JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Proteinbiochemie und Bioinformatik

Introduction to statistical genetics

David Fournier
dfournie@uni-mainz.de

21.03.2017

Outline

I Definitions

- SNP
- QTL
- Types of associations

II Allelic association

1. Marker association
2. Case/control allelic test
3. Genetic association

III Genome-wide association studies

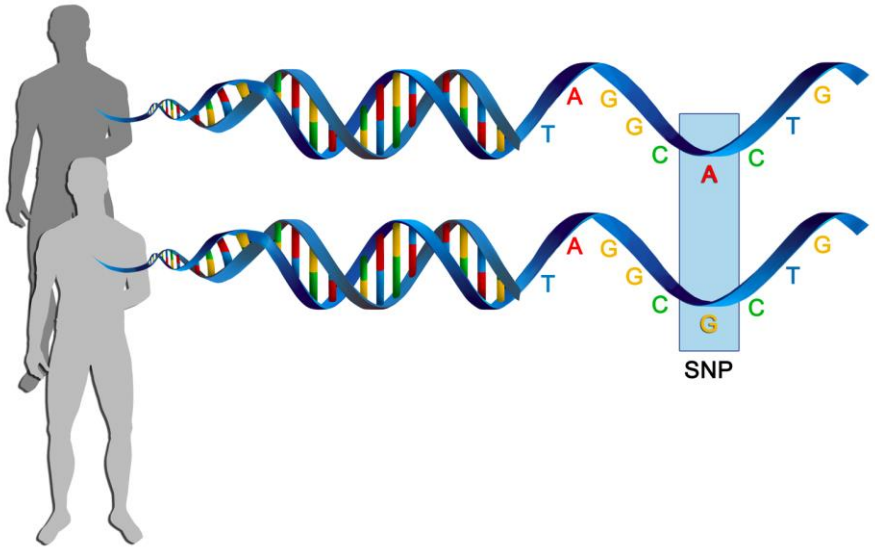
1. With markers
2. Manhattan plots

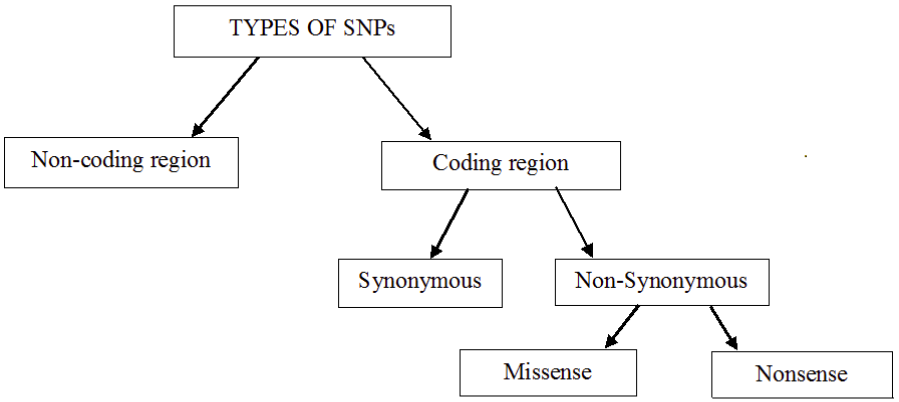
Single Nucleotide Polymorphism

Definition

Variation in a single nucleotide.

- 1,419,190 known in the human genome
- Major allele is the wild-type, minor allele the mutant
- Quite rare in most cases (< 0.1%)





Quantitative trait locus

Definition.
A quantitative trait loci (QTL) mapping is a locus whose mutation influences the variation of a quantitative trait.

- Can be a chromosome region, or a punctual position, such as a Single Nucleotide Polymorphism (SNP).

- Involves common alleles, i.e. with minor allele frequencies ≥ 1% – 5%
- The effect of most SNP for phenotype is weak or null.
- Some quantitative traits can be largely influenced by a single gene as well as by environmental factors

Types of association

Marker association:
An entire region of a chromosome is associated to the phenotype (several SNP).

Allelic association:
An allele of a gene (for instance a or A for gene A) is associated to a phenotype (one SNP).

Genotypic/genetic association:
Both alleles on the two chromosomes are associated to the phenotype, for instance genotypes AA, Aa or aa at gene A (one SNP).

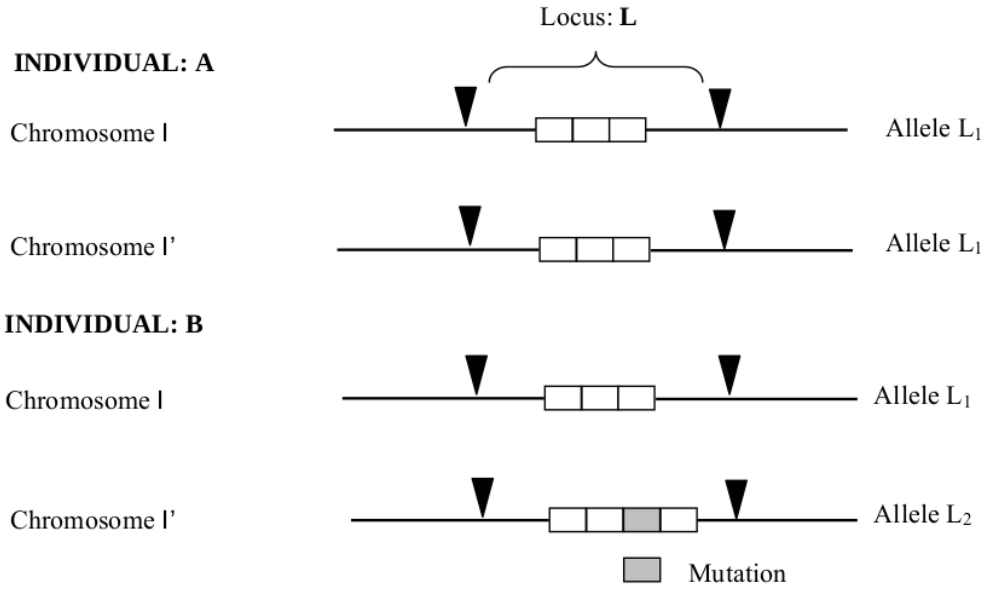
Genome-wide association studies (GWAS):
Genotypic association studies performed for all available SNP on the genome.

