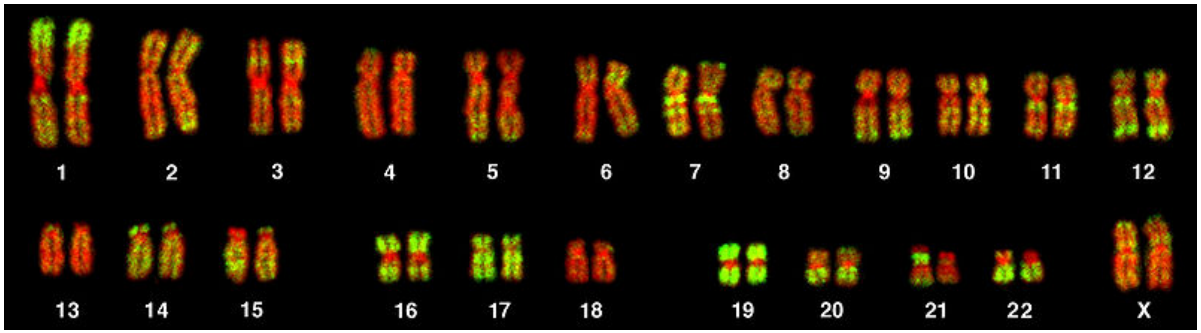
Statistical method for mapping disease genes

Marianne Jonker

April, 2010

Chromosomes

The genetic code of a human is divided over 46 chromosomes, which are present in the nucleus of every cell. These chromosomes form 23 pairs, 22 of which are called autosomes and 1 pair are the sex chromosomes.



An egg or sperm cell carry half the genetic information of an individual.

Genes - Alleles - Genotype

Genes

Genes are located at the chromosomes and hold genetic information, for instance a gene for the color of blossom or a breast cancer gene.

Alleles

The genetic information on a gene is called allele. For instance, the allele "red" (A) or "white" (a) at the gene for the color of blossom. Organisms carry two alleles.

Genotype

A pair of alleles at the same locus is called genotype; for instance aa, aA and AA. Of every pair, one was passed on by the father and one by the mother.

Genes - Alleles - Genotype

Genes

Genes are located at the chromosomes and hold genetic information, for instance a gene for the color of blossom or a breast cancer gene.

Alleles

The genetic information on a gene is called allele. For instance, the allele “red” (A) or “white” (a) at the gene for the color of blossom. Organisms carry two alleles.

Genotype

A pair of alleles at the same locus is called genotype; for instance aa, aA and AA. Of every pair, one was passed on by the father and one by the mother.

Genes - Alleles - Genotype

Genes

Genes are located at the chromosomes and hold genetic information, for instance a gene for the color of blossom or a breast cancer gene.

Alleles

The genetic information on a gene is called allele. For instance, the allele “red” (A) or “white” (a) at the gene for the color of blossom. Organisms carry two alleles.

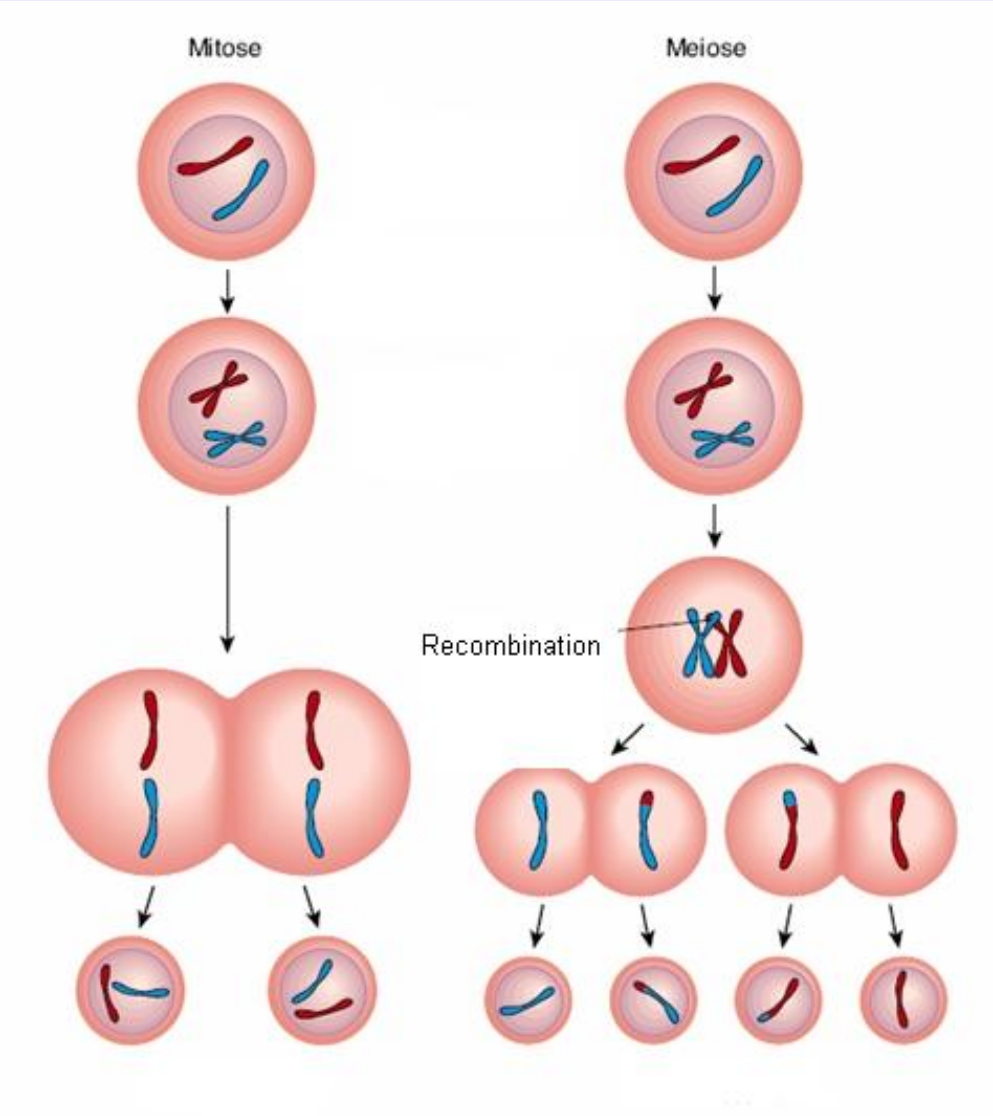
Genotype

A pair of alleles at the same locus is called genotype; for instance aa, aA and AA. Of every pair, one was passed on by the father and one by the mother.

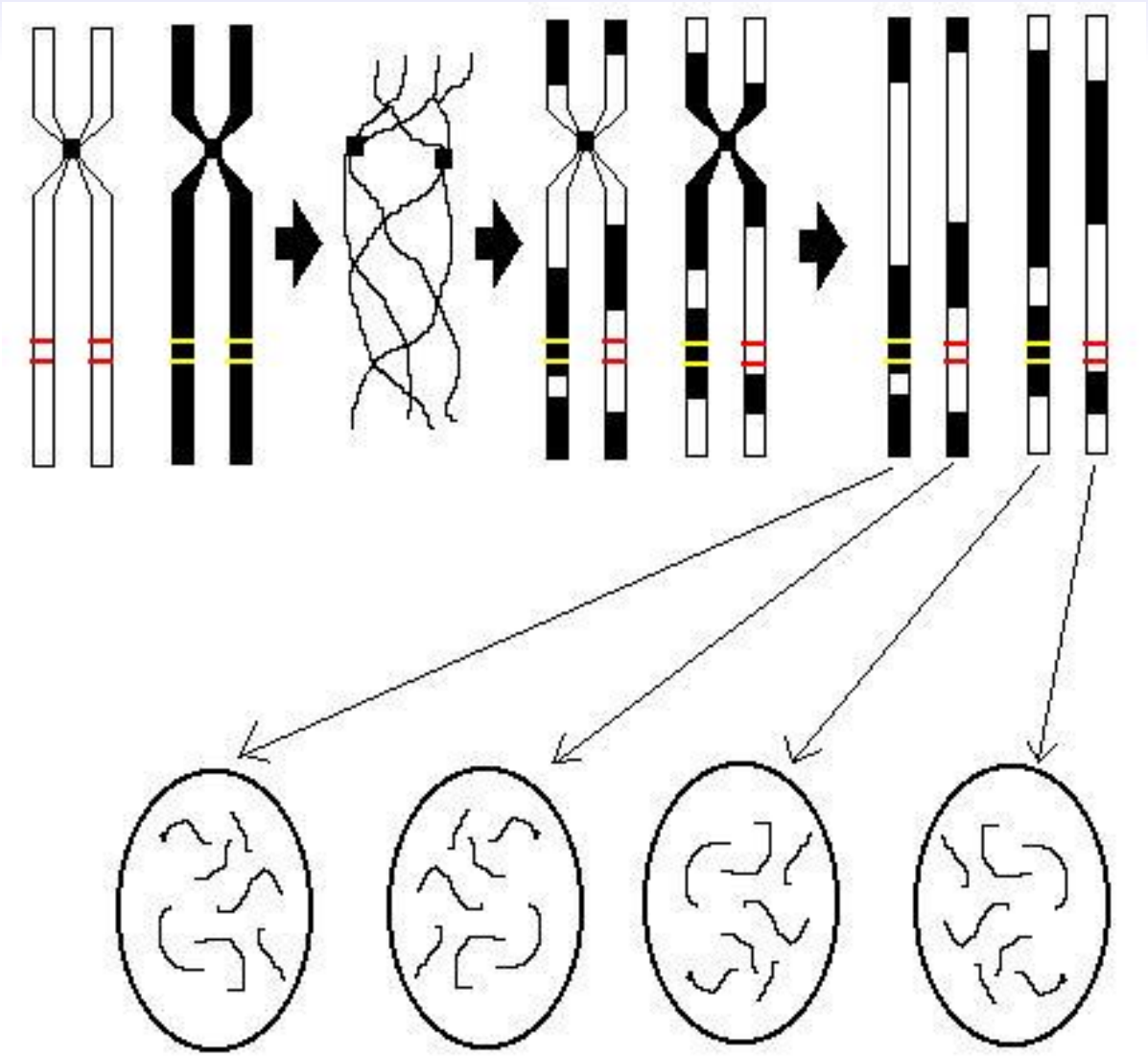
Phenotype

A phenotype is any observable characteristic or trait; like hair color, body height, or the color of blossom. Phenotypes result from the expression of an organism's genes as well as the influence of environmental factors and possible interactions between the two.

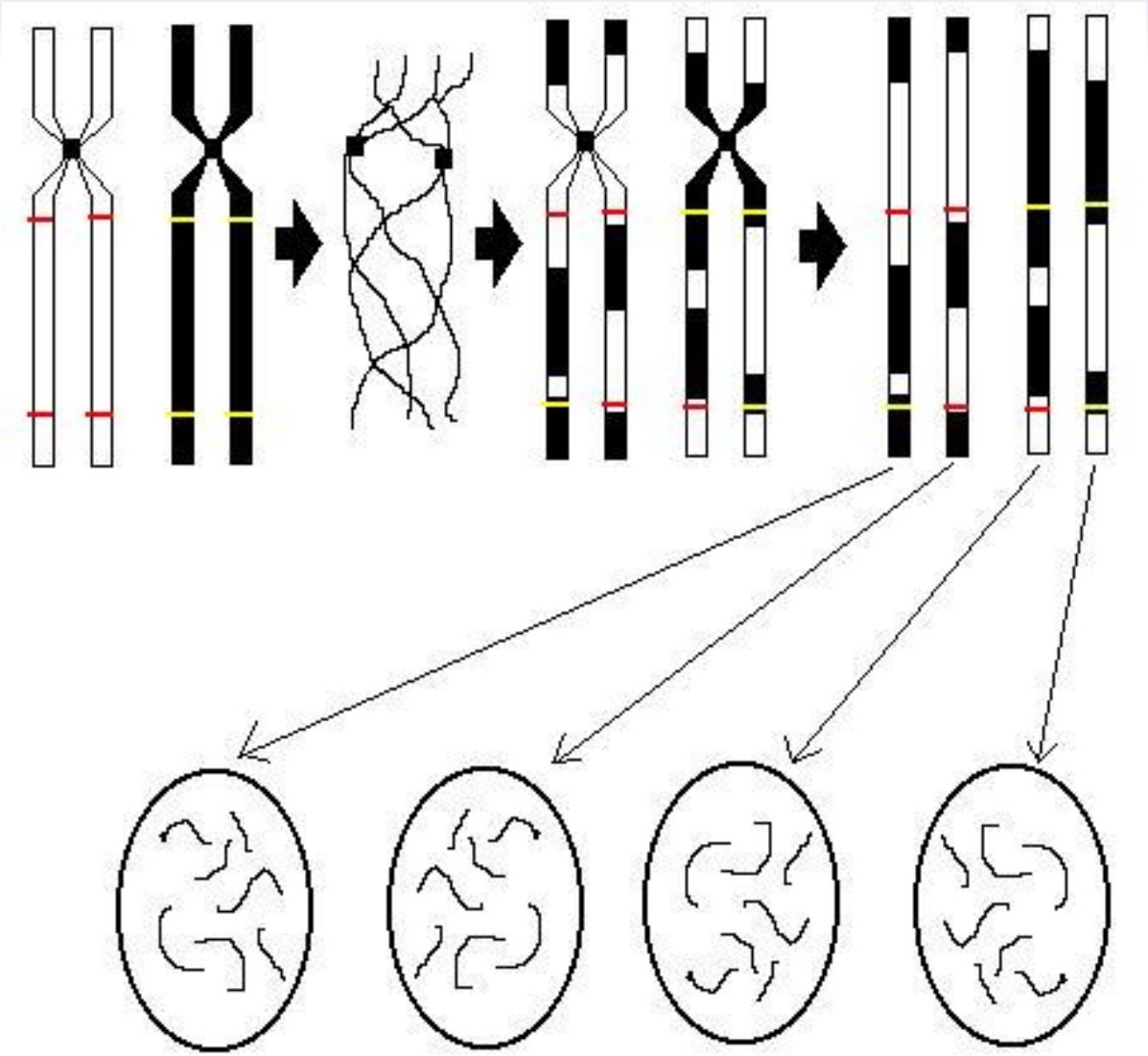
Mitose - Meiose



Recombination



Recombination



Linkage and markers

Two loci are recombinant (non-recombinant) if an odd (even) number of crossovers occur between them. The recombination fraction is the probability that two loci become recombinant during meiosis.

Alleles at loci closely located on the same chromosome will tend to inherit together. So: some combinations of alleles on these short segments may be preserved over a large number of generations.

A marker is a gene or DNA sequence with a known location on a chromosome. Genetic markers can be used to find the location of the disease-gene.

Linkage and markers

Two loci are recombinant (non-recombinant) if an odd (even) number of crossovers occur between them. The recombination fraction is the probability that two loci become recombinant during meiosis.

Alleles at loci closely located on the same chromosome will tend to inherit together. So: some combinations of alleles on these short segments may be preserved over a large number of generations.

A marker is a gene or DNA sequence with a known location on a chromosome. Genetic markers can be used to find the location of the disease-gene.

Linkage and markers

Two loci are recombinant (non-recombinant) if an odd (even) number of crossovers occur between them. The recombination fraction is the probability that two loci become recombinant during meiosis.

Alleles at loci closely located on the same chromosome will tend to inherit together. So: some combinations of alleles on these short segments may be preserved over a large number of generations.

A marker is a gene or DNA sequence with a known location on a chromosome. Genetic markers can be used to find the location of the disease-gene.

What do we know, want and do?

Relations:

$$\text{marker allele} \Longleftrightarrow \text{allele at disease gene} \Longrightarrow \text{Phenotype}$$

We are interested in:

$$\text{marker allele} \Longleftrightarrow \text{allele at disease gene} \Longrightarrow \text{Phenotype}$$

We measure:

$$\text{marker allele} \Longleftrightarrow \text{allele at disease gene} \Longrightarrow \text{Phenotype}$$

What do we know, want and do?

Relations:

$$\text{marker allele} \Longleftrightarrow \text{allele at disease gene} \Longrightarrow \text{Phenotype}$$

We are interested in:

$$\text{marker allele} \Longleftrightarrow \text{allele at disease gene} \Longrightarrow \text{Phenotype}$$

We measure:

$$\text{marker allele} \Longleftrightarrow \text{allele at disease gene} \Longrightarrow \text{Phenotype}$$

What do we know, want and do?

Relations:

$$\text{marker allele} \Longleftrightarrow \text{allele at disease gene} \Longrightarrow \text{Phenotype}$$

We are interested in:

$$\text{marker allele} \Longleftrightarrow \text{allele at disease gene} \Longrightarrow \text{Phenotype}$$

We measure:

$$\text{marker allele} \Longleftrightarrow \text{allele at disease gene} \Longrightarrow \text{Phenotype}$$

Association study

Phenotype: Cases and Controls

The analysis is based on the fact that for markers in close vicinity of a disease gene, some marker alleles are over represented among cases and under represented among controls.

Suppose the marker has alleles m and M. The m allele-frequencies among cases and controls are denoted as q_{case} and q_{con}. We want to test the hypothesis:
$H_0 : q_{case} = q_{con}$ versus $H_1 : q_{case} \neq q_{con}$.

Association study

Phenotype: Cases and Controls

The analysis is based on the fact that for markers in close vicinity of a disease gene, some marker alleles are over represented among cases and under represented among controls.

Suppose the marker has alleles m and M. The m allele-frequencies among cases and controls are denoted as q_{case} and q_{con}. We want to test the hypothesis:

$H_0 : q_{case} = q_{con}$ versus $H_1 : q_{case} \neq q_{con}$.

Association study

Phenotype: Cases and Controls

The analysis is based on the fact that for markers in close vicinity of a disease gene, some marker alleles are over represented among cases and under represented among controls.

Suppose the marker has alleles m and M. The m allele-frequencies among cases and controls are denoted as q_{case} and q_{con}. We want to test the hypothesis:

$H_0 : q_{case} = q_{con}$ versus $H_1 : q_{case} \neq q_{con}$.

Data

Sample cases and controls.
The m marker allele frequencies among the cases and controls can be estimated by the sample frequencies:

$$\hat{q}_{con} = \frac{\#\{m \text{ alleles controls}\}}{\#\{\text{alleles controls}\}} \qquad \hat{q}_{case} = \frac{\#\{m \text{ alleles cases}\}}{\#\{\text{alleles cases}\}}$$

Association study, the test

Hypothesis: $H_0 : q_{case} = q_{con}$ versus $H_1 : q_{case} \neq q_{con}$.

Test-statistic: $$T = \frac{\hat{q}_{case} - \hat{q}_{con}}{\sqrt{\widehat{\mathrm{Var}}(\hat{q}_{case} - \hat{q}_{con})}}.$$

Reject H_0 for $|T| \geq \xi_{\alpha/2}$.
The threshold $\xi_{\alpha/2}$ was chosen so that if H_0 is true, the probability that H_0 is rejected is at most α.

If the test is performed for many markers, quite a number of the tests will be falsely rejected.

Association study, the test

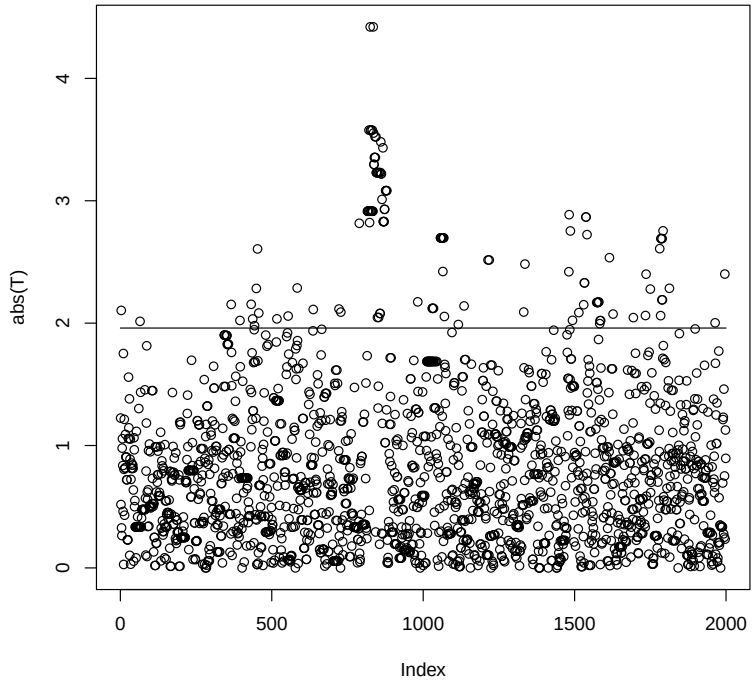
Hypothesis: $H_0 : q_{case} = q_{con}$ versus $H_1 : q_{case} \neq q_{con}$.

Test-statistic: $$T = \frac{\hat{q}_{case} - \hat{q}_{con}}{\sqrt{\widehat{\mathrm{Var}}(\hat{q}_{case} - \hat{q}_{con})}}.$$

Reject H_0 for $|T| \geq \xi_{\alpha/2}$.
The threshold $\xi_{\alpha/2}$ was chosen so that if H_0 is true, the probability that H_0 is rejected is at most α.

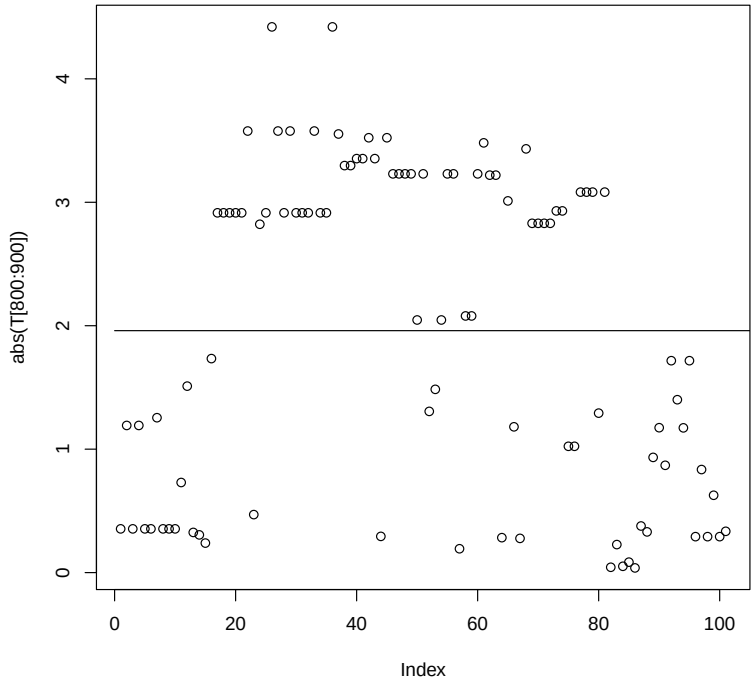
If the test is performed for many markers, quite a number of the tests will be falsely rejected.

Simulation study



Where is the disease gene?

Simulation study, markers 800-900

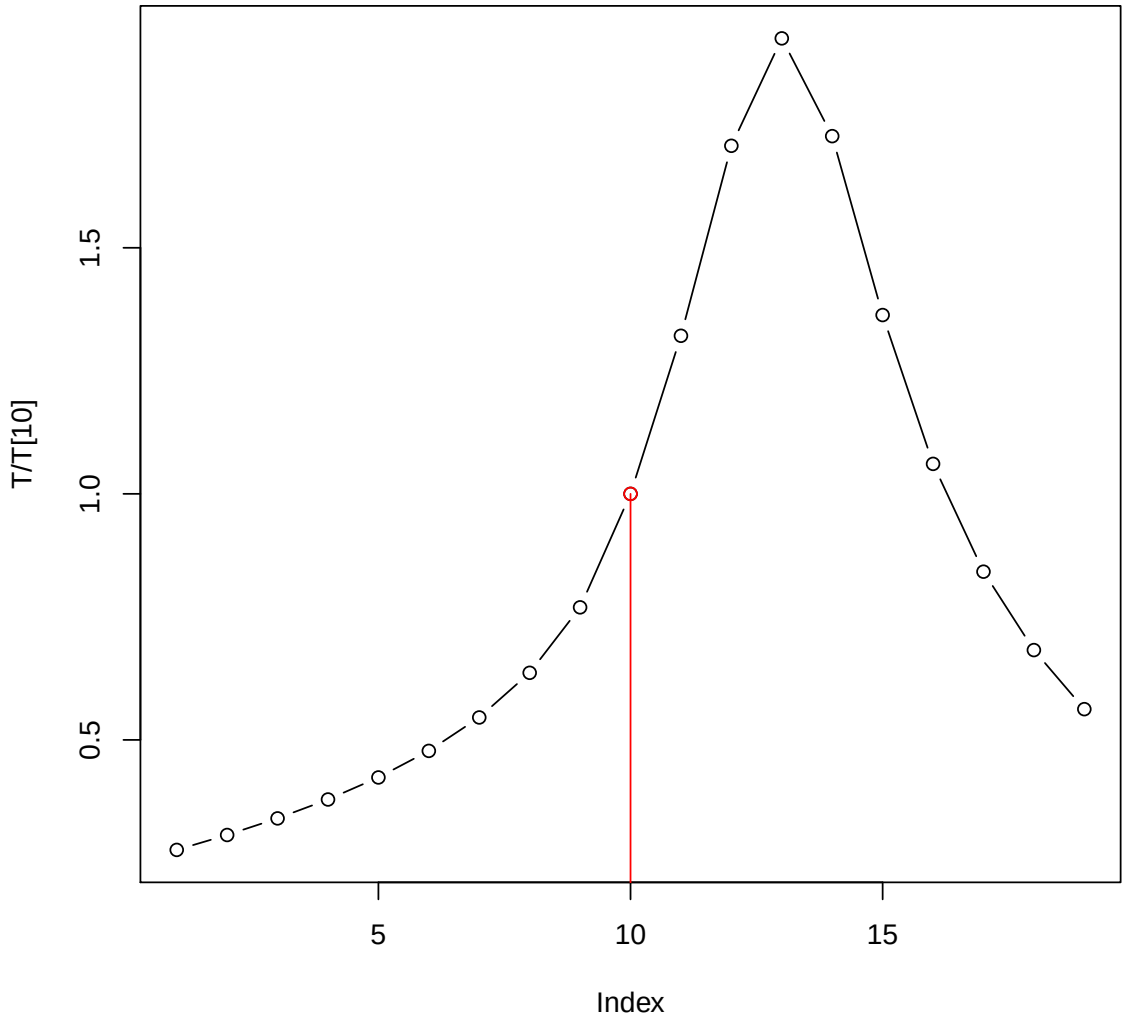


Where is the disease gene?

Question

Is the disease gene located closest to the marker with the highest value of the test-statisitc?

Numerical Example



Question

Is the disease gene located closest to the marker with the highest value of the test-statisitc? NO

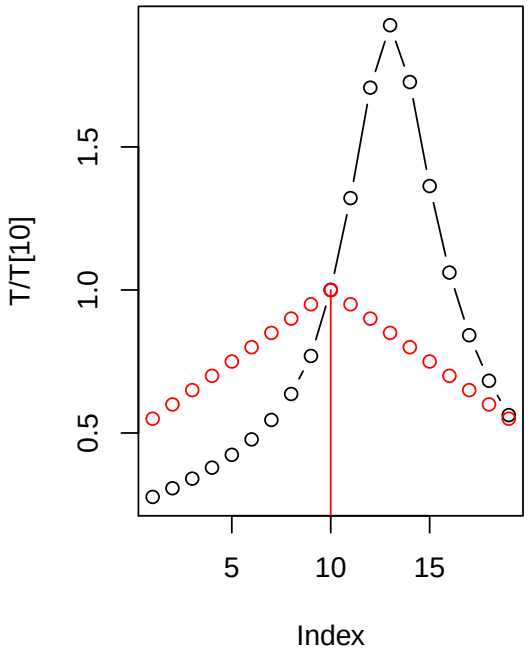
New questions:
Why not?
How can we find the disease gene?

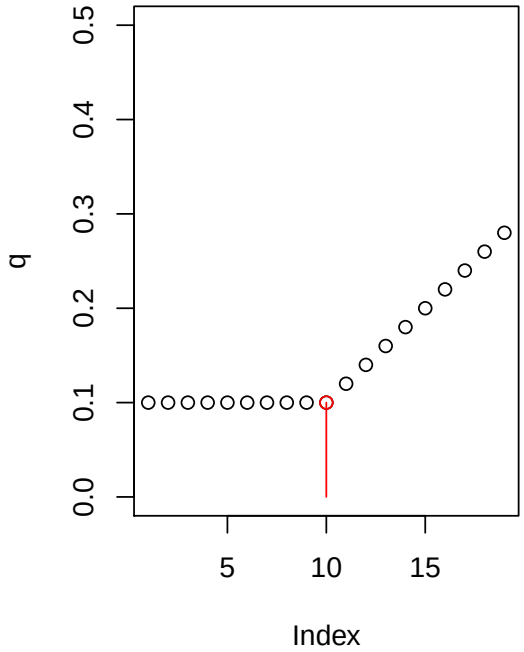
Question

Is the disease gene located closest to the marker with the highest value of the test-statisitc? NO

New questions:
Why not?
How can we find the disease gene?

Numerical Example, Why not?





Estimation of Δ

It can be proven that, for large numbers of cases and controls, the correlation between the marker and the disease gene is equal to

$$C\Delta \approx Q\,T$$

with Q a function of the marker allele frequencies in the samples of cases and controls, T the test-statistic and C a constant.

Conclusions

- A case-control study can be used for localizing disease genes
- Especially if the allele frequencies at the disease gene is small, the marker might not be located closest to the disease gene.
- Estimates of Δ should be considered too.

Conclusions

- A case-control study can be used for localizing disease genes
- Especially if the allele frequencies at the disease gene is small, the marker might not be located closest to the disease gene.
- Estimates of Δ should be considered too.

Conclusions

- A case-control study can be used for localizing disease genes
- Especially if the allele frequencies at the disease gene is small, the marker might not be located closest to the disease gene.
- Estimates of Δ should be considered too.

Reference

On testing for association in a case-control study

by M.A. Jonker, Z. Bochdanovits and A.W. van der Vaart.

Thank you