CS-E5875 High-Throughput Bioinformatics
2.11.2018

Exercise 1

After the exercise session, you should submit a short report of your solutions to the assignments 1 and 2 to MyCourses. **Assignment 0 does not have to be reported, as it is an R tutorial**. Explain the steps in your analysis, show your implementation when needed, include the plots and results and discuss your results. The deadline for the report submission is on **9th November at 12.00**.

0. (0 points) During this course we will do some programming using R language. Therefore, we start with an introduction to R programming. In this assignment, we will go through some basic R operations that would be of use to you in the future. If you want to use your own laptop for the exercises, please install R and RStudio beforehand from `https://cran.rstudio.com/` and `https://www.rstudio.com` respectively. The file `handout.pdf` gives an overview of R and some sample scripts that are useful for bioinformaticians. The R code snippets can also be found from file `handout.R`. We will go through the script together in class.

1. (2 points) Re-create the example shown in slides 16-19 in the Introduction lecture. First, compute the probabilities of obtaining x heads out of 30 coin flips using a fair coin (i.e. probability of heads= 0.5), where $x = 0, ..., 30$. Then plot the probabilities $p(x)$ against x. What is a probability of observing 20 or more heads? If you use $p \leq 0.05$ as a threshold to reject a null hypothesis, H_0 ="the coin is fair" (against alterative hypothesis H_A="the coin is not fair", i.e. probability of heads $\neq 0.5$), how many heads do you need to observe to reject the null hypothesis?

   (Useful R functions: dbinom, plot)

2. (4 points) Hypothesis testing using t-test and multiple correction. Use set.seed(1234).

(a) Generate normally distributed expression data for 100 genes for two groups A and B with 8 replicates for each group, where $\mu_A = \mu_B = 0$ and $\sigma_A^2 = \sigma_B^2 = 3$.
(Useful R functions: rnorm)

(b) For each gene, test the null hypothesis, $H_0 : \mu_A = \mu_B$, using t-test and plot the p-values in a histogram. How many genes have $p \leq 0.05$?
(Useful R functions: t.test, hist)

(c) Implement the Bonferroni method for adjusting p-values for multiple testing. (You can implement this as an R function, see for example `http://www.statmethods.net/management/userfunctions.html`).

Adjust the p-values using your implementation. How many genes have adjusted $p \leq 0.05$? Comment on differences with uncorrected p-values, if any.
(Useful R functions: pmin)

(d) Same as (c), but implement the Benjamini-Hochberg method. Adjust the p-values using your implementation. How many genes have adjusted $p \leq 0.05$? Comment on differences with uncorrected p-values, if any.

(e) Generate expression values for 90 genes for groups A and B, where $\mu_A = \mu = B = 0$ and $\sigma_A^2 = \sigma_B^2 = 2$ with 8 replicates for each group. Generate similar expression values for 10 genes but with $\mu_A = 0$, $\mu_B = 5$ and $\sigma_A^2 = \sigma_B^2 = 3$.

(f) For each gene, test the null hypothesis, $H_0 : \mu_A = \mu_B$, using t-test, when alternative hypothesis is $H_A : \mu_A \neq \mu_B$ and when alternative hypothesis is $H_A : \mu_A < \mu_B$. How many genes have $p \leq 0.05$?

(g) Correct the p-values for multiple testing using your implementation of both methods. How many genes have adjusted $p \leq 0.05$? Comment on differences, if any.