Marker association

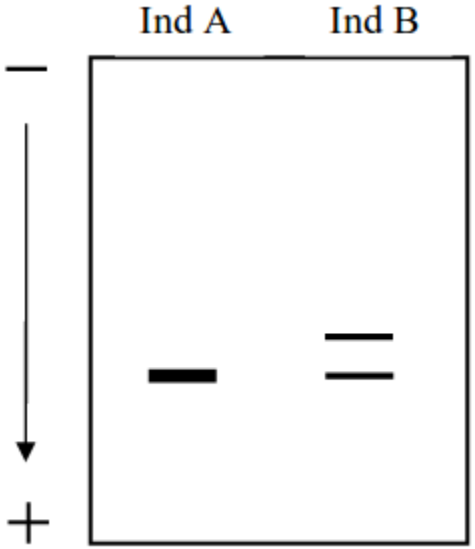
Genetic markers

Genetic markers

Enzyme digestion



Gel electrophoresis

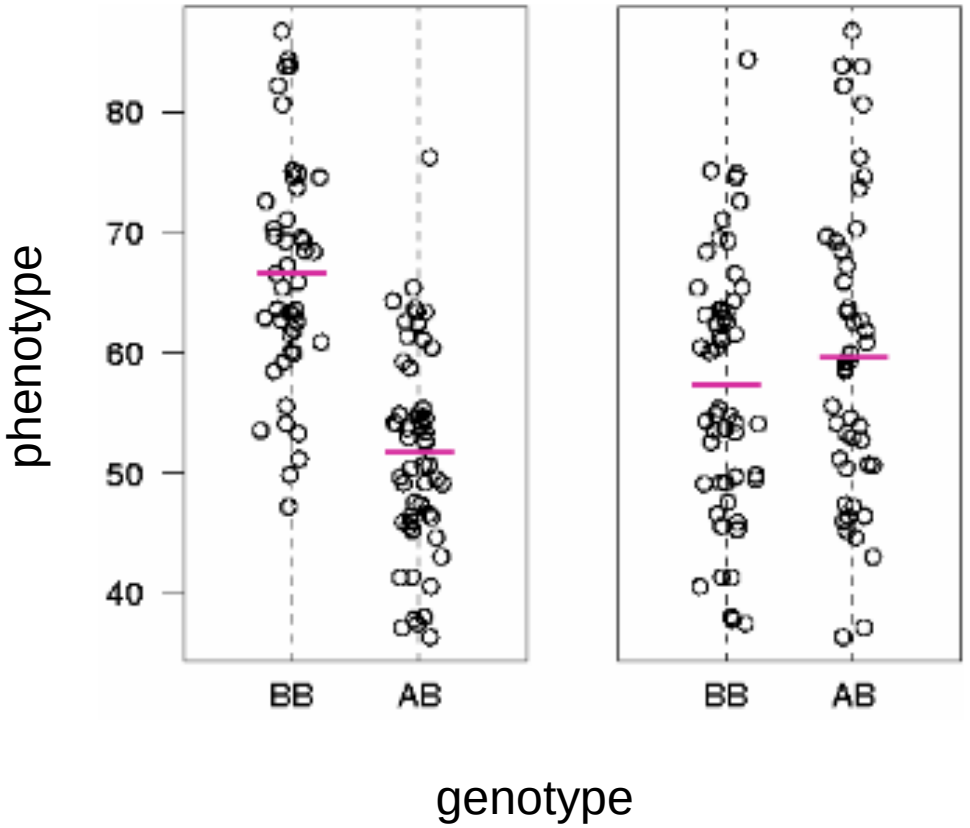


A is homozygote for L
B is heterozygote for L

Indirect measure of a mutation
(SNP or more complex)

Single QTL analysis

- T-test for two mice
- Analysis of variance (ANOVA) by comparing difference of means for more mice

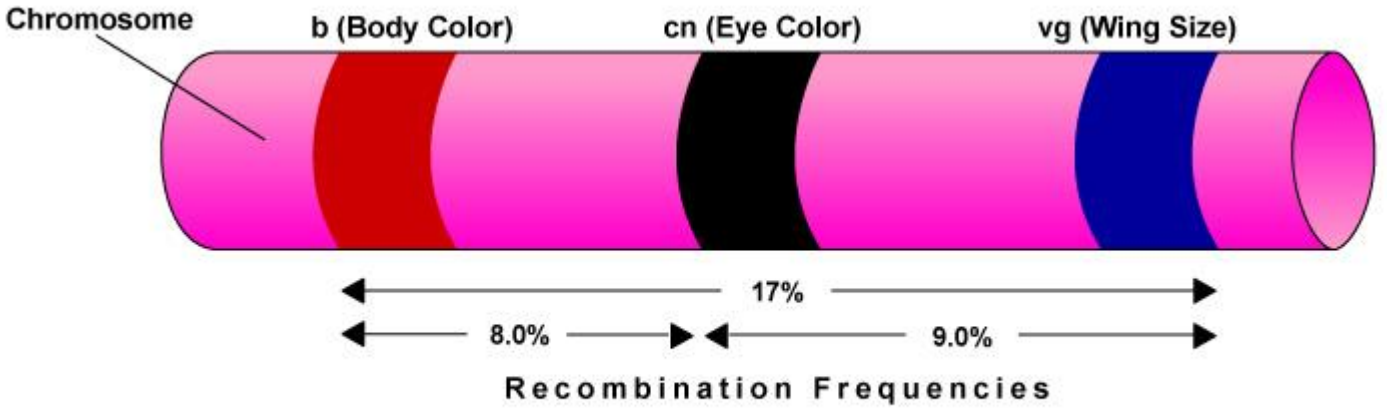


Interval mapping

Definition:
Method to associate two alleles on a chromosome using a maximum likelihood ratio (LOD score).

- Helps to decide which QTL are the best (not possible with simple mapping - at least in the context of genetic markers), contrary to t-test or ANOVA.

- Uses a genetic map:



LOD score

Definition:
Logarithm of odds or LOD score is the log in base 10 of the likelihood ratio "QTL versus no QTL models".
A high LOD score for a pair of alleles (or interval) means that they recombine more often that random (more 50% of times). At a given interval, if LOD score increases with a given condition, then the two alleles have a good chance to be associated with the disease.

