Assignments to Lecture 6 – Statistical testing

Danny Arends

Please note *this data is from a current analysis, and is not public (yet)*.

arrays.txt contains information on the different microarrays:

- Filename – The original name of the file from the company
- CompID – ID used by the company
- Strain – Mouse strain on the array
- Tissue – Which tissue was on the array
- Individual – Our identifier of the individual

arraydata.txt contains the probe level measurements for each microarrays and sample, each row contains: Microarray probe name, Sequence of the probe, Intensity measurements

0) Load in the provided datasets using either the read.table or read.csv functions.

Preprocessing and Normalization

1)

a) Create a boxplot showing the expression levels across the different arrays

b) Log2 transform the expression data from arraydata.txt, and create another boxplot

c) Using the normalize.quantiles function (from the preprocessCore package) to normalize the data across different arrays

d) Create a boxplot showing the data before (1a) and after normalization

Correlation

We can use plots and correlation on the normalized data to have a quick look at our data.

2)

a) Create plots that compares each array versus the other arrays using correlation (so 1 plot for 1 array compared to all the other arrays) in total you should get 16 plots
Note: Using a for loop or the apply function

b) Calculate the correlation (Using the spearman method) matrix between the arrays and create an image plot showing how arrays are correlated. What do you learn from looking at the image plot ?

Basic data analysis

Using the arrays.txt file we can split the data into two groups: HT (hypothalamus) and GF (Gonadal fat), we see that the same individuals have been measured in both tissues. During this analysis we will not bother adjusting for the strain of the individuals (F1, BFMi and B6n) however in a normal analysis this does need to be taken into account.

3)

a) Split the normalized expression data into two groups (HT, and GF) using the data from the arrays.txt file

b) For each probe on the array:

   i) Calculate the mean and standard deviation in both groups

   ii) Perform a t.test between the two groups for each probe

   iii) Store the results in a matrix with 5 columns, so that the matrix looks like:

   meanHT, meanGF, sdHT, sdGF, p-value

   Note: Do not forget to add the probeID as rowname

   Note: You can get the p-value from the t-test using:

```
res <- t.test(x, y)
pval <- res$p.value
```

c) How many probes are significantly ($p < 0.05$) differentially expressed between Hypothalamus and Gonadal fat, when correcting for multiple testing using Bonferonni ?

d) Use the p.adjust function to correct the t-test pvalues using the Benjamini & Hochberg Procedure, how many probes are now significantly differential expressed ?