Curricular Unit of Computational Biology/ Bioinformatics
2019/20, Instituto Superior Técnico, Lisboa, Portugal

LAB 6 - APPLIED OMICS

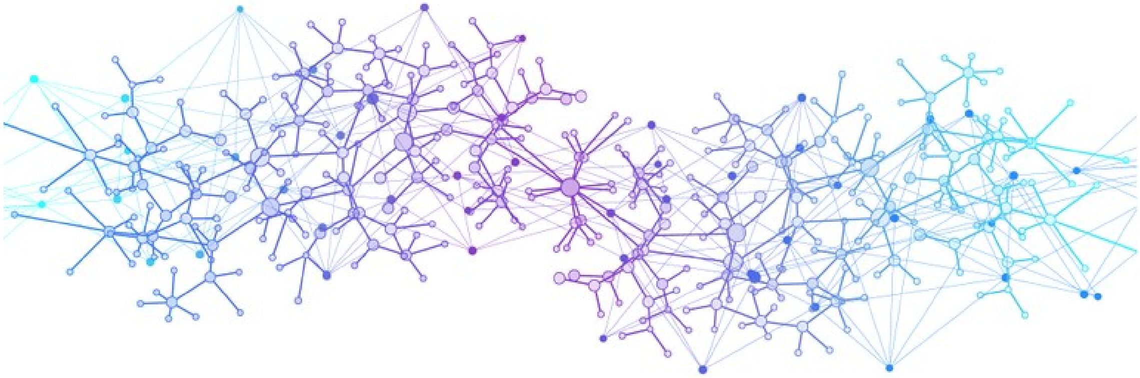
André Miranda (83811)[1], Diogo Ramalho (86407)[2], José Correia (83828)[3] and Manuel Madeira (83836)[4]

[1] *Instituto Superior Técnico,*
Integrated Masters in Biomedical Engineering
andre.f.miranda@tecnico.ulisboa.pt

[4] *Instituto Superior Técnico,*
MSc program in Computer Science and Engineering
diogo.ribeiro.ramalho@tecnico.ulisboa.pt

[2] *Instituto Superior Técnico,*
Integrated Masters in Biomedical Engineering
jose.r.c.correia@tecnico.ulisboa.pt

[3] *Instituto Superior Técnico,*
Integrated Masters in Biomedical Engineering
manuel.madeira@tecnico.ulisboa.pt



1 Group I

1.1 Exercise *(a)*

In order to perform an initial quality control validation of raw sequences that come straight of a high throughput sequencer, we resorted to the *FastQC* software, using the provided *fastq* files - TCGA-C8-A138-011.fastq (from now on called file 1) and TCGA-C8-A138-012.fastq (file 2) - as inputs. The results yielded are exposed on Fig 1.1.

