Assignment 2 — Total mark: 25

Due by 12:00pm by email [redacted] *on Sunday, May 31rst, 2013. Read the instructions carefully. The assignment will be submitted in the form of a .doc or .docx file. You will give the* `R` *code for each question and comment/interpret it. Comments on each question should not exceed 200 words and the total document should not exceed 10 pages. When including a figure, do not forget to add a succinct legend mentioning exercise number, question number and type of plot.*
Note that each question (1 - 4) can be answered independently.

The following questions relate to gene expression values published in "Comprehensive molecular portraits of human breast tumours"[1] which we have discussed in the Lecture.
The data is provided in an `expressionSet` object called `eSet`. The `eSet` object comprises gene expression values and phenotypic data.

1. The data. [Total mark: 3]

```
> library("Biobase")
> # load data
> load('Data/eSet.RData')
```

   (a) What phenotypic data is contained within the `eSet` object? Which function did you use to extract this information? [0.5pt]

   (b) Extract and store the expression data into a variable called '`data`'. What are the dimensions of `data`? How many genes and how many tumours comprise the data? [0.5pt]

   (c) We would like to filter out genes that are invariant across samples. Compute the standard deviation of each gene contained within `data`. Generate a histogram for the standard deviation values of all genes in `data`. [1pt]

   (d) Using a standard deviation cutoff of 0.7 produce a filtered data set called '`data2`'. How many genes have been retained for further analysis? [1pt]

2. Analysis of three therapeutic groups. [Total mark: 8].
   Breast cancer is a heterogeneous disease that can be categorized into three basic therapeutic groups: (i) Eostrogen Receptor (ESR1) postive; (ii) HER2 receptor (ERBB2); and (3) triple negative (characterized by a lack of expression of ESR1, HER2 and the Progesterone Receptor (PGR)).

   Supervised clustering of mRNA expression data has reproducibly established that breast cancers encompass five distinct disease entities, often referred to as the intrinsic subtypes of breast cancer. Tumours are classified into their respective subtypes by using the so called PAM50 classifier which is a list of 50 genes capable of descriminating between subtypes.

   (a) Extract `PAM50.mRNA` values from the `PhenoData` slot of `eSet` and create a factor variable called '`subtype`. Use the `summary()` function to determine how many tumours fall within each tumour subtype. [1pt]

[1] http://www.nature.com/nature/journal/v490/n7418/full/nature11412.html

(b) Produce a series of boxplots showing the relative expression of ESR1, ERBB2 and PGR in the five intrinsic subtypes of breast cancer. Use the factor variable `subtype` to partition the data. For each plot perform an ANalysis Of VAriance to test the null hypothesis that the population means of the subtypes are equal. Based on the significance level obtained for the test what do you conclude? What assumptions about the data have you made in applying the ANalysis Of VAriance tests? [3pt]

(c) Apply Tukey's 'Honest Significant Difference' method to the expression values for ESR1, ERBB2 and PGR across the subtypes. Interpret the results. [1pt]

(d) Which intrinsic subtype is triple negative? Which intrinsic subtypes are positive for ESR1 and PGR? [1pt]

(e) Test the association between ESR1 status and subtype. Would you use the chi-squared test or the Fisher's exact test and why? To answer the question use the factor variable `er.status` which can be generated as follows:

```
> er.status <- factor(eSet$ER.Status, levels=c("Negative", "Positive"), labels=c(0,1))
```

Are your results consistent with the results obtained above? [2pt]

3. Basal-like vs. Luminal A analysis. [Total mark: 12]

Patients diagnosed with basal-like tumours have a poor prognosis and do not respond to adjuvant tamoxifen (an antagonist of the eostrogen receptor) therapy. You will perform a differential gene expression analysis between tumours classified as Basal-like and Luminal A by answering the following questions:

(a) Create a factor variable called '`diff` with levels corresponding to tumours classified as either Basal-like or Luminal A (Hint: Use the `PAM50.mRNA` phenotypic data). Use the `summary()` function to determine how many tumours fall within each tumour subtype. [1pt]

(b) Before applying a two sample t-test check the assumption that each gene value follows a normal distribution. What do the Q-Q plots for ESR1, PGR and HER2 look like? Use an appropriate test to check the normality of all genes in data2. How many genes do not have a normal distribution? Use the `var.test()` function to assess if the variance is equal between tumour subtypes for each gene. How many genes have an unequal variance between the two groups of patients? [3pt]

(c) Perform scale normalisation on the samples of `data2` by subtracting the median and dividing by the Median Absolute Deviation (MAD) (see Section 2.4 in the Bioconductor Prac. material). Produce boxplots to compare the data distribution before and after scale normalisation (N.B. Only provide plots for the first 20 samples). Repeat the normality and variance tests used in Question 2b on the standardised data. Has the scale normalisation made a difference? [3pt]

(d) Perform a two-sample t-test on the standardised data to identify differentially expressed genes. Pay special attention to the `var.equal` argument - based on the results of the tests perfomed in question 2c what should the value of this argument be? How many genes are differentially expressed if we use a significance level of 0.01 as a cutoff? [2pt]

(e) Using the `p.adjust()` function, apply the Benjamin Hochberg multiple correction procedure. For a significance level of 0.01, after multiple correction, how many genes are differentially expressed? Did you find the genes ESR1 and PGR in your list? What are the adjusted p-values of these genes? Order the adjusted p-values in ascending order i.e. from lowest to highest and select the top 300

genes for further analysis. Store the names of the top 300 genes in a vector called `DE.genes`. [1pt]

(f) Perform hierarchical clustering using the `heatmap()` function from library `gplots` on the top 300 differentially expressed genes (Hint: Use the linkage method "ward" in `hclust`). Interpret the results. [2pt]

4. The PAM50 classifier. [Total mark: 2]
As stated above tumours are classified into their respective subtypes by using the so called PAM50 classifier which is a list of 50 genes capable of descriminating between subtypes. A vector containing these 50 genes is provided and is called `pam50.genelist`.

(a) How many of the PAM50 genes are found in your top 300 list and what are their names? [1pt]

(b) Using the `pam50.genelist` and the original matrix '`data`' perform hierarchical clustering using the `heatmap()` function from library gplots (Hint: Use the linkage method "ward" in hclust). How many major clusters are produced and do they correspond to the known intrinsic subtypes of breast cancer? [1pt]