Integrative Transcriptomics
Dr. K. Nieselt, Faculty of Science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lecture: Microarrays Bioinformatics — WS 2017/18

Assignment No. 10

(5 points + 2 Bonus)

Hand out:	Thursday, January 25
Hand in due:	Thursday, February 1, 10:00
Tutorial date:	Tuesday, January 30, 10:15-11:45
Direct inquiries to:	fabian.moertter@student.uni-tuebingen.de

Theoretical Assignments

1. **Hypothesis testing and p-values of normal distribution** (5p)

   (a) A sample x with k replicates is collected. One wants to test, whether the sample's mean significantly differs from a given value μ_0. It is assumed that the sample is drawn from a normally distributed population.

   i. State the null hypothesis and set up the respective test statistic.
   ii. Think of an application of this test in the context of gene expression data.

   (b) Assume that expression values of an experiment are *normally* distributed with mean ($\log_2$) value 7.3 and standard deviation 3.2.

   i. Which expression range contains expression values with 95% (99%) probability?
   ii. What is the expression range for the first quartile?
   iii. What is the probability that a gene has an expression value above 10?

Practical Assignments

The material is available from `ILIAS` (`A10.zip`).

2. **t-Test: application to expression data**

   Use the Golub data set from the material for this assignment. Recall that this data represents an expression screening of two types of leukemia, AML and ALL. This data set contains already normalized expression data. The column names indicate the class affiliation.

   (a) Write your own function, that implements the two-sample t-test (here you are asked not to make use of the existing R function `t.test`). The function parameters should contain the *kind of test* (one-sided or two-sided) and the *significance level* α. The input and output should be similar to the existing R functions that allow the user to conduct a t-test, especially the p-value and t-statistics should be part of the output.

   (b) Apply your function to the Golub data to compute a two-sided t-test, in order to compute the number of differentially expressed genes between the ALL and AML samples for different significance levels ($\alpha = 0.05, 0.01, 0.001$).

   (c) Next apply your function to the Golub data to compute two one-sided t-tests, in order to compute up-regulated genes in the ALL samples and/or down-regulated genes in ALL again for different significance levels ($\alpha = 0.05, 0.01, 0.001$).

   (d) Output the number of differentially expressed genes and number of up/down-regulated genes in form of a table for each level α. What do you find: Are more genes up-regulated than down-regulated in ALL? Does the number of differentially expressed genes and/or up-/down-regulated genes vary significantly for the different α levels?

3. **Volcano plot** (2 Bonus)

(a) Implement a function that produces the so-called volcano plot, ie., a scatter plot of the $(-\log(\text{p-value}))$ against the fold change. The p-value can come from any statistical test.

(b) Apply your function to the vector of computed mean fold changes of the genes between the ALL and AML samples in the Golub data and the p-values of the two-sided t-test of the Golub data ($\alpha = 0.05$). The plot should also visualize for a given significance threshold the unchanged, down-regulated (in ALL) and up-regulated (in ALL) genes by differently colored points.

(c) Explain the interpretation of the volcano plot and discuss your results.

Please read the questions carefully. If there are any questions, you may ask them during the tutorial session or via e-mail to your tutor. You will usually get an answer in time, but late e-mails (e.g. on Thursday morning before class) might not be answered in time. Please upload your solutions in the Ilias system. Please pack your source code, the plots, as well as the theoretical part into **one single archive file (zip)**. Source code should compile correctly.