Framework:

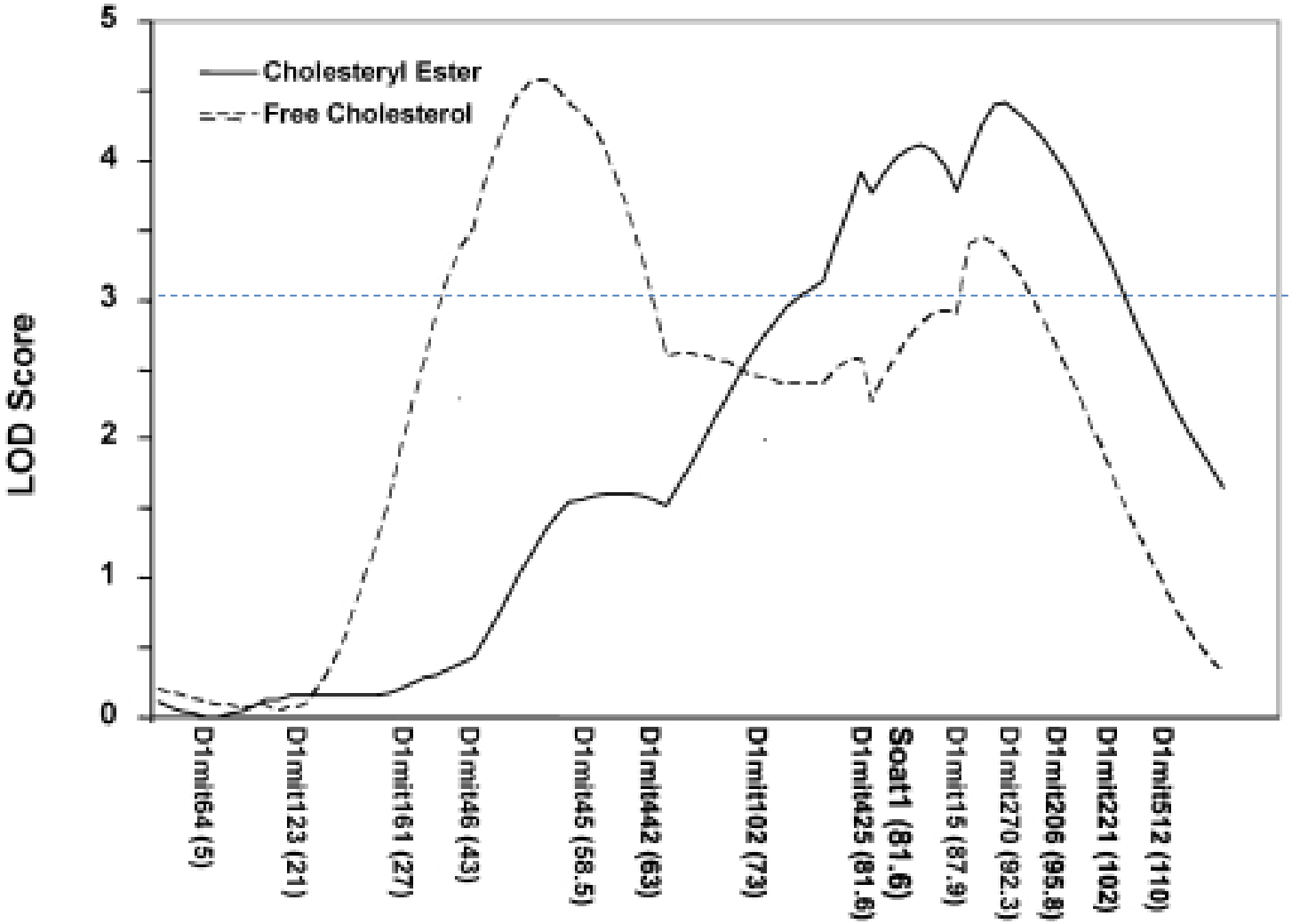
Need a record of the individuals pedigree, i.e. ancestors.
Establish a estimate of the recombination rate of each locus you are studying.

LOD score:

$$LOD = \log_{10} \frac{P(x|\theta_1)}{P(x|\theta_0)} = \log_{10} \frac{(1-\theta)^{NR} \times \theta^{R}}{0.5^{(NR+R)}}$$

With $P(x|\theta_1)$ likelihood of having the two alleles knowing a recombination rate θ in the Previous generation;
$P(x|\theta_0)$ the null model where the recombination rate is random, i.e. 0.5 chance to happen;
NR is the number of non-recombinant individual in this generation;
R is the number of recombinant individual in this generation.