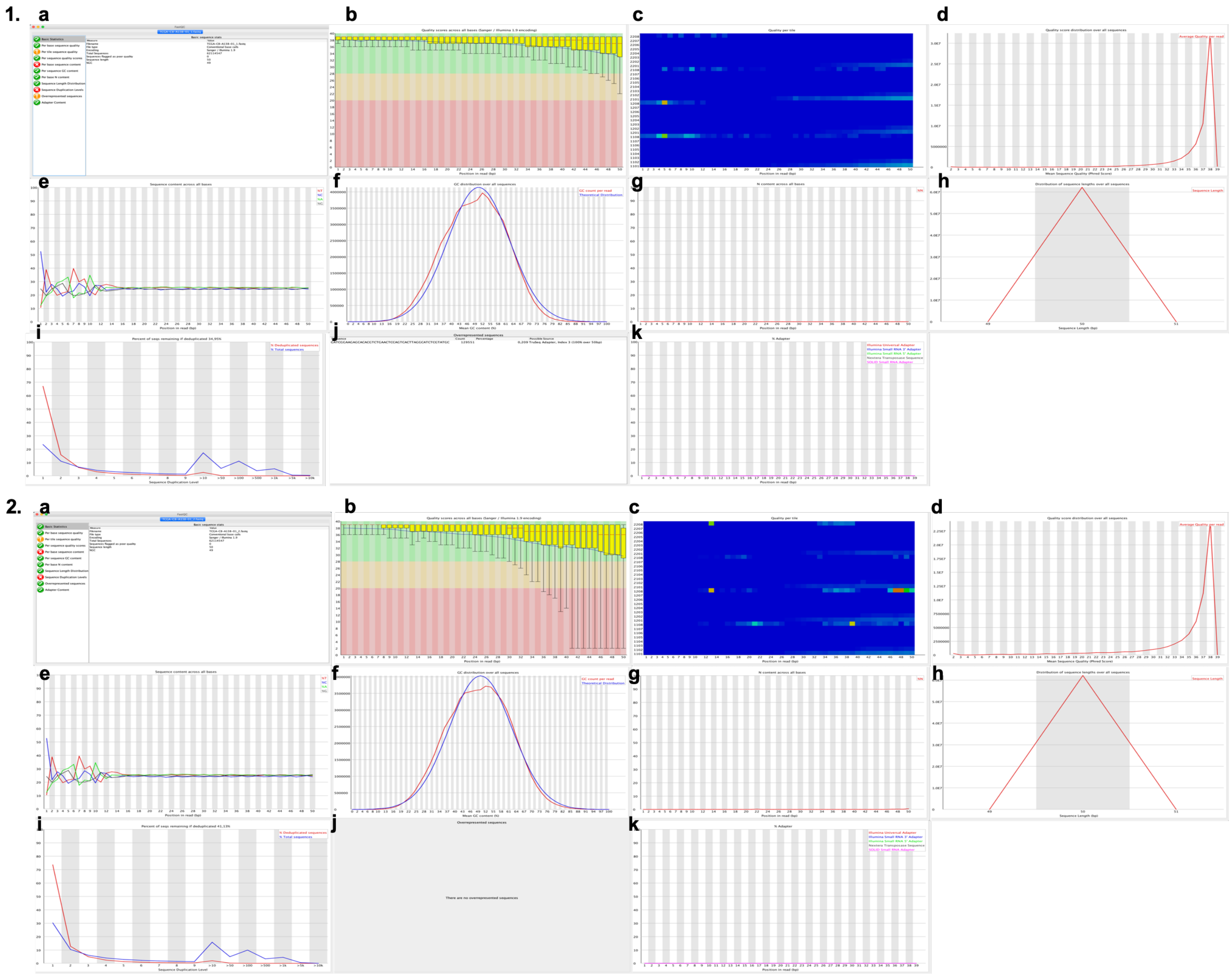


Figure 1.1: *FastQC* results obtained. **1** refers to TCGA-C8-A138-011.fastq results, while **2** the TCGA-C8-A138-012.fastq ones. The graphs shown are correspondent to: **a** - Basic sequence statistics; **b** - Quality per base sequence; **c** - Quality per tile ; **d** - Quality scores per sequence; **e** - Content per base sequence; **f** - GC content per sequence; **g** - N content per base; **h** - Sequence length distribution; **i** - Sequence duplication levels; **j** - Overrepresented sequences; **k** - Adapter content;

A quick look to both **a** graphs, allow us to confirm that both *fastq* files show similar basic stats, the only difference is referent to the exclamation mark that file 1 presents for the overrepresented sequences (while for file 2, this section is classified with a green mark). Generally speaking, both files raise concerns on the quality per tile sequence, content per base sequence and sequence suplication levels (graphs **c, e and i**, respectively).

Proceeding to the analysis of each of the remaining graphs:

- **Quality per base sequence - (b) graphs**: the quality information was encoded accordingly to Sanger/Illumina 1.9, and, for both files, admissible qualities are observable. Nevertheless, it might be stated that for file 2 the results are worse, inclusively with the more glaring feature being the 10^{th} percentiles for positions in reads higher than 40 reaching values way below 20 (which is usually considered to be the threshold separating good quality results from bad ones). Moreover, another interesting features which is present in both files is that, in general, the quality decreases as the position in the read increases. Both observations can be explained by the fact that the technique used to read the sequences is very efficient, but will always cause some deterioration for each read, as the clearance of the flowcell is not perfect at each ”iteration” of the process and so, it is natural that the quality of reads decreases along the position in sequence for each of the *fastq* files, as well as it decreases from the file 1 to file 2 as it already took the rebounds of the sequencing procedure of the former.

