Lecture 9

Introduction to bioinformatics (MVE510)

Autumn, 2020

Repetion

- Analysis of RNA-seq data consist of three main steps
  1. Quantification of gene expression
  2. Normalization
  3. Identification of differentially expressed genes.
- Linear models are flexible statistical tools that can be used to analyze gene expression data from RNA-seq experiments.
- Each gene is analyzed separately – this results in a large number of linear models, coefficients and p-values.
- Visualization of the results for all gene simultaneously can be used to verify the model and identify genes that are differentially expressed.

Multiple testing

- In bioinformatics we are working with high-dimensional data.
- Genome sequencing
  - Large number of positions in the genome where each tested for the presence of mutations.
- Transcriptomics
  - Expression of thousands of genes are measured where each gene is tested for differentially expression.
- Metagenomics
  - Bacterial communities contains tens of thousands of species where each species is tested for differentially abundance.

High-dimensional data = many tests

The statistical hypothesis test

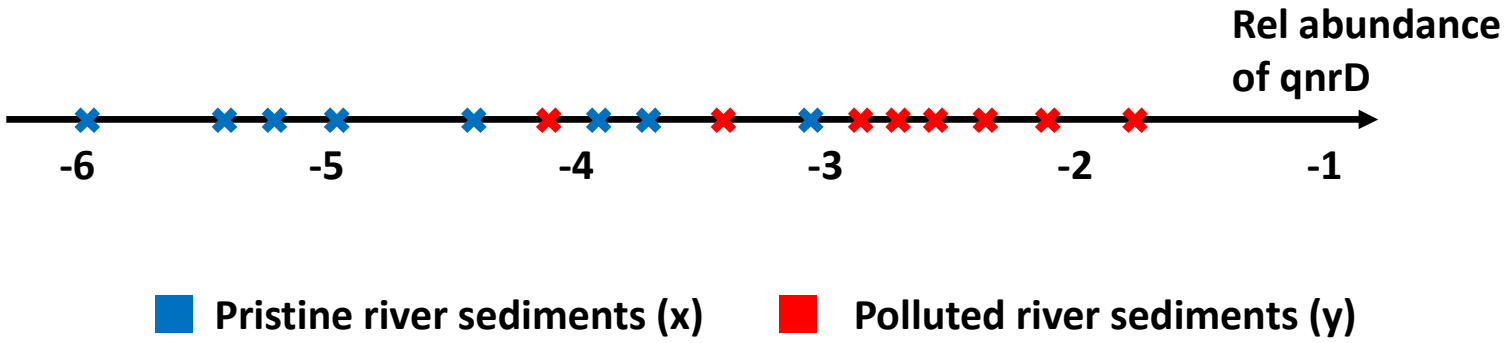
1. Formulate assumptions of the data.