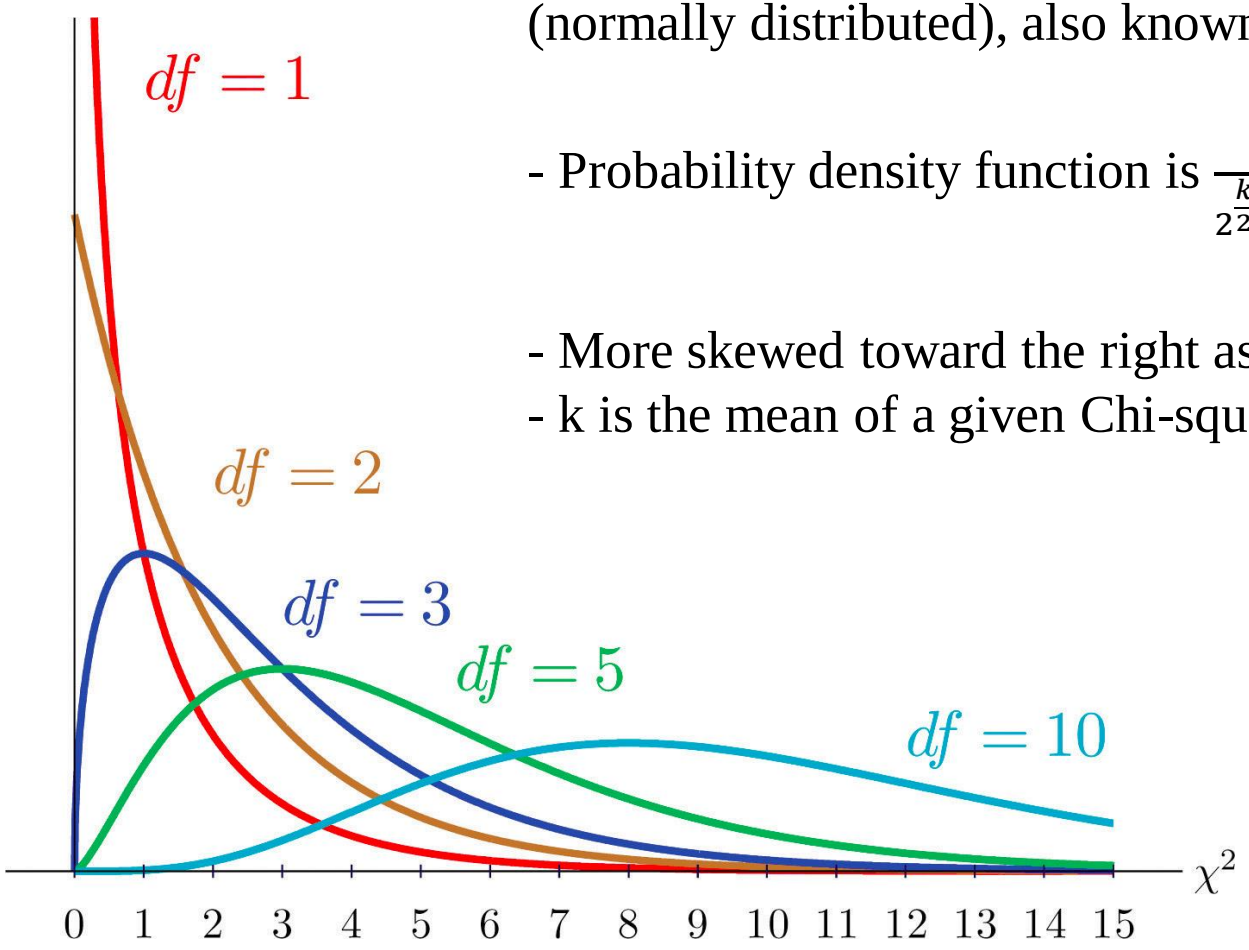
Interval mapping



Significance threshold

Case/control study allelic test

Chi-square (χ^2) distribution



- Distribution of the average of square of k random variables (normally distributed), also known as degrees of freedom (df)

- Probability density function is $\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$

- More skewed toward the right as k raises.
- k is the mean of a given Chi-square distribution.

Pearson's χ^2 test

Definition

Tests whether **the observed data** from a **contingency table** that shows the frequency distribution of different variables is different from a **theoretical uniform distribution of data** across the same table (theoretical or expected values).

Example.

A screening test for a given infectious disease is tested. Some people are Sick, some are healthy. The result is the following: Tested positive and sick (true Positive or TP), Tested negative and healthy (true negative or TN), positive and Healthy (false negative or FN), negative and sick (false positive of FP).

The resulting contingency table is the following:

	Disease	Control	Total
Test+	n_{TP}	n_{FP}	n_P
Test-	n_{FN}	n_{TN}	n_N
Total	n_{sick}	$n_{control}$	n

n_P: $n_{TP} + n_{FP}$
(nb of positive tests)

n_N: $n_{TN} + n_{FN}$
(nb of negative tests)

n_{sick}: $n_{TP} + n_{FN}$
(nb of sick ppl)

$n_{control}$: $n_{FP} + n_{TN}$
(nb of healthy ppl)

$N = n_P + n_P + n_c + n_s$
(total nb of outcomes)

Pearson's χ^2 test

Expected frequency table:
This is a theoretical table assuming that the distribution of items is the same For different columns on one hand and the distribution if the same for Different rows.

In the previous example, the expected table should have the following Characteristics:
$n_{TP}/n_{FN} = n_{FP}/n_{TN} = n_P/n_N$ and $n_{TP}/n_{FP} = n_{FN}/n_{TN} = n_{sick}/n_{control}$

Formula to fill cell (i,j) in the table:
$E_{i,j} = (n_i + n_j)/n$ with n_i total of the row and n_j total of the column in the original data.

	Disease	Control	Total
Test+	$(n_P * n_s)/n$	$(n_P * n_s)/n$	n_P
Test-	$(n_N * n_s)/n$	$(n_N * n_c)/n$	n_N
Total	n_{sick}	$n_{control}$	n

Pearson's χ^2 test

Value of the statistics for a Pearson χ^2 test :

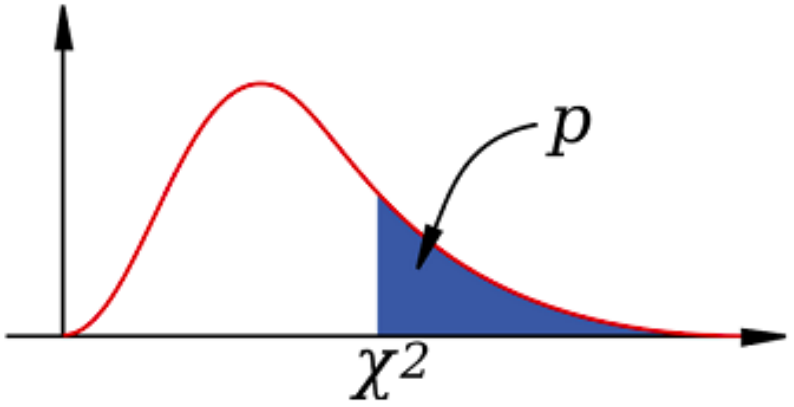
$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

With n the number of cells (4 for a 2x2 table),
O_i the observed data;
E_i the expected value in cell i.