- **Quality per tile sequence - (c) graphs**: this graph allows the observation of the quality scores from a spatial point of view, as it shows the deviation from the average quality for each tile. The colours are on a cold to hot scale, with cold colours being positions where the quality was at or above the average for that base in the run, and hotter colours indicate that a tile had worse qualities than other tiles for that base. Therefore, a perfect plot should be blue all over. In our case, it is possible to observe some regions of poor quality (leading to the concern mark), with a mean Phred score less than 2 below the mean for that base across all tiles, but not more less than 5 (as that would raise a failure mark). Some common reasons for problems at this level are the presence of bubbles, smudges or debris in the flowcell or as a result of the flowcell being overloaded. An important aspect to be pointed out is the fact that a bad quality read does not have a great negative impact in the workflow, as the probability of that wrong read sequence aligning with sequences of interest is really small and so it simply won’t align. Nevertheless, if a significative part of the flowcell lead to bad quality reads, by knowing the spatial information about where they came from it is possible to filter out those reads, avoiding the deterioration of the results, as the ”concentration” in the remaining area of all the sequences is constant (they’re randomly spread by the flowcell).

- **Quality scores per sequence - (d) graphs** - it is basically the observation of the (b) graph from the right side, it is possible to observe that the distribution of the average quality per read is quite high for both files (strongly deviated to the right). This weird shape of distribution (*a priori* it wouldn’t be expected to have a probabilistic distribution of this shape), in fact, is a result of the scale used in x-axis, as well as the scale having a technologically imposed upper limit (around 40 to 42): Phred Score, which log-scales the aforementioned probabilistic distribution (of base-calling error, based on intensity of cluster, signal-to-noise ratio, etc...). Again, it is still possible to observe that the left tail of the distribution extends itself more significantly on the file 2 analysis, reflecting its lower quality in comparison to the file 1 results.

- **Content per base sequence - (e) graphs**: in this graphs, parallel lines are supposed to be observed as a result of the fact that for a diverse library, the position that we’re looking at doesn’t influence the base call. In our files, it is possible to observe that behaviour for positions in read higher than 16bp. This is due to the fact that RNA-seq often shows biased composition in the first bases of cDNA due to chemical preferences of the ’random’ sequences used to break up the library, not actually being all that random, leading to the concern mark obtained for this section. This is in fact very typical of Illumina technology and so it is not an error by itself. After these initial positions, it is possible to observe that the lines remain more or less parallel with %A and %T slightly above 25%, while %G and %C are slightly below. This is a consequence of the fact that the genomes evolved in such a way that usually regions that have more AT content are usually coding regions, while regions with more CG content are regulatory ones (intergenic regions, introns - *e.g.* CpG islands). As we are working with RNA, which is comes from coding regions, the bias towards a bigger AT content than CG can be used as good quality control check.

- **GC content per sequence - (f) graphs**: is is possible to observe that both distributions are more or less similar. The theoretical distribution is a normal one, centered around 50% (with mean and standard deviation that my real library has). Thus, it is possible to observe that both curves are shifted to the left as a consequence of the smaller %C and %G and higher %A and %T. Nevertheless, the presence of a small peak with one spike coming out above the 50% of GC content might be a consequence of the aforementioned biased priming, where 10 out of the first 16 positions show a preference by C or G (62.5%) nucleotides over A and T.

- **N content per base - (g) graphs**: allows us to understand if are there any uncalled/undetermined bases in our library. In our case, it is possible to conclude that they’re not significant.

- **Sequence length distribution - (h) graphs**: in our case, it is possible to observe that our library is all the same length - 50bp - for both files.

- **Sequence duplication levels - (i) graphs**: it is possible to observe similar profiles of duplication for both files and, in fact, those profiles show a concerning amount of duplicated sequences, namely from >10 to >5k sequence duplication levels. It is also possible to observe that if we deduplicated the library, the percentage of remaining sequences would be 34.95% for file 1 and 41.13% for file 2. All these suspicious results lead to concerning mark for this section, as non-unique sequences make up more than 50% of the total. Nevertheless, the fact that these bins have different sizes cause the unexpected behaviour for higher duplication levels. There are two possible reasons for excessive duplication: due to coverage or due to the amplification of contamination, and both lead to different duplication distributions. The former usually presents smoother distributions as it has to do with the distribution of expression of the genes, while the latter shows big spikes of duplication for higher duplication levels. Given this, despite all these concerns raised by the software, as we observe the case of the smooth distribution without sudden spike, we assign the observed behaviour to the coverage reason, not compromising the results for further analysis.

