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]
This is the function to use to extract the phenotypic data:

```
> varLabels(eSet)
```

```
 [1] "Gender"                          "Age.at.Initial.Pathologic.Diagnosis"
 [3] "ER.Status"                       "PR.Status"
 [5] "HER2.Final.Status"               "Tumor"
 [7] "Tumor..T1.Coded"                 "Node"
 [9] "Node.Coded"                      "Metastasis"
[11] "Metastasis.Coded"                "AJCC.Stage"
[13] "Converted.Stage"                 "Survival.Data.Form"
[15] "Vital.Status"                    "Days.to.Date.of.Last.Contact"
[17] "Days.to.date.of.Death"           "OS.event"
[19] "OS.Time"                         "PAM50.mRNA"
[21] "SigClust.Unsupervised.mRNA"      "SigClust.Intrinsic.mRNA"
```

(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]

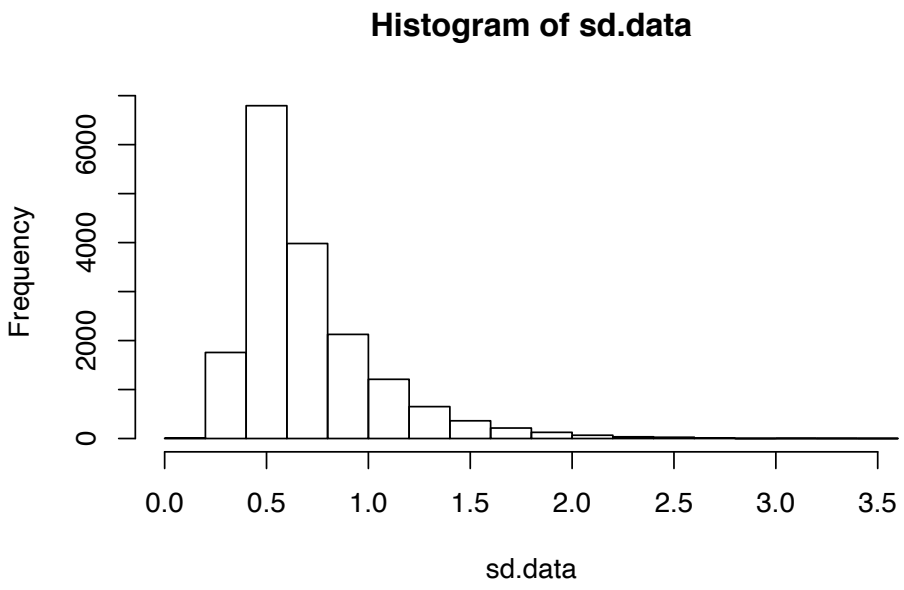
```
> data <- exprs(eSet)
> dim(data)
```

We have 348 tumours and 17373 genes.

(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]
We use the `apply()` function to do so.

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

```
> sd.data <- apply(data, 1, sd)
> hist(sd.data)
```



(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]

```
> selected <- names(which(sd.data >= 0.7))
> data2 <- data[selected,]
> dim(data2)
> length(selected)
```

There are 6467 genes left after prefiltering.

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.

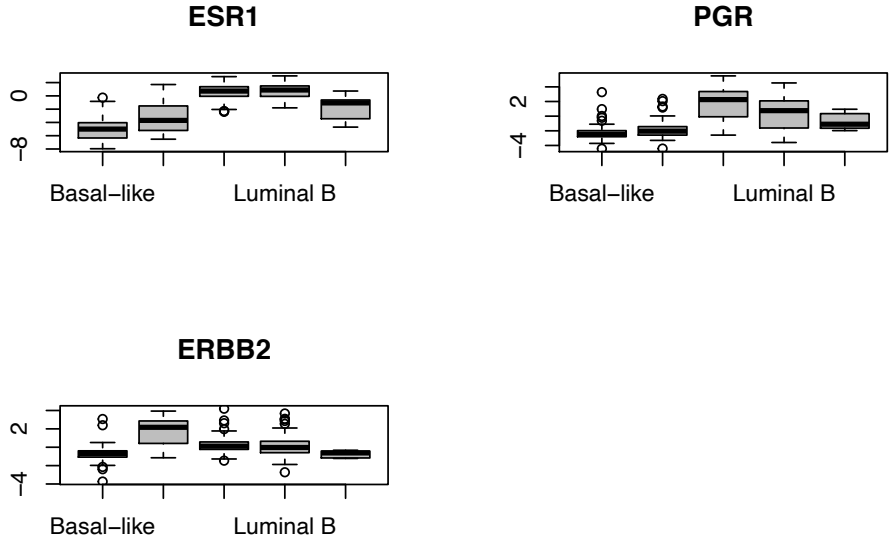
(a) Extract `PAM50.mRNA` values from the `PhenoData` slot of `eSet` and create a factor variable called '`subtype`. Use the `summary()` function to determine how many tumours fall within each tumour subtype. [1pt]

```
> subtypes <- factor(eSet$PAM50.mRNA) # or factor(pData(eSet)[,"PAM50.mRNA"])
> summary(subtypes)
```