2. Formulate statistical hypotheses:
   Null hypothesis,
   Alternative hypothesis.

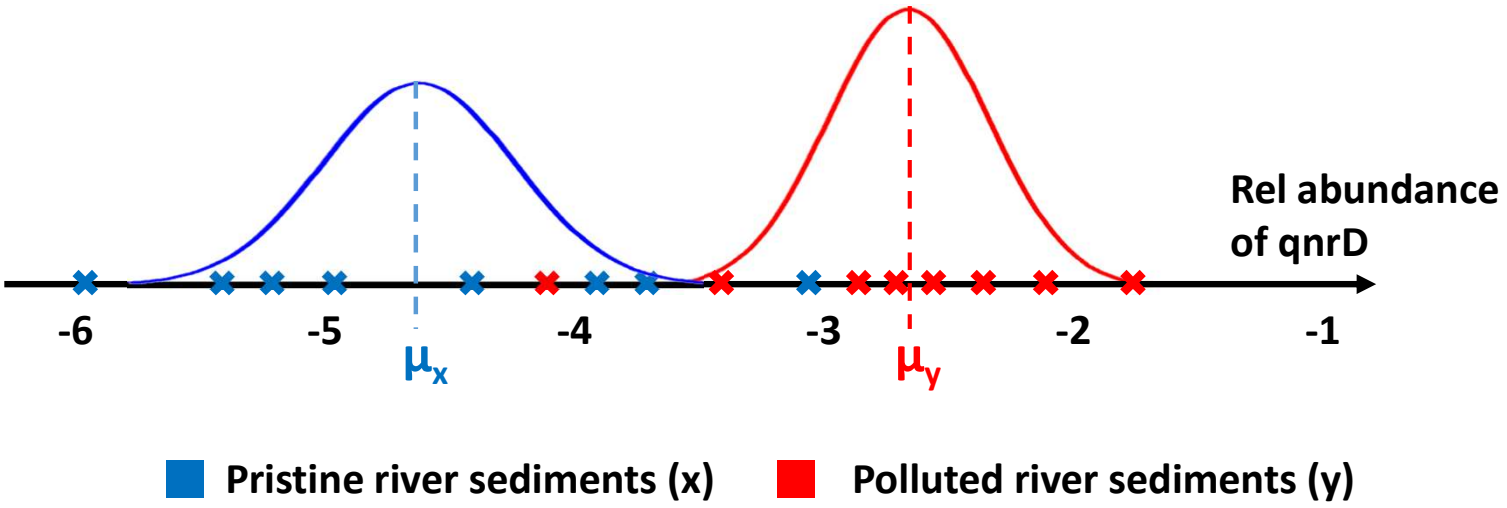
3. Apply a decision rule that decides when null hypothesis should be rejected.

A familiar example: the t-test

- Antibiotic resistance genes (qnrD) in pristine and polluted river sediments
- Measured by qPCR and quantified relative 16s rDNA. Data was then log-transformed.
- Replicates: 8+8