Pearson's χ^2 test

Interpretation of the statistics:

- Computation of the degree of freedom of your model: $(n_{row} - 1)(n_{col} - 1)$
- For a given number of degree of freedom (for instance 2 for a 3x3 table), H_0 is the hypothesis that the data to be the same as the expected values and the p-value is the probability or likelihood $P(x|\theta)$ to have the observed values (your data *x*) given model "the observed data are not different from expected values".
- The value of the statistics can be reported on the χ^2 distribution associated to The correct number of degrees of freedom. The area under the curve which Is after

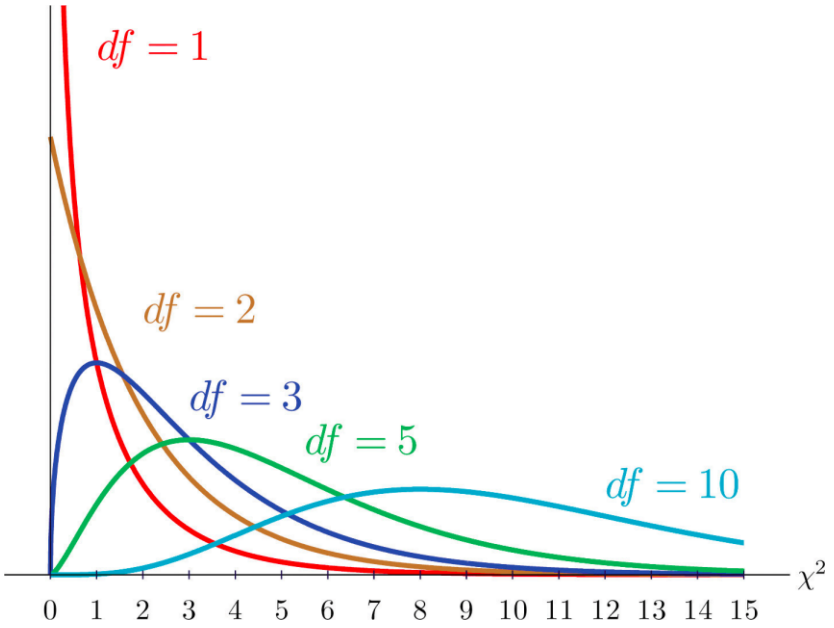


Pearson's χ^2 table

Percentage Points of the Chi-Square Distribution

Accept H0 Reject H0

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09



Pearson's χ^2 test

Example/exercise:

Find below a contingency table representing statistics about a screening test.
Fill the expected/theoretical table so the distribution in the table is homogenous.
Calculate the value of the X^2 test.
How many degrees of freedom is considered? Are observed data significantly different from expected ?

	Disease	Control	Total
Test+	11	111	122
Test-	1	1111	1112
Total	12	1222	1234

	Disease	Control	Total
Test+	?	?	122
Test-	?	?	1112
Total	12	1222	1234

Pearson's χ^2 test

Example/exercise:

Find below a contingency table representing statistics about a screening test.
Fill the expected/theoretical table so the distribution in the table is homogenous.
Calculate the value of the X^2 test.
How many degrees of freedom is considered? Are observed data significantly different from expected ?

	Disease	Control	Total
Test+	11	111	122
Test-	1	1111	1112
Total	12	1222	1234

	Disease	Control	Total
Test+	**1,19**	**120,81**	122
Test-	**10,81**	**1101,19**	1112
Total	12	1222	1234

12x(122/1234)

$$X^2 = \frac{(11-1{,}19)^2}{1{,}19} + \frac{(111-120{,}81)^2}{120{,}81} + \frac{(1-10{,}81)^2}{10{,}81} + \frac{(1111-1101{,}19)^2}{1101{,}19}$$

$X^2 = 90{,}66$; df = 1

> 3.84 : H0 rejected

Pearson's χ^2 test

H_0: no relationship between allele and case/control status

	Case	Control	Total
Allele a	n_a^1	n_a^2	n_a
Allele A	n_A^1	n_A^2	n_A
Total	$2N^1$	$2N^2$	T

N1 and N2: number of cases and controls with genotype A or a

$n_a^1 = 2\, n_{aa}^1 + n_{aA}^1$; $n_A^1 = 2\, n_{AA}^1 + n_{aA}^1$;

$n_a^2 = 2\, n_{aa}^2 + n_{aA}^2$; $n_A^2 = 2\, n_{AA}^2 + n_{aA}^2$

Pearson's χ^2 test

Exercise.
Here is the distribution of genotypes across cases and controls:

Case AA	Case Aa	Case aa	Control AA	Control Aa	Control aa
100	20	3	850	200	31

Fill tables for observed and expected values. Are expected and observed values significantly different?

Hints:

N1 and N2: number of cases and controls with genotype A or a

$n_a^1 = 2\, n_{aa}^1 + n_{aA}^1$; $n_A^1 = 2\, n_{AA}^1 + n_{aA}^1$; $n_a^2 = 2\, n_{aa}^2 + n_{aA}^2$; $n_A^2 = 2\, n_{AA}^2 + n_{aA}^2$

Pearson's χ^2 test

Exercise.
Here is the distribution of genotypes across cases and controls:

Case AA	Case Aa	Case aa	Control AA	Control Aa	Control aa
100	20	3	850	200	31

Fill tables for observed and expected values. Are expected and observed values significantly different?

	Case	Control	Total
Allele a	26	262	288
Allele A	220	1900	2120
Total	246	2162	2408

	Case	Control	Total
Allele a	29,42	258,58	288
Allele A	216,58	1903,42	2120
Total	246	2162	2408

$$X^2 = \frac{(26-29{,}42)^2}{29{,}42} + \frac{(262-258{,}58)^2}{258{,}58} + \frac{(220-216{,}58)^2}{216{,}58} + \frac{(1900-1903{,}42)^2}{1903{,}42}$$

$= 0{,}398 + 0{,}0452 + 0.0540 + 6{,}14e^{-4} = 0{,}498$; df= 1

p-value associated: $0.25 < X^2 < 0.5$ H0 not rejected, distribution close to expected

