0U})}{n_{0U}/(n_{0A}+n_{0U})} = \frac{n_{2A}/n_{2U}}{n_{0A}/n_{0U}}$$

2 x 2 contingency tables



Marker allele	Affected	Unaffected	Total
D	n_{1A}	n_{1U}	$n_{1.}$
d	n_{0A}	n_{0U}	$n_{0.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

Odds ratio for allele D relative to d:

$$\mathrm{OR}(D:d) = \phi_{D|d} = \frac{P(Y=1|D)/P(Y=0|D)}{P(Y=1|d)/P(Y=0|d)} = \frac{n_{1A}/n_{1U}}{n_{0A}/n_{0U}}$$

Under the null hypothesis of no disease-marker association:

$$X^2 = \sum_{i=0,1}\sum_{j=A,U} \frac{(n_{ij} - \mathrm{E}(n_{ij}))^2}{\mathrm{E}(n_{ij})}, \text{where } \mathrm{E}(n_{ij}) = \frac{n_{i.}n_{.j}}{n_{..}} \ \sim \chi^2_1$$

Fisher's exact test

- The chi-square test of independence is an approximation that holds for large sample sizes and/or common SNPs.
- The Fisher's exact test does not rely on the large sample approximation and it is typically used when a cell of the contingency table is less than or equal to 5.
- Fisher's exact test looks at a table and determines how many tables are as extreme or more extreme than the observed table under the null hypothesis of no association.
- Fisher showed that, under the null hypothesis, the probability of obtaining the frequencies observed in a 2x2 contingency table given the row and column sums is given by the hypergeometric distribution:

$$p = \frac{\binom{n_{.A}}{n_{1A}}\binom{n_{.U}}{n_{1U}}}{\binom{n_{..}}{n_{1.}}} = \frac{n_{.A}!n_{.U}!n_{1.}!n_{0.}!}{n_{..}!n_{1A}!n_{0A}!n_{1U}!n_{0U}!},$$

where $n! = 1 \times 2 \times 3 \times \cdots \times (n-1) \times n.$

Least Square method for parameter estimation

In order to solve:

$$\arg\min \sum_{i=1}^{N} (y_i - x_i\beta_1 - \beta_0)^2$$

We differentiate with respect to each parameter, set it to 0, and solve it for each parameter:

$$F(\beta_1, \beta_0) = \sum_{i=1}^{N} (y_i - x_i\beta_1 - \beta_0)^2$$

$$\frac{\partial F(\beta_1, \beta_0)}{\partial \beta_1} = 0 \,, \quad \frac{\partial F(\beta_1, \beta_0)}{\partial \beta_0} = 0$$

To obtain:

$$\beta_1 = \frac{\sum_{i=1}^{N} (y_i - \overline{y})(x_{i,} - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})(x_i - \overline{x})} \qquad \beta_0 = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \beta_1$$

Testing the significance of association using linear models

- We want to test whether to reject the null hypothesis that there is no linear relationship between the given SNP and phenotype:
  - Null hypothesis H_o: $\beta_1=0$
  - Alternative hypothesis H_1: $\beta_1 \neq 0$

- We compute the test statistic:

$$T = \frac{\hat{\beta}_1}{se.(\hat{\beta}_1)} \qquad se.(\hat{\beta}_1) = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{N}(y_i - x_i\beta_1 - \beta_0)^2}}{\sqrt{\sum_{i=1}^{N} x_i x_i}}$$

- The test statistic T asymptotically follows the standard normal distribution N(0,1)
- Use this to:
  - Construct a 95% confidence interval for β_1 : $\hat{\beta}_1 \pm 1.96 \times se(\hat{\beta}_1)$
  - Assess the significance of association by calculating a p-value = $2 \times \left(1 - \Phi(|T|)\right)$, where Φ is the cumulative distribution function of N(0,1).

Association between a genetic variant and a binary trait

- Logistic regression is an extension of the linear regression model that allows for case/control data to be modelled.

- Suppose N independent random variables Y_1, Y_2, …, Y_N with value 0/1 corresponding to controls/cases.

- We model the probability of individual i being a case given he carries $x_i = 0, 1, 2$ copies of the minor allele at genotype x_i being a case π_i as:

$$\text{logit}\,(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i \;\Rightarrow\; \pi_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

- Inference can be carried out using similar techniques to those used for linear regression models.

Logistic regression model

- If individual i carries 0 copies of the minor allele, $logit(\pi_i) = \alpha$.
- If individual i carries 1 copy of the minor allele, $logit(\pi'_i) = \alpha + \beta$.
- If individual i carries 2 copies of the minor allele,$logit(\pi''_i) = \alpha + 2\beta$.

- Subtracting these equations:

$$\beta = \text{logit}\ (\pi_i') - \text{logit}\ (\pi_i) = \ln\{\pi_i'/(1-\pi_i')\} - \ln\{\pi_i/(1-\pi_i)\} = \ln\left\{\frac{\pi_i'/(1-\pi_i')}{\pi_i/(1-\pi_i)}\right\}$$

- Therefore β is the logarithm of the odds ratio of an individual being a case when he/she carries 1 copy of the minor allele compared to carrying none.

- The odds of being a case when carrying 1 copy of the minor allele is $\exp(\beta)$ times the odds of being a case when carrying 0 copies of the minor allele.
- Similarly, the odds of being a case when carrying 2 copies of the minor allele is $\exp(2\beta)$ times the odds of being a case when carrying 0 copies of the minor allele.

- The disease model from this coding is called additive on the log scale, or multiplicative on the raw scale. Different codings imply a recessive or dominant disease model.