- **Overrepresented sequences - (j) graphs**: while for file 1 there is one overrepresented sequence, *i.e.*, there’s a sequence which represents more than 0.1% of the total sequences in the library, for file 2 that is not the case. That overrepresented sequence is GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGC, has 129551 counts (0.209% of the total sequences) and the software proposes that the source for this concern is a TruSeq Adapter, Index 3, which shows a 100% overlap with this sequence for all the 50bp, being a very credible hypothesis.

- **Adapter content - (k) graphs**: by observing the graphs, it is possible to observe that there's no adapter which is present in an excessively high amount and so no warning is issued, as it'd only happen if any sequence was present in more than 5% of all reads.

1.2 Exercise (b)

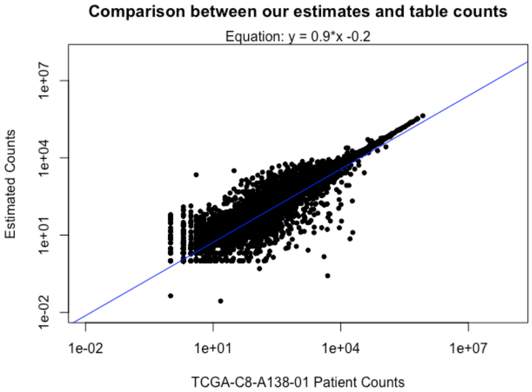


Figure 1.2: Comparison between the results obtained from the *fastq* files and the ones provided in the *TCGA BRCAGeneReadCounts.txt*. Each point represents a gene and while in the y-axis the results plotted are the ones obtained from the former, in the x-axis the results plotted are the ones obtained from the latter.

The chosen aligner was *kallisto*, which is a program that allows the quantification of transcript abundances from bulk and single-cell RNA-seq data. It takes advantage of the pseudoalignments for rapidly determining the compatibility of reads with targets, without the need for the whole alignment.

Therefore the first step taken was to index the human transcriptome. Here, we used the one provided by *ensemble.org*, *Homosapiens.GRCh38.cdna.all.fa*, as an input for *kallisto*, which is also able to perform this task. Afterwards, the output of this indexing procedure was used as an input, along with the *fastq* mentioned in the previous exercise, again in *kallisto* to perform the alignment itself, yielding a file with the number of reads for each *m*RNA belonging to human transcriptome. Finally, the match between those *m*RNAs and the respective human genes was already made using *R*, resorting to a file available on *BioMart*.

By observing Fig. 1.2, it is possible to conclude that the results yielded by the TCGA pipeline used are correlated with ours, specially for the higher expression region. In fact, the correlation between both pipelines for lower expression region is less robust, as it is a noisier region. Nevertheless, some differences are always expected as annotations and algorithms with different extents of conservativity can be used and so, for example, alignments against very rare exons might be considered (or not). Therefore, more counts for given genes are expected for more conservative approaches.

In our case, it is possible to observe that we usually consider higher numbers of counts for the same genes, which means that our approach was more conservative than the one adopted by the TCGA. As our pipeline aligns directly against the transcriptome while the TCGA approach aligns firstly against the genome and only afterwards passes to the transcriptome, it is more efficient at aligning reads that cover exon-intron reads, but, on the other hand, it loses reads in not annotated regions. These facts, result in the "equilibrium" observed, *i.e.*, high correlation between both approaches

2 Group II

2.1 Exercise (a)

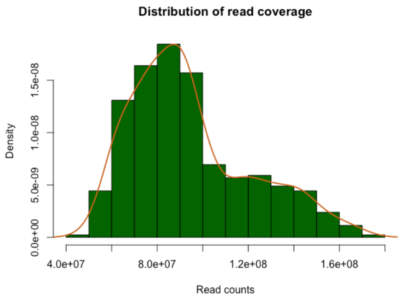


Figure 2.1: Distribution of the read coverage for the library used.

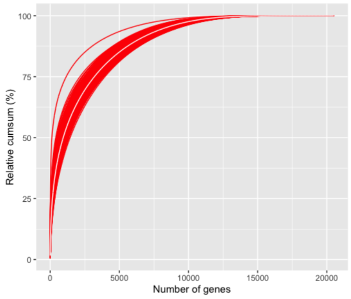


Figure 2.2: Cumulative distribution for all genes, for each sample. The curve in white is the mean cumulative distribution across all samples.