A familiar example: the t-test

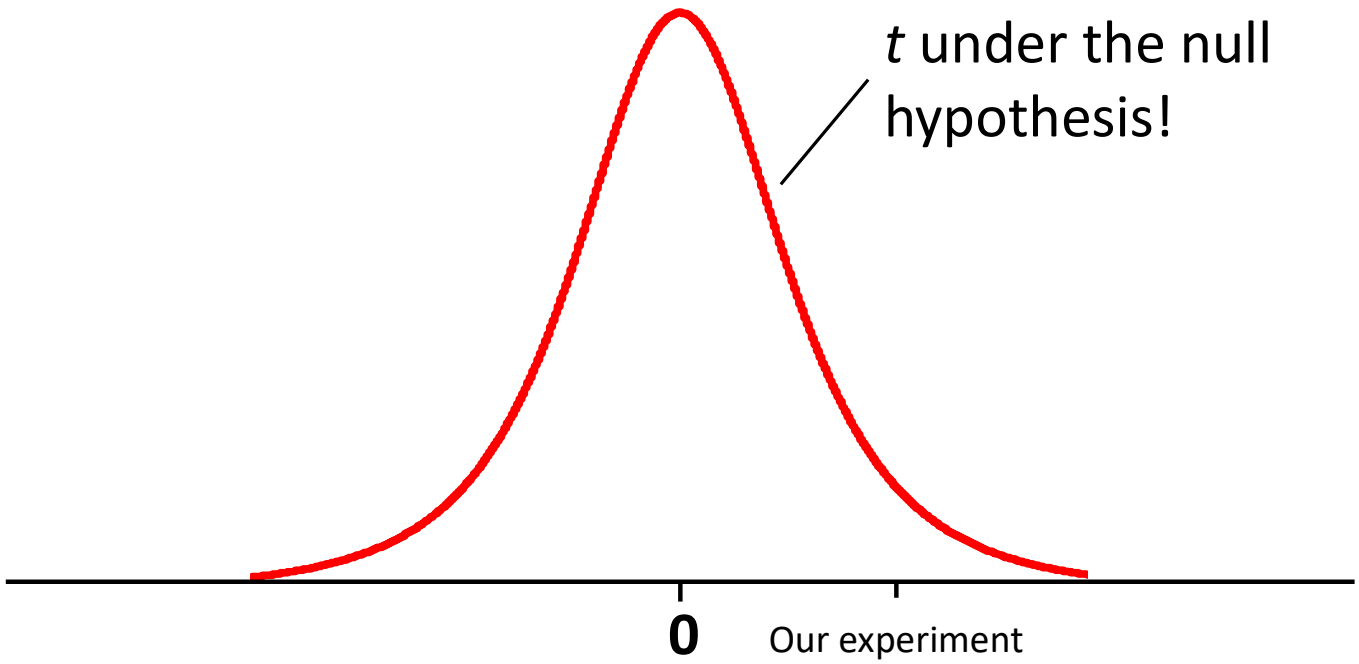


H_0: qnrD has the same abundance in pristine and polluted sediments ($\mu_y=\mu_x$)

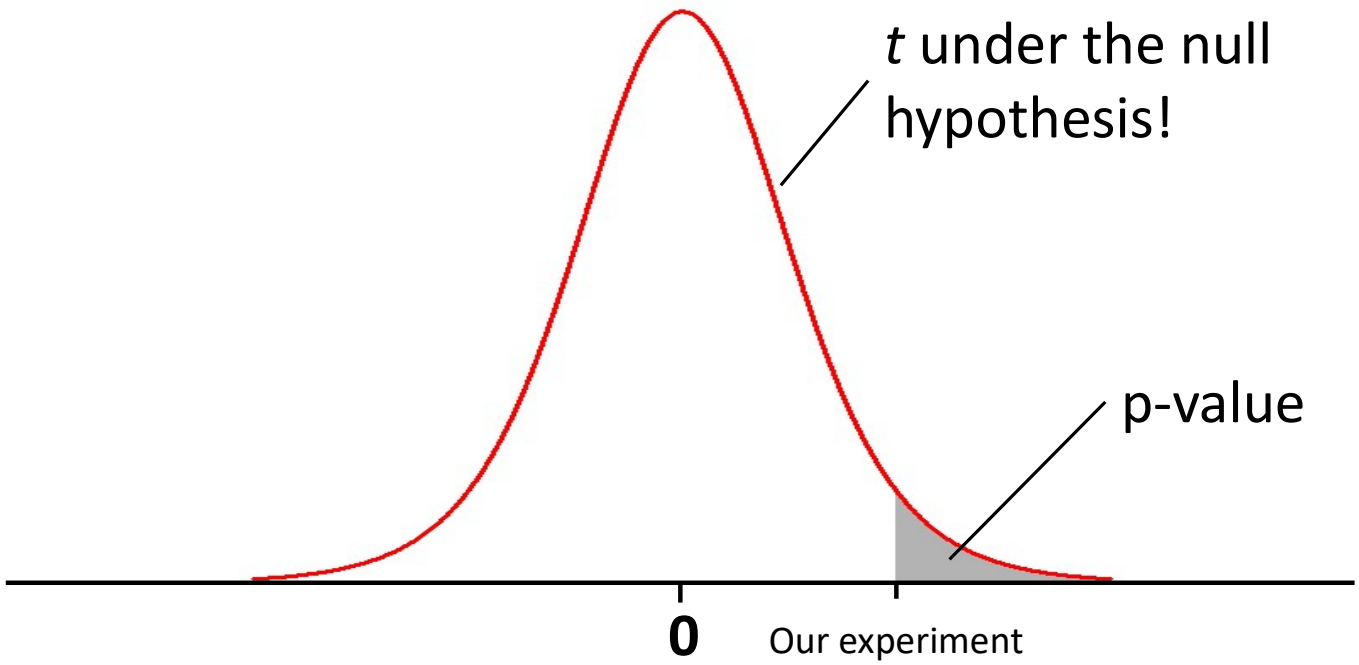
H_A: qnrD is more abundant in polluted sediments ($\mu_y>\mu_x$)