Genetic association testing

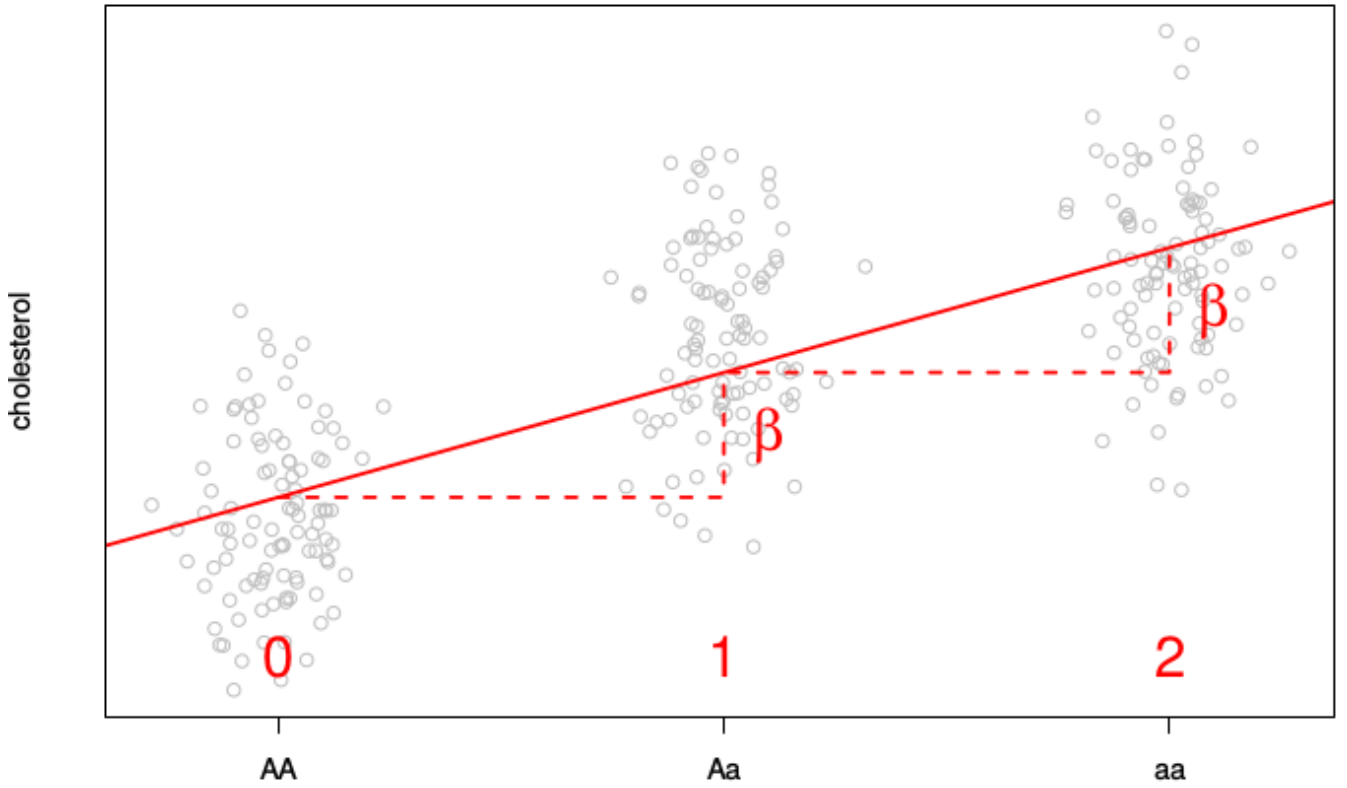
QTL mapping

Definition:
For traits that are heritable, i.e., traits with a non-negligible genetic component that contributes to phenotypic variability, identifying (or mapping) QLT that influence the trait is often of interest.

Linear regression models are commonly used for QTL mapping
Linear regression models will often include a single genetic marker (e.g., a SNP) as predictor in the model, in addition to other relevant covariates (such as age, sex, etc.), with the quantitative phenotype as the response