(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]

```
> esr.values <- data2["ESR1",]
> pgr.values <- data2["PGR",]
> her2.values <- data2["ERBB2",]
> par(mfrow=c(2,2))  # divide the window into 2 rows 2 cols
> boxplot(esr.values ~ subtypes, main = 'ESR1', col = 'grey')
> boxplot(pgr.values ~ subtypes, main = 'PGR', col = 'grey')
> boxplot(her2.values ~ subtypes, main = 'ERBB2', col = 'grey')
> par(mfrow=c(1,1))
```



Wen applying an ANOVA, we assume that (3.3.3 in Lectures)

i. The samples are *independent* random samples, i.e. the results from one sample do not affect the measurements observed in another sample.

ii. Each sample is selected from a *normal* population.

iii. The mean and variance for population or group $k = 5$ are, respectively, μ_k and σ_k^2, $k = 1, \ldots, K$. The K *variances are equal*: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2 = \sigma^2$.

```
> summary(aov(esr.values ~ subtypes))

             Df Sum Sq Mean Sq F value Pr(>F)
subtypes      4 1940.4   485.1   247.9 <2e-16 ***
Residuals   343  671.2     2.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(aov(pgr.values ~ subtypes))

             Df Sum Sq Mean Sq F value Pr(>F)
subtypes      4  950.2  237.55   71.04 <2e-16 ***
Residuals   343 1146.9    3.34
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(aov(her2.values ~ subtypes))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
subtypes      4  161.0   40.24   43.63 <2e-16 ***
Residuals   343  316.4    0.92
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on these ANOVA outputs, we conclude that for each gene, there is at least one tumour subtype for which the mean expression levels for that gene is differentially expression from the other subtypes.

(c) Apply Tukey's 'Honest Significant Difference' method to the expression values for ESR1, ERBB2 and PGR across the subtypes. Interpret the results. [1pt]
The Tukey's 'Honest Significant Difference' tests the equality of the mean between pairwise subtypes. The adjusted p-value for multiple testing is indicated in the column 'p adj'.

```
> TukeyHSD(aov(esr.values ~ subtypes), conf.level=0.95)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = esr.values ~ subtypes)

$subtypes
                                 diff        lwr        upr     p adj
HER2-enriched-Basal-like   1.75065356  0.9934249  2.5078822 0.0000000
Luminal A-Basal-like       5.65741206  5.0930071  6.2218170 0.0000000
Luminal B-Basal-like       5.75393648  5.1177917  6.3900812 0.0000000
Normal-like-Basal-like     3.19822560  1.4187830  4.9776682 0.0000127
Luminal A-HER2-enriched    3.90675849  3.2389456  4.5745714 0.0000000
Luminal B-HER2-enriched    4.00328292  3.2738302  4.7327357 0.0000000
Normal-like-HER2-enriched  1.44757204 -0.3673211  3.2624651 0.1869447
Luminal B-Luminal A        0.09652443 -0.4300295  0.6230783 0.9870680
Normal-like-Luminal A     -2.45918645 -4.2024577 -0.7159152 0.0012304
Normal-like-Luminal B     -2.55571088 -4.3235123 -0.7879094 0.0008469
```

We can see that ESR is not differentially expressed between the subtypes Normal-like vs HER2 enriched and Luminal A vs Luminal B (adjusted p-value > 0.05), and differentially expressed between any other pairs of subtypes.