$$t = \frac{\bar{y} - \bar{x}}{\sqrt{\mathrm{SE}_y^2 + \mathrm{SE}_x^2}}$$

A familiar example: the t-test



A familiar example: the t-test



Facts about the p-value

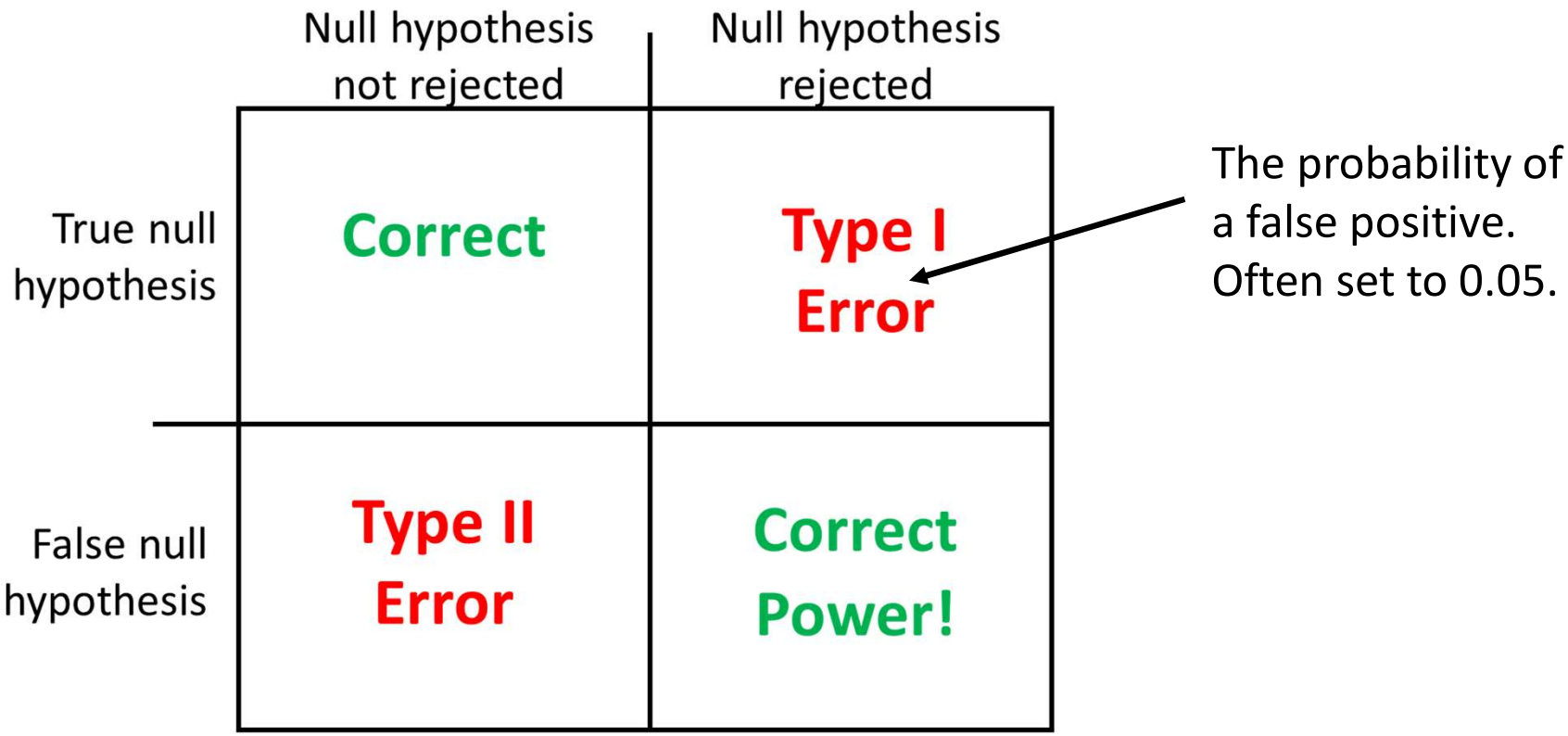
1. Measures deviation from the null hypothesis ("significance").