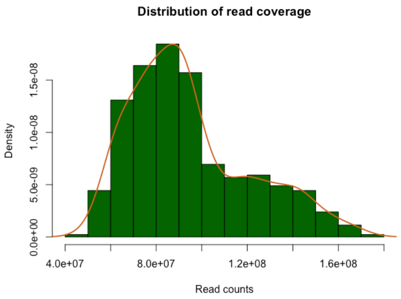


Figure 2.3: Number of distinct genes obtained when sampling (without replacement) the outlier sample and the mean sample.

A summary of the distributions of read coverage is observed in Fig. 2.1. It shows an approximate log-normal shape and it can be observed that there are very few samples with a huge or minimal number of read count, but actually they range in an interval of 180 million read counts approximately. Nevertheless, the majority of samples have a number of samples closer to the lower bound, presenting an amount read counts around 80 millions.

Regarding the library complexity, it was analysed using two approaches: in the first (Fig. 2.2), we observed the cumulative distribution for all genes for each sample in the dataset and in the second one we picked two samples of that dataset and registered the number of distinct genes obtained when sampling them.

In Fig. 2.2, it is possible to conclude that for all samples there is a small number of genes with the higher amount of read counts, while the remaining present a much lower number (the last 5000 genes seem that they almost do not have significant counts, as for every sample the curve is already stabilized after gene 15000 - and our total number of genes is 20502). Given this, the proportion between genes that have more and less read counts vary across samples and it is possible to detect an outlier regarding this aspect: there's one sample with an extremely small amount of genes with an extremely high number of read counts - the 1250 genes reach 75% of the read counts of that sample. We'll track this sample for next steps, referring to it as the maximal outlier. The mean sample is a virtual sample that simulates the mean of the cumulative distributions across all samples. This virtual sample allows us to have an idea of the different cumulative curves distribute themselves over the overloaded area.

Regarding Fig. 2.3, the samples chosen to this analysis are the maximum outlier and the sample that resorting to 125 genes had the minimal cumulative probability across all the samples, from now on called the minimal sample. In the figure it can be observed that for the same amount of sampled reads, the minimal sample always presents a higher amount of distinct genes obtained than the maximal one. This is due to the fact that the maximal sample has a smaller group of genes with a higher amount of read counts and so a high amount of being sampled, whereas with the minimum sample it goes the other way around. Therefore, the results obtained are accordingly to what was expected. Nevertheless, it should be pointed out that for a high enough number of sampled reads (*i.e* that would exhaust the great majority of reads of the most expressed genes of the maximal sample), the number of distinct genes would start to be higher for the maximal outlier, as from that point on all the genes would have a more similar probability of being sampled, while the minimal outlier would still be stuck to its more expressed genes, as they wouldn't still be exhausted at that point.

From the two analysis above (read coverage and library complexity), it is clear that a normalization is required across the different samples so that we can compare from a quantitatively point of view the gene expression among them, without being biased by factors as their different total number of reads (which is a result of the experimental setup having different qualities in data acquisitions), etc.

2.2 Exercise (b)

The aforementioned normalization was obtained resorting to *voom* (used in *R*). This was the normalization tool chosen since it allows us, generally speaking, the data ready for linear modelling. It normalizes the counts for each gene and each sample not only taking into account the total number of read counts for each sample when in comparison with the total read counts of the other samples but also by estimating the mean-variance relationship and using it to compute appropriate observational-level weights. It yields the counts per gene and per sample transformed onto log2-counts per million (logCPM) - it is not exactly a linear scaling transformation, but it is not a bad compromise.

In order to compare the effects of the normalization on each sample counts, box-plots were used. Figs. 2.4 and2.5 show the results obtained.

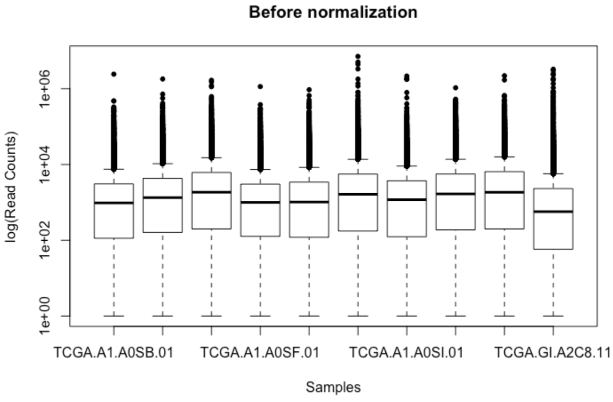


Figure 2.4: Box-plot obtained for 9 random samples and the maximal outlier before normalization

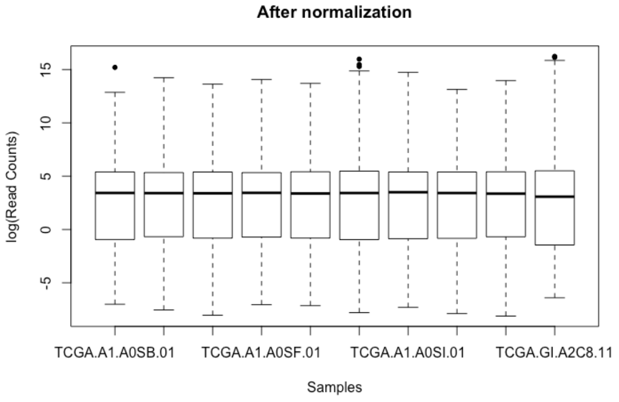


Figure 2.5: Box-plot obtained for 9 random samples and the maximal outlier after normalization

The differences before and after the normalization are clear: all the medians and quartiles are much more aligned after the normalization, showing that the normalization procedure was successful, since now it is possible to proceed to the comparison of expression profiles between samples.

Nevertheless, it is worth mentioning that the maximum outlier ("TCGA.GI.A2C8.11" - the rightmost box-plot in both Figures) presents an odd behavior in comparison to the other samples. Before normalization it is clear that it is the one with a lower median and with more point above it and even after the normalization it is the only showing a median slightly different (smaller) when in comparison with all the other. Its quartiles after normalization are also quite different from the obtained for other samples (like if they were "pushed-up"). We'll keep tracking this point so that we can conclude something about it.

2.3 Exercise (c)

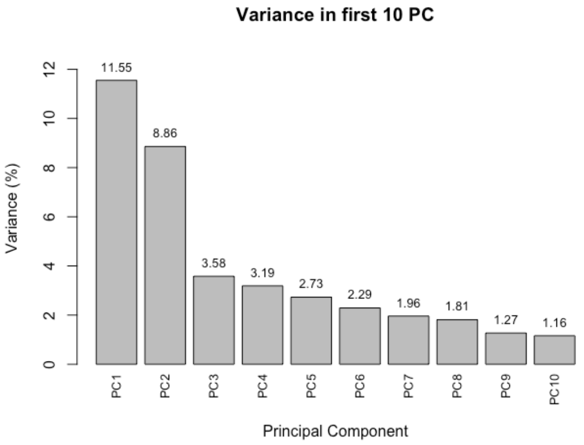


Figure 2.6: Percentage of variance explained by each principal component.

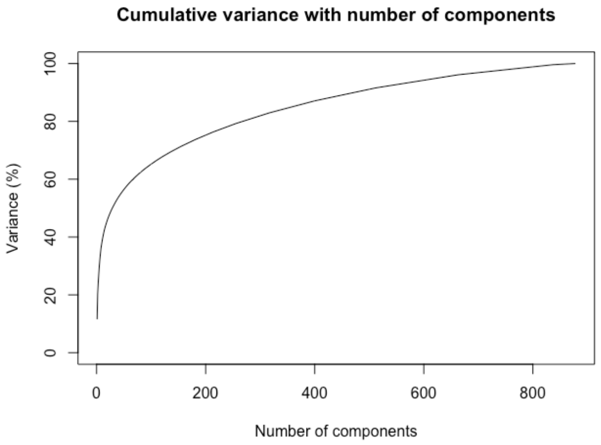


Figure 2.7: Cumulative variance through the number of components

Firstly, we can observe in Fig 2.6 that the 2 first Principal Components explain around 20% of the variance, which is a lot since there are 878 components. Also, in Fig 2.7 we can see that very few components explain a lot of the variance. Since the first 2 Principal components explain 20% of the variance, we expected to find clusters associated to some trait in both components. As we can see in Fig 2.8, the tumour samples are in the right side of the plot, explained by the 1st Principal Component. In Figs. 2.10 and 2.9, we can see that the samples that are positive for Estrogen and Progesterone receptors are in the upper side of the plot, explained by the 2nd Principal Component. We can also see through Fig. 2.11 there are clusters being formed by PAM50 molecular subtypes. We can also see that other traits such as "Read Coverage" or "Year of Diagnosis" do not form clusters. We couldn't find batch effects for other variables either Furthermore, we can see in Fig. 2.14 that the normalization indeed work, because the max outlier is not an outlier in this PCA dispersion graphic. In Fig. 2.15 we can see the 10 genes that better explain the variance in PC1 and in PC2. They all have a positive rotation, which means that these genes are better expressed in samples on the right side of the plot for PC1 and, on the upper side for PC2. As we can distinguish correctly clusters on the PC1 and PC2, we can conclude that there are no batch effects strong enough to influence the effects of biological traits.

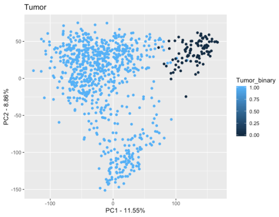


Figure 2.8: PCA analysis - Normal Vs Tumour

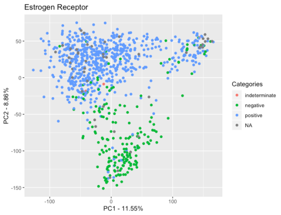


Figure 2.9: PCA analysis - Estrogen receptors

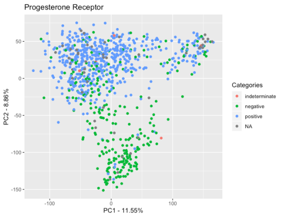


Figure 2.10: PCA analysis - Progesterone receptors

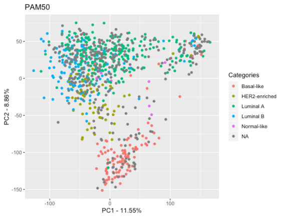


Figure 2.11: PCA analysis - PAM50

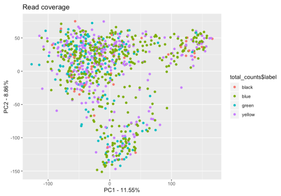


Figure 2.12: PCA analysis - Read Coverage

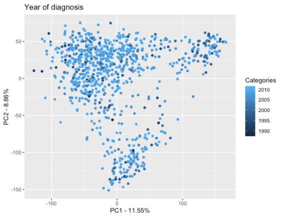


Figure 2.13: PCA analysis - Year of Diagnosis

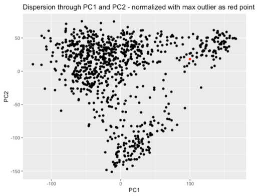
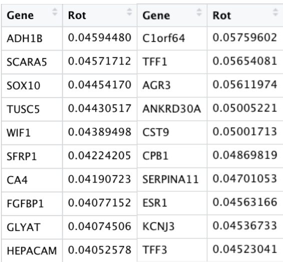


Figure 2.14: PCA analysis - maximal outlier in red, remaining samples in black.



Gene	Rot	Gene	Rot
ADH1B	0.04594480	C1orf64	0.05759602
SCARA5	0.04571712	TFF1	0.05654081
SOX10	0.04454170	AGR3	0.05611974
TUSC5	0.04430517	ANKRD30A	0.05005221
WIF1	0.04389498	CST9	0.05001713
SFRP1	0.04224205	CPB1	0.04869819
CA4	0.04190723	SERPINA11	0.04701053
FGFBP1	0.04077152	ESR1	0.04563166
GLYAT	0.04074506	KCNJ3	0.04536733
HEPACAM	0.04052578	TFF3	0.04523041

Figure 2.15: Fist two columns show the top 10 genes for the 1st principal component with the respective rotation, while the last two the top 10 genes and respective rotations for the 2nd principal component

2.4 Exercise (d)

What are the main differences in expressed genes and activated pathways between primary tumours and normal breast samples? How does the age of patients affect those differences?

In order to study the main differences in expressed genes between primary tumours (samples ending in ".01") and normal (".11") breast samples, the metastic (".06") samples were neglected. These results are exposed in Fig. 2.16. An analysis to study the interaction of the age with the previous discrimination was also carried out, with the results being presented Fig. 2.17.