```
> TukeyHSD(aov(pgr.values ~ subtypes), conf.level=0.95)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = pgr.values ~ subtypes)

$subtypes
                                 diff        lwr        upr     p adj
HER2-enriched-Basal-like   0.5142575 -0.4755584  1.5040735 0.6120797
Luminal A-Basal-like       4.0351468  3.2973816  4.7729120 0.0000000
Luminal B-Basal-like       2.7439859  1.9124454  3.5755263 0.0000000
Normal-like-Basal-like     1.5642702 -0.7617387  3.8902791 0.3498935
Luminal A-HER2-enriched    3.5208893  2.6479537  4.3938249 0.0000000
Luminal B-HER2-enriched    2.2297283  1.2762198  3.1832368 0.0000000
Normal-like-HER2-enriched  1.0500127 -1.3223356  3.4223610 0.7433985
Luminal B-Luminal A       -1.2911610 -1.9794490 -0.6028729 0.0000045
Normal-like-Luminal A     -2.4708766 -4.7496040 -0.1921493 0.0260110
Normal-like-Luminal B     -1.1797156 -3.4905078  1.1310765 0.6280621
```

For the PRG gene, it is not differentially expressed between the subtypes:

- HER2 enriched vs basal like
- Normal-like vs basal like
- Normal-like vs HER2 enriched
- Normal-like-Luminal B

and differentially expressed between any other pairs of subtypes.

```
> TukeyHSD(aov(her2.values ~ subtypes), conf.level=0.95)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = her2.values ~ subtypes)

$subtypes
                                  diff        lwr        upr     p adj
HER2-enriched-Basal-like     2.45650803  1.9366294  2.9763867 0.0000000
Luminal A-Basal-like         0.84837485  0.4608802  1.2358695 0.0000000
Luminal B-Basal-like         0.76676001  0.3300120  1.2035080 0.0000218
Normal-like-Basal-like      -0.08133799 -1.3030222  1.1403462 0.9997507
Luminal A-HER2-enriched     -1.60813318 -2.0666231 -1.1496433 0.0000000
Luminal B-HER2-enriched     -1.68974802 -2.1905570 -1.1889390 0.0000000
Normal-like-HER2-enriched   -2.53784603 -3.7838689 -1.2918231 0.0000005
Luminal B-Luminal A         -0.08161484 -0.4431227  0.2798931 0.9719860
Normal-like-Luminal A       -0.92971285 -2.1265634  0.2671378 0.2096378
Normal-like-Luminal B       -0.84809801 -2.0617899  0.3655939 0.3105925
```

For the ERBB2 gene, it is not differentially expressed between the subtypes:

- Normal-like vs basal like
- Luminal B vs Luminal A
- Normal-like-Luminal A
- Normal-like-Luminal B

and differentially expressed between any other pairs of subtypes.

(d) Which intrinsic subtype is triple negative? Which intrinsic subtypes are positive for ESR1 and PGR? [1pt]
Triple negative is characterized by a lack of expression of ESR1, HER2 and PGR. According to the boxplots, the intrinsic subtype is 'Basal-like', which is also concordant with the Tukey's tests. Subtypes positive for ESR1 and PGR are Luminal A and Luminal B (for the latter, the expression of PGR is differentially expressed between these subtypes and basal-like).

(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]

```
> table(subtypes, er.status)

              er.status
subtypes          0   1
  Basal-like     55   9
  HER2-enriched  20  20
  Luminal A       7 145
  Luminal B       0  81
  Normal-like     1   4
```

```
> fisher.test(subtypes, er.status)