2. When the null hypothesis is true, the p-value is a random value between 0 and 1.

3. Summarizes the entire testing procedure.

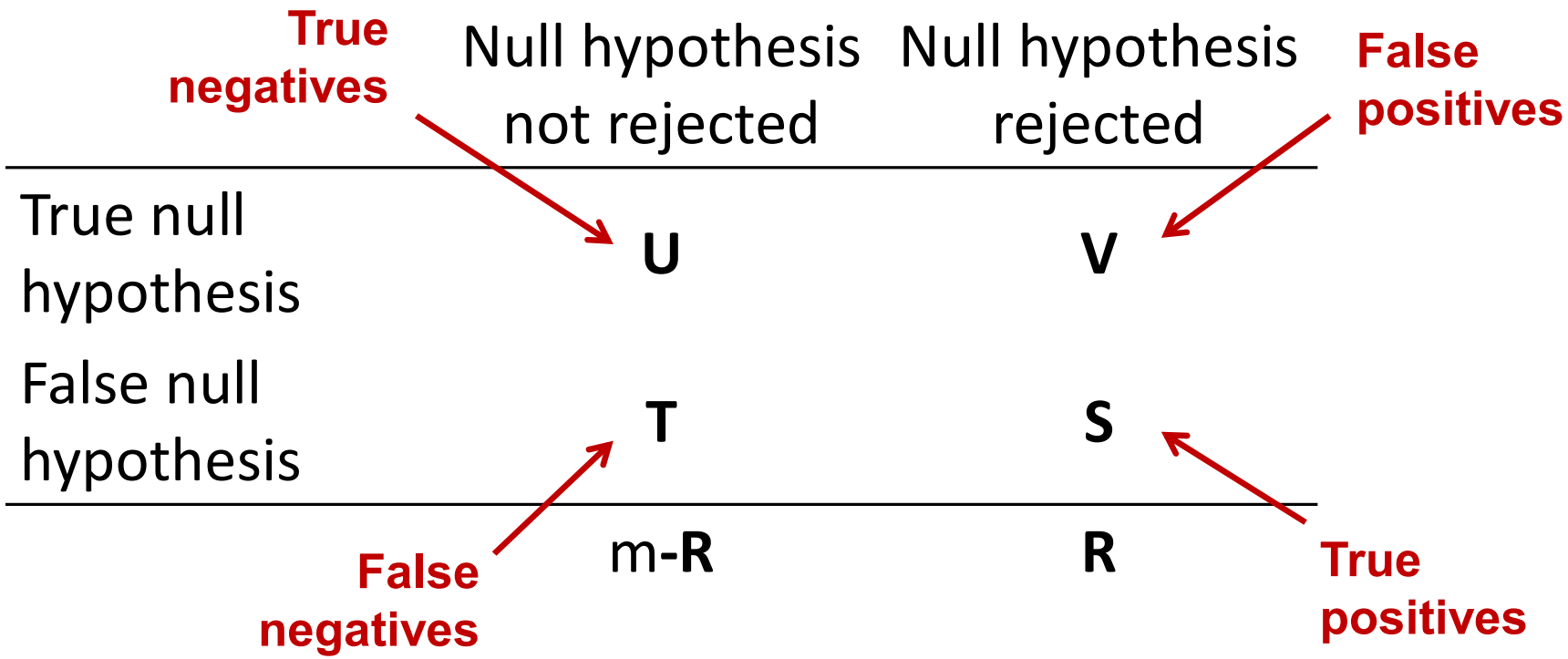
"The absence of evidence is not the evidence of absence"

The outcome of a test



Multiple testing

Assume that we perform m tests.



Multiple testing

Assume that we perform m tests.

Performing m tests at a level α where H_0 is true result in $m \times \alpha$ false positives (in average).

Correcting for multiple testing means controlling the number of false positives!