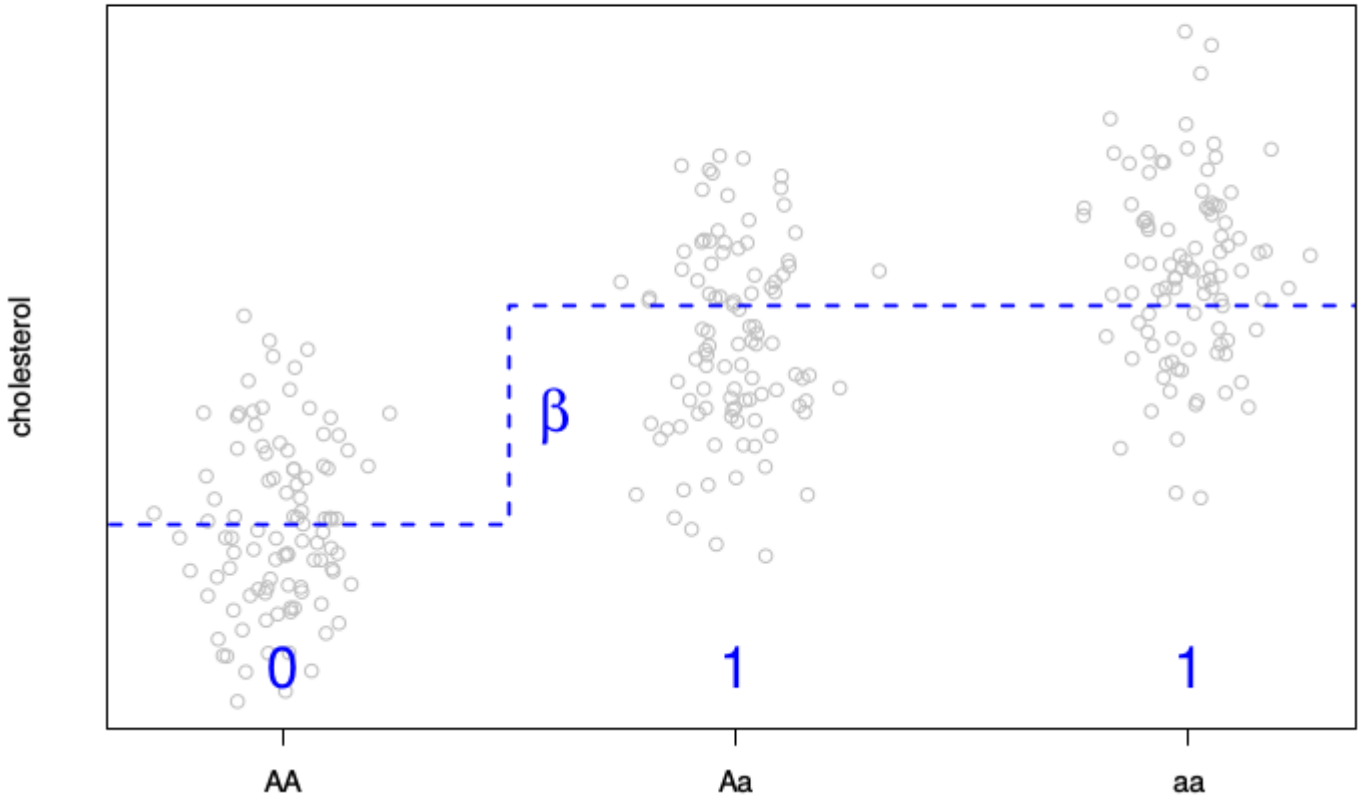
Linear regression models

$$y = \beta_0 + \beta \times \#\text{minor alleles}$$



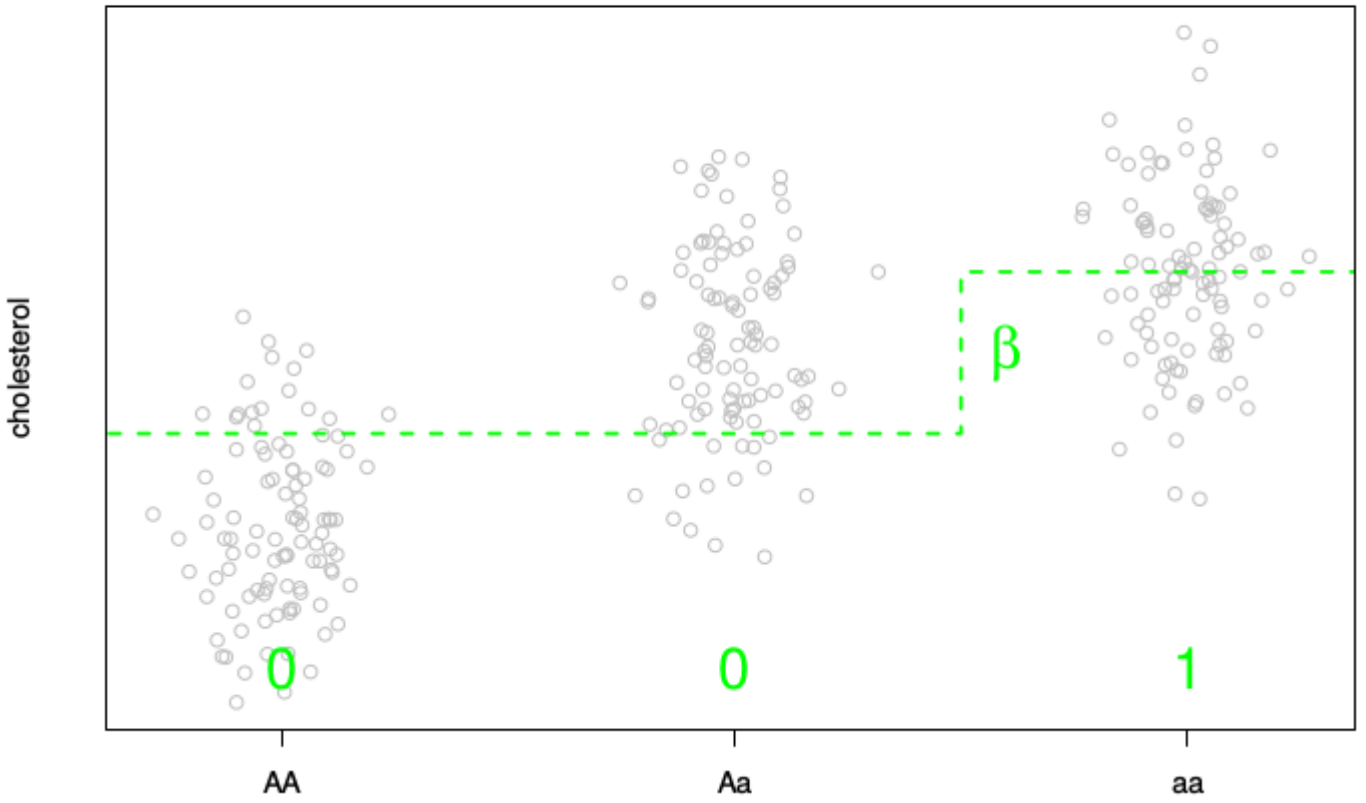
“Dominant” model

$$y = \beta_0 + \beta \times (G \neq AA)$$



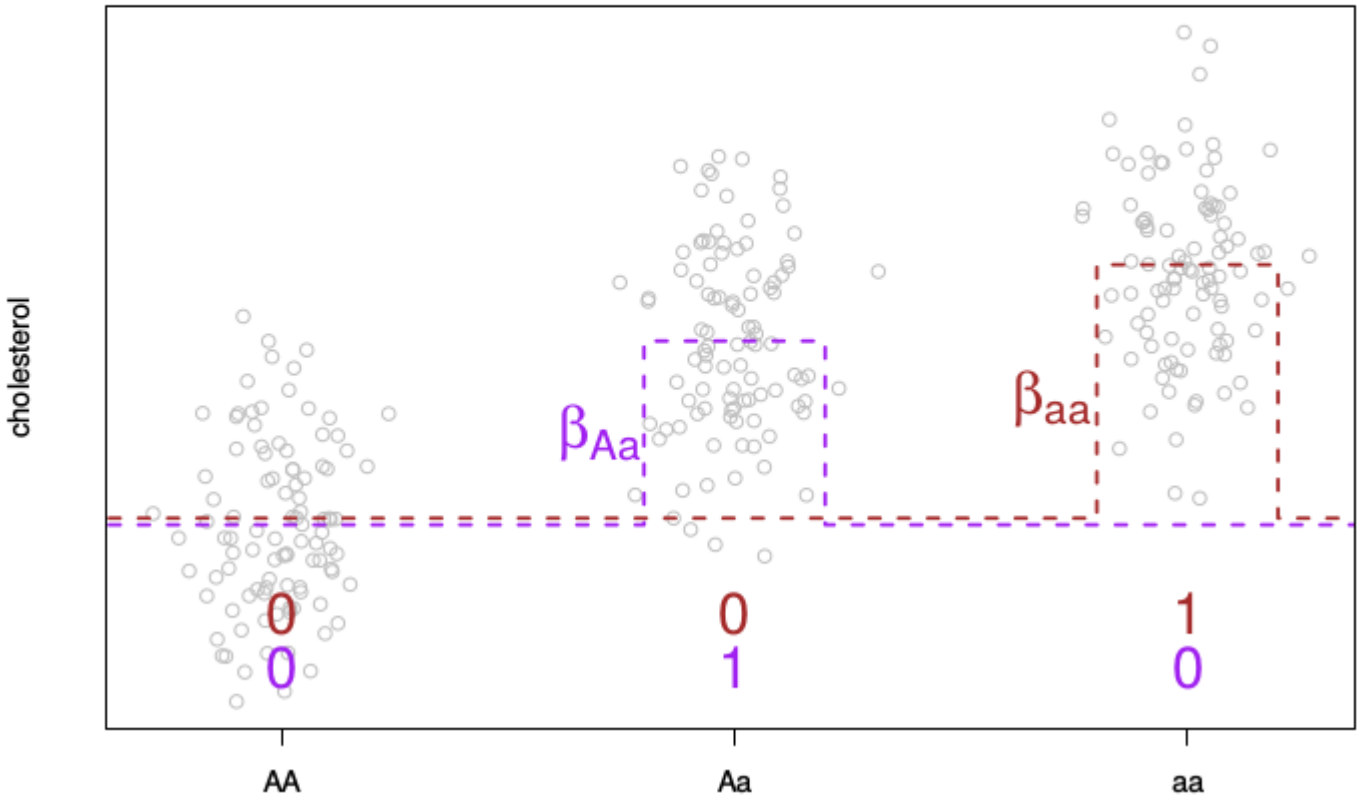
“Recessive” model