        Fisher's Exact Test for Count Data

data:  subtypes and er.status
p-value < 2.2e-16
alternative hypothesis: two.sided
```

According to Remark 10 in the Lectures, since we have few counts less than 1, and for a total number of cells equals to 10, we have 3 cells with counts less than 5 (i.e. 30% of the cells). It is therefore appropriate to use a Fisher's exact test. This test indicates that we can reject the null hypothesis that the ESR1 status and the subtype are independent. Therefore, there is a significant association between these two categorical variables.
the table `table(subtypes, er.status)` above indicates that positive ER status tumours are more likely Luminal A or Luminal B subtypes, which was also indicated in the Tukey's test for ESR1 (highest adjusted p-value).

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:

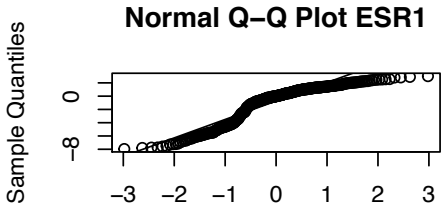
(a) Create a factor variable called '`diff` with levels corresponding to tumours classified as either Basal-like or Luminal A (Hint: Use the `PAM50.mRNA` phenotypic data). Use the `summary()` function to determine how many tumours fall within each tumour subtype. [1pt]

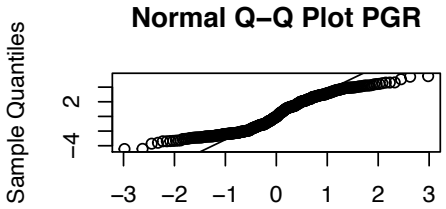
```
> diff <- factor(eSet$PAM50.mRNA, levels=c("Basal-like", "Luminal A"))
> summary(diff)

Basal-like  Luminal A       NA's
        66        154        128
```

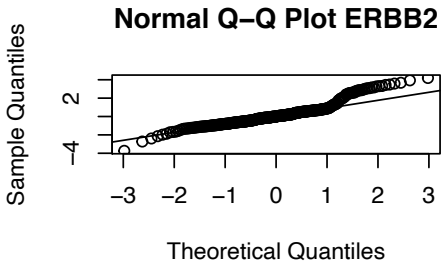
We have 66 basal-like tumours and 154 Luminal A tumours (the rest os classified as 'Not Applicable').

(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]

```
> par(mfrow=c(2,2))
> qqnorm(data2["ESR1",],main = "Normal Q-Q Plot ESR1")
> qqline(data2["ESR1",])
> qqnorm(data2["PGR",],main = "Normal Q-Q Plot PGR")
> qqline(data2["PGR",])
> qqnorm(data2["ERBB2",], main = "Normal Q-Q Plot ERBB2")
> qqline(data2["ERBB2",])
> par(mfrow=c(1,1))
```







The Q-Q plots indicate that there is a number of genes which do not have a normal distribution (low and high tails).

We use a Shapiro test to test whether the genes have a normal distribution for each subtype and given the high number of genes, we apply a Benjamini Hochberg multiple testing correction on the p-values:

```
> #define a new function
> shapiro.function = function(x){shapiro.test(x)$p.value}
> # for luminal A subtype
> pval.shapiro.lumA <- apply(data2[,diff=="Luminal A"], 1, shapiro.function)
> #apply multiple testing correction
> pval.shapiro.lumA.adj = p.adjust(pval.shapiro.lumA, method = 'BH')
> sum(pval.shapiro.lumA.adj > 0.05)
[1] 3188
> # for Basal-like subtype
> pval.shapiro.basal <- apply(data2[,diff=="Basal-like"], 1, shapiro.function)
> #apply multiple testing correction
> pval.shapiro.basal.adj = p.adjust(pval.shapiro.basal, method = 'BH')
> sum(pval.shapiro.basal.adj > 0.05)
[1] 4466
```

In the Luminal A subtype, we have 3188 genes which do not have a normal distribution, and for basal-like tumour, we have 4466 genes which do not have a normal distribution.

We next test the equality of variance for each gene between the two subtypes:

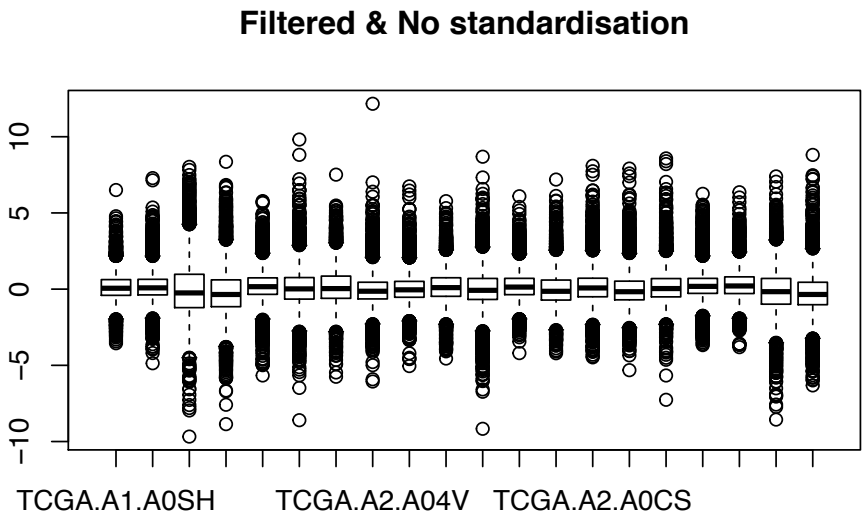
```
> var.function <- function(x, diff){
+   var.test(x[diff=="Basal-like"], x[diff=="Luminal A"], alternative='two.sided')$p.value
+ }
> pval.vartest <- apply(data2, 1, var.function, diff=diff)
> #apply multiple testing correction
> pval.vartest.adj = p.adjust(pval.vartest, method = 'BH')
> sum(pval.vartest.adj < 0.05)
```

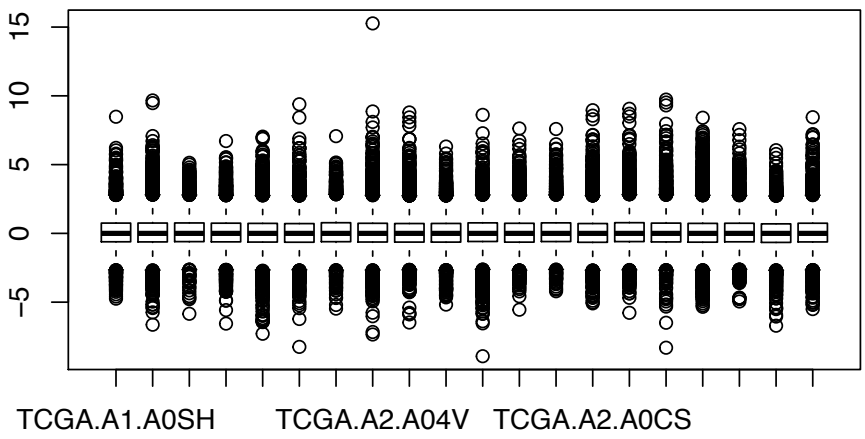
```
[1] 3000
```

There are 3000 genes for which we reject the null hypothesis that their variance between the two subtypes are equal.

(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]

```
> mads <- apply(data2, 2, mad)
> median <- apply(data2, 2, median)
> data.m <- sweep(data2, 2, median)
> data.standardised <- sweep(data.m, 2, mads, FUN="/")
> par(mfrow=c(2,1))
> boxplot(data2[,1:20], main="Filtered & No standardisation")
> boxplot(data.standardised[,1:20], main="Filtered & Standardised: Column Median/Mad")
> par(mfrow=c(1,1))
```





These boxplot show the effect of the standardisation, all the box and whiskers plots are now well aligned (same variance and same median for each tumour).

```
> # for Luminal A subtype
> pval.shapiro.lumA2 <- apply(data.standardised[,diff=="Luminal A"], 1, shapiro.function)
> #apply multiple testing correction
> pval.shapiro.lumA.adj2 = p.adjust(pval.shapiro.lumA2, method = 'BH')
> sum(pval.shapiro.lumA.adj2 > 0.05)
[1] 3232
> # for Basal-like subtype
> pval.shapiro.basal2 <- apply(data.standardised[,diff=="Basal-like"], 1, shapiro.function)
> #apply multiple testing correction
```

```
> pval.shapiro.basal.adj2 = p.adjust(pval.shapiro.basal2, method = 'BH')
> sum(pval.shapiro.basal.adj2 > 0.05)
```

```
[1] 4355
```

We now have for the Luminal A subtype 3232 (rpreviously 3188) genes which do not have a normal distribution, and for basal-like tumour, we have 4355 (previously 4466) genes which do not have a normal distribution.
What is important to assess though is whether there is a major proportion of these genes that are the same:

```
> # for Luminal A, we have in common:
> length(intersect(which(pval.shapiro.lumA.adj> 0.05), which(pval.shapiro.lumA.adj2> 0.05)))
```

```
[1] 2718
```

```
> # for basal like, we have in common:
> length(intersect(which(pval.shapiro.basal.adj> 0.05), which(pval.shapiro.basal.adj2> 0.05)))
```

```
[1] 4112
```

We next assess the equality of variance between the two subtypes:

```
> pval.vartest2 <- apply(data.standardised, 1, var.function, diff=diff)
> #apply multiple testing correction
> pval.vartest.adj2 = p.adjust(pval.vartest2, method = 'BH')
> sum(pval.vartest.adj2 < 0.05)
```

```
[1] 3248
```

There are 3248 (previously 3000) genes for which we reject the null hypothesis that their variance between the two subtypes are equal.

(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]
According to the results obtained previously, it is preferable to set the argument `var.equal = FALSE` in the following t-test:

```
> pval <- apply(data.standardised, 1, function(x) {t.test(x~diff,
+     alternative='two.sided', var.equal=FALSE)$p.value})
> sum(pval <= 0.01)
```

```
[1] 4460
```

There are 4460 genes that are differentially expressed with a 0.01 significance threshold (without applying a multiple testing correction).

(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]

```
> pval.adj <- p.adjust(pval, method = 'BH')
> names(pval.adj) <- rownames(data.standardised)
> sum(pval.adj <=0.01)
```

```
[1] 4350
```

After applying a multiple test correction, we have 4350 genes that are differentially expressed between the two subtypes.
For the two genes of interest, below are their adjusted p-values (highly significant):

```
> pval.adj["ESR1"]
        ESR1
4.194506e-46
> pval.adj["PGR"]
         PGR
2.667089e-42
```

We store the top 300 DE genes:

```
> top300 <- sort(pval.adj, decreasing=FALSE)[1:300]
> DE.genes <- names(top300)
```

(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]

```
> library(gplots)
> # data.300 <- data.standardised[DE.genes,]
> # hc <-hclust(dist(t(data.300)), method="ward")
> # hlg <-hclust(dist(data.300), method="ward")
> # patient.colors <- as.numeric(subtypes)
> # colors <- c("red", "green", "blue", "orange", "purple")
> # heatmap(data.300, Rowv=as.dendrogram(hlg), Colv=as.dendrogram(hc), scale='none',
> #         col=greenred(100), ColSideColors=colors[patient.colors])
>
```

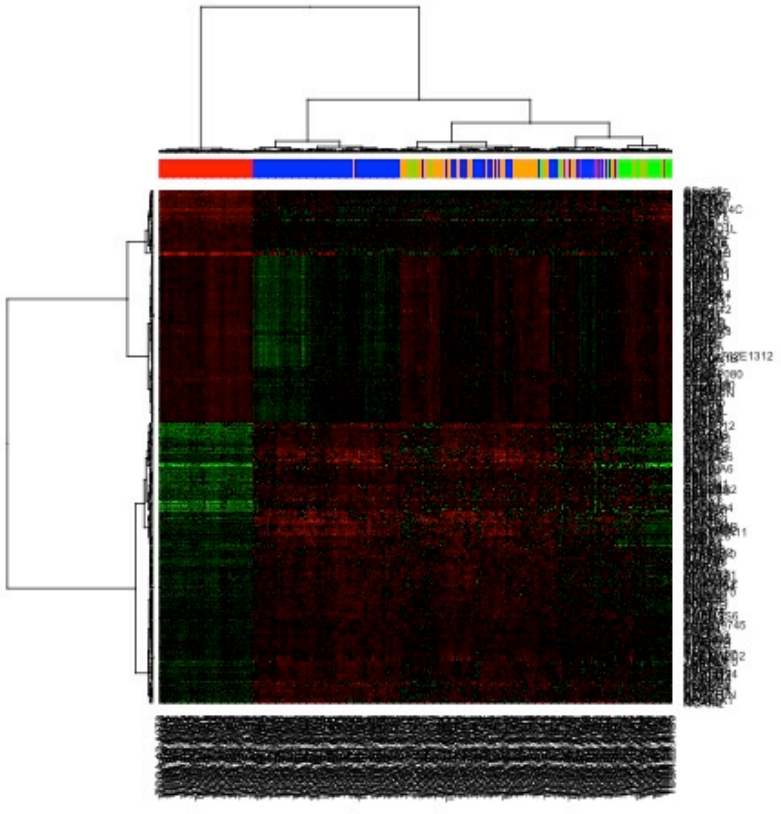


Figure 1: Hierarchical clustering of the 300 top DE genes with linkage method = Ward and Euclidian distance.

The heatmap shows that these genes are able to cluster the HER2-enriched subtype well (red color on top of the heatmap). Most of the Luminal A subtypes are clustered together (blue color), but some of them are merged with the Luminal B (green) and the Normal-like subtypes (orange). The Normal-like and the Luminal B subtypes seem quite distinct too.

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]

```
> sum(DE.genes %in% pam50.genelist)
[1] 24
> DE.genes[which(DE.genes %in% pam50.genelist)]
 [1] "NAT1"    "MELK"    "ANLN"    "KIF2C"   "ESR1"    "MLPH"    "SLC39A6"
 [8] "CDC20"   "CEP55"   "PGR"     "ORC6L"   "EXO1"    "CENPF"   "MAPT"
[15] "UBE2C"   "FOXC1"   "MKI67"   "PTTG1"   "MYBL2"   "BIRC5"   "CCNE1"
[22] "PHGDH"   "FOXA1"   "CDC6"
```

   (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]

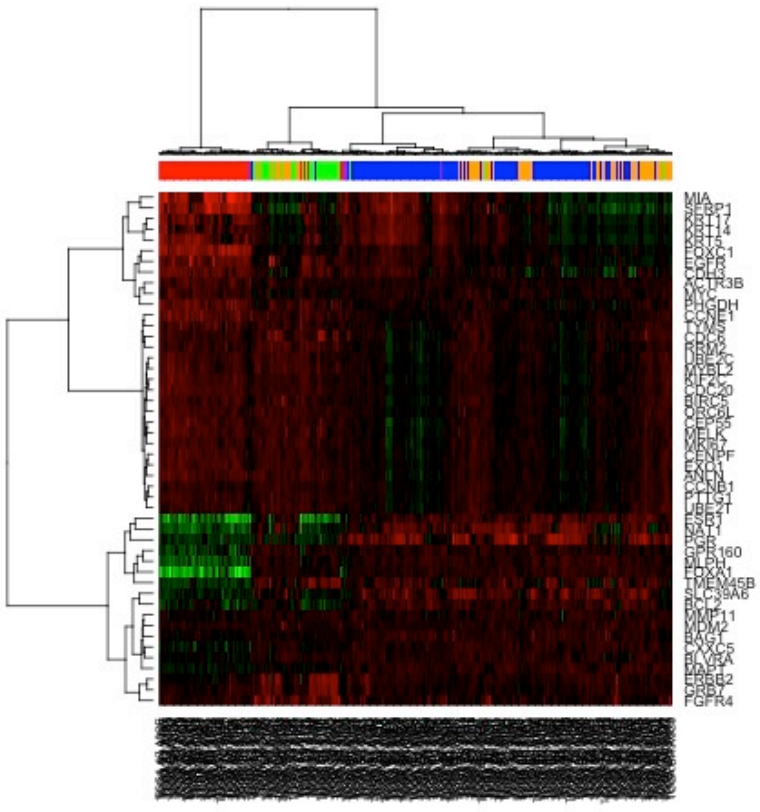


Figure 2: Hierarchical clustering of the pamr50 genes with linkage method = Ward and Euclidian distance.

We observe similar clustering for the HER2-enriched (red) subtype, which seems to be the easiest subtype to differentiate from the other subtypes. Differences arise for the cluster of the other subtypes: we observe more Luminal A tumours (blue) mixed with the Normal-like subtype (orange). The Luminal B (green) tumours are mixed with the some of the Normal-like tumours.