Modeling Cell Fate

Prof. Dr. Volkhard Helms
Maryam Nazarieh, Thorsten Will
Summer Semester 2015

Saarland University
Chair for Computational Biology

Exercise Sheet 6

Due: Juli 28, 2015 10:15 am

Submission

- You are advised to work in groups of two people. If necessary, we will suggest teammates.
- Submit your solutions on paper in Room 3.02, E2.1. Alternatively you may send an email with a single PDF attachment to thorsten.will@bioinformatik.uni-saarland.de. Late submissions will not be considered.
- The PDF should contain written answers to all questions as well as the source code and the output of your scripts. Everything that is missing will not be graded.
- Do not forget to mention your names/matriculation numbers.
- Discussion of this exercise will be on August 4th at 12:45 in the lecture room (E2.1 007).

Exercise 6.1: Mixed questions (30 points)

(a) What is Knudson's two-hit hypothesis? Briefly explain it in your own words. (10 points)

(b) Imagine you have a set of hypotheses that you wish to test simultaneously. Normally you test each hypothesis separately using some level of significance α. However, consider a case where you have 20 hypotheses to test and a significance level of 0.05. What is the probability of observing at least one significant result just due to chance? (10 points)

(c) Bonferroni correction is a method used to counteract the multiple testing problem mentioned above. Briefly explain the principle and mention at least one alternative procedure. How can you expand a statistical test in a very simple way to account for the correction? (10 points)

Exercise 6.2: Handling TCGA data (30 points)

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze a huge amount of cancer-related data. It contains clinical information, genomic characterization data, and high level sequence analysis of many tumor genomes.
The data hub can be found at `https://tcga-data.nci.nih.gov/tcga/`.

(a) TCGA samples are labeled with a special barcode, for example TCGA-KL-8324-01A-11R-2315-07. Outline what the barcode can tell you. How can you distinguish normal and tumor tissue based on this code? (5 points)

(b) Download all Kidney Chromophobe (KICH) RNASeqV2 expression data from TCGA. How large is the data and how many files are part of the archive? Describe briefly what this data contains and how it was generated/processed (tool(s) sufficient). Which identifier(s) were used to name genes? (15 points)

(c) The KICH data comprises matched data for normal and (solid) tumor tissue from the same patients. Use the included metadata in **FILE_SAMPLE_MAP.txt** to relate barcodes to data files and to find out which files belong to the same patients. How many matched sample pairs are present? (10 points)

Exercise 6.3: Differential gene expression (40 points)

In this exercise you will implement a simple tool that detects differentially expressed genes. The file **expr_data.csv** in the supplementary data to this assignment contains gene expression data for matched pairs of normal and tumor tissue precompiled from TCGA. The group of healthy samples is found in the first half of the data, the data for cancerogenic samples in the second half. The order of the patients is the same in both categories.

(a) What is the advantage of matched sample pairs? What is the disadvantage? (5 points)

(b) In this exercise you will apply the Wilcoxon signed-rank test. An appropriate alternative would be the paired T-test. What is the difference between the two tests and what makes them special compared to, for example, the Mann-Whitney U test? (10 points)

(c) Read in the expression data and check for every gene if the average expression is higher or lower in the tumor tissue. Use a one-sided Wilcoxon signed-rank test and Bonferroni correction to test if the difference is significant. You are free to use an implementation of the test from any package (for example, from scipy.stats in Python).
Furthermore, make sure you compute one-sided p-values. If the implementation returns two-sided p-values, how can you convert them in this case? Also, how many hypothesis are tested in this exercise (important for Bonferroni correction)?
Report how many genes are significantly up-/down-regulated in tumors and report the five genes with lowest p-values in each case (up/down). (20 points)

(d) Check if the 50 most significantly up-regulated genes are enriched in any KEGG pathways. Use the DAVID webservice (http://david.abcc.ncifcrf.gov/) or GeneTrail2 (http://genetrail2.bioinf.uni-sb.de/) for this task. What is the appropriate set of background genes in our case and why? Report your results. (5 points)