$$y = \beta_0 + \beta \times (G == aa)$$



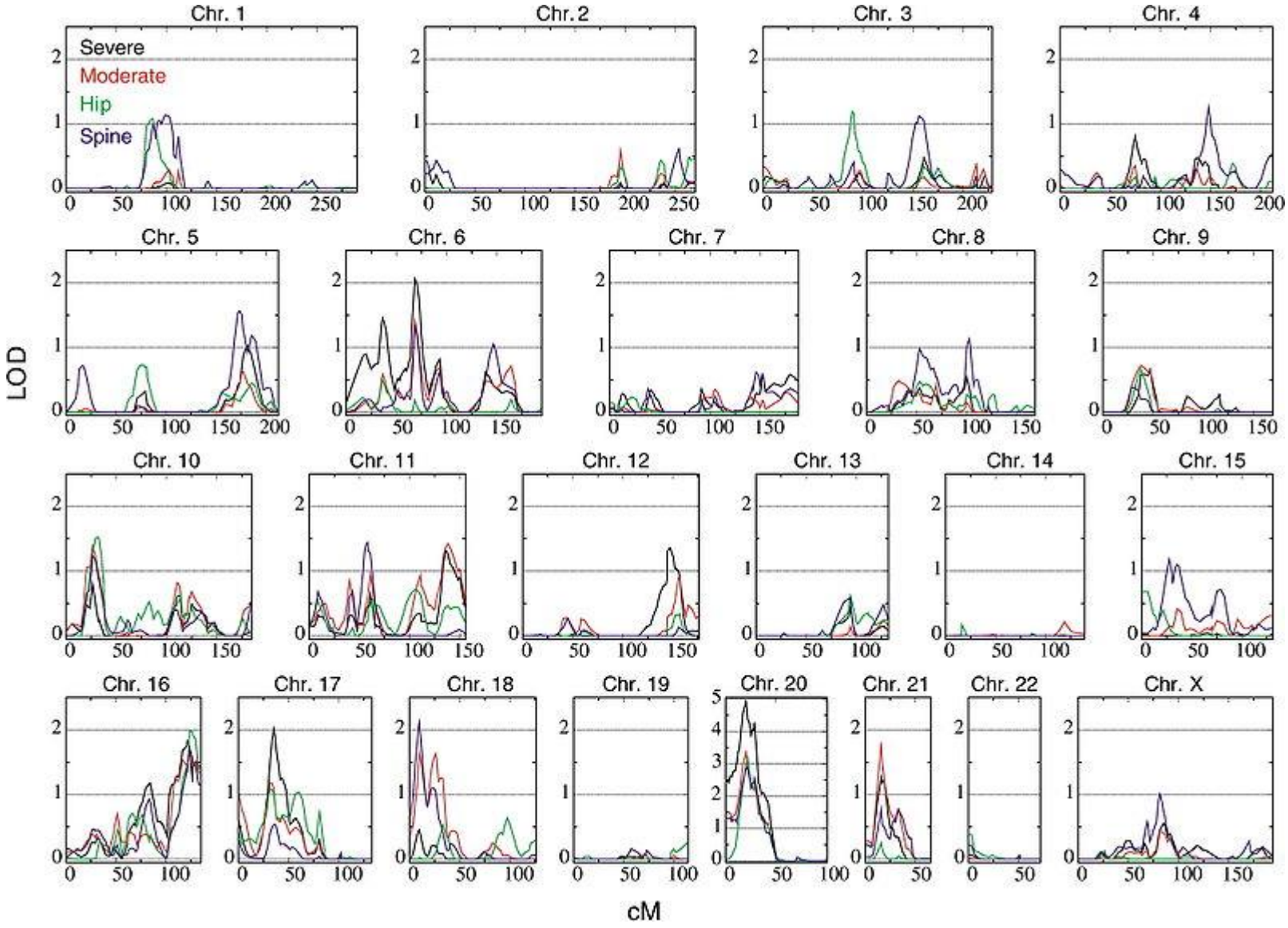
Model with 2 degrees of freedom

$$y = \beta_0 + \beta_{Aa} \times (G == Aa) + \beta_{aa} \times (G == aa)$$



Genome-wide association studies

Genome-wide scan for QTL at markers positions



Manhattan plot of genome-wide SNP associations