Multiple testing - FWER

Family-wise error rate (FWER)

The probability of at least one false positive, i.e.

$$\mathrm{FWER} = \mathrm{Prob}(V > 0)$$

Under the null hypotheses (H_0 is true for the m tests)

$$\mathrm{Prob}(V > 0) = \mathrm{Prob}(\text{at least one false positive}) = 1 - (1 - \alpha)^m$$

Multiple testing: FWER

$$\text{Prob}(\text{at least at least one false positive}) = 1 - (1 - \alpha)^m$$

Number of performed tests	Probability of at least one false positive
1	0.05
2	0.098
5	0.226
10	0.401
100	0.99

The Bonferroni correction

Algorithm: Bonferroni correction

Divide the significance level α (p-value cut-off) by the number of performed test (i.e. m).

A Bonferroni adjusted p-value can be calculated by multiply each p-value with the number of performed tests (i.e. m).

It can be shown that Bonferroni corrected p-values always control the Family Wise Error Rate (using Boole's inequality).

Interpretation: If we perform m tests with and use the significance cut-off α/m, then $\mathrm{Prob}(V > 0) \leq \alpha$.

Bonferroni correction: example

In a study, the entire genome of *Escherichia coli* was sequenced and compared to a reference. In total, 4,639,221 position was tested ('called') for single nucleotide polymorphisms (SNP). At a significance level of 0.05 we expect 232,961 false positives. In total 237,961 test had a p-value <0.05.

If the tests are corrected for multiple testing using the Bonferroni method, the significance cut-off is set to $0.05/4{,}639{,}221 = 1.07\times10^{-8}$.

Alternatively, a Bonferroni-adjusted p-value can be calculated as

$$p_{adj} = \min(1, p \times 4{,}639{,}221).$$

Positions with a $p_{adj} < 0.05$ are then considered significant.

After correction with Bonferroni, 3 tests were significant.

The Bonferroni correction

As always, there is no free lunch…..

A stricter p-value cut-off comes at the cost of reduced statistical power!

Bonferroni and controlling the FWER is often considered too conservative!

Multiple testing: FDR

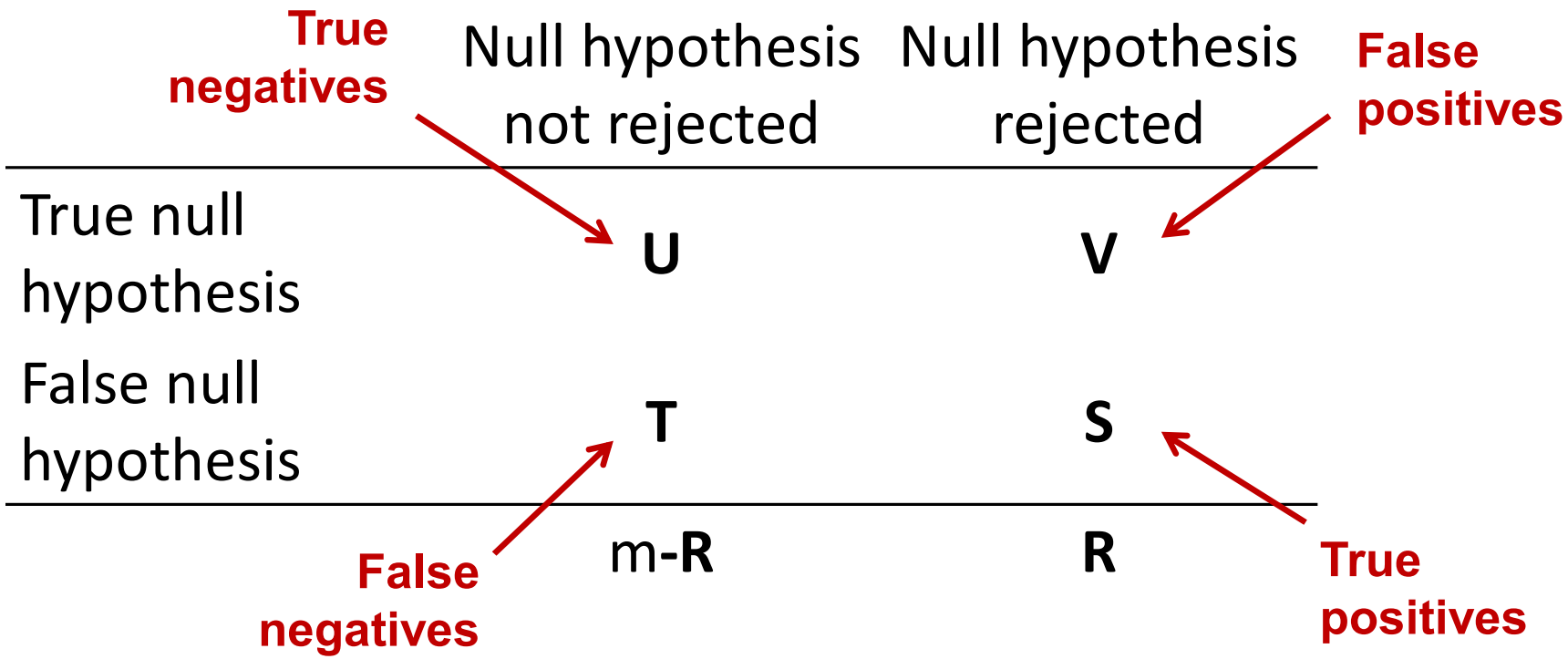
False discovery rate (FDR)

- The number of false positive in relation to the total number of rejected null hypotheses (significant tests)

$$\mathrm{FDR} = \mathrm{Exp}[\mathrm{V}/\mathrm{R}]$$

Multiple testing

Assume that we perform m tests.



Multiple testing: FDR

Algorithm: Benjamini-Hochberg q-values

1. Order the p-values from the m tests as $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$.
2. The Benjamini-Hochberg q-value for test (i) is

$$q_{(i)} = \frac{m}{i} p_{(i)}$$

3. Reject all tests where $q_{(i)} <$ cut off.

Multiple testing: FDR

- Benjamini-Hochberg correction controls the false discovery rate (FDR). Proof is complicated and based on several assumptions that are fullfilled if
  - The individual tests are independent.
  - P-values from tests where the null hypothesis is true are uniformly distributed (i.e. that the statistical distributional assumptions are true).

- The FDR cut-off can be set to the proportion of false positives that you can handle (0.05-0.20 is common in genomics).

Multiple testing – an example

P-value	Bonferroni correction	BH q-values
0.005	0.05	0.05
0.01	0.10	0.05
0.02	0.2	0.067
0.10	1	0.25
0.20	1	0.40
0.30	1	0.50
0.40	1	0.57
0.50	1	0.63
0.60	1	0.67
0.70	1	0.70

False discovery rate: example

In a study, the entire genome of *Escherichia coli* was sequenced and compared to a reference. In total, 4,639,221 position was tested ('called') for single nucleotide polymorphisms (SNP). At a significance level of 0.05 we expect 232,961 false positives. In total 237,961 test had a p-value <0.05.

The Benjamini-Hochberg algorithm were used to calculate q-values for all tests.

In total 34 tests had a q-value < 0.05. These tests have an estimated false discovery rate of 5% (i.e. 1 of 20 tests are, on average, false positives)

In total 257 tests had a q-value < 0.20. These tests have an estimated false discovery rate of 20% (i.e. 1 of 5 tests are, on average, false positives).

Multiple testing – a few tips

- For high-dimensional genomics data
  - Correction for multiple testing should always be done to ensure that there is not too many false positives. In research, correction for multiple testing is often a requirement.
  - Bonferroni is often too strict and a small proportion of false positives is often acceptable. Estimation of the false discovery rate is therefore often preferred.
  - It is good to minimize the number of comparisons. Only test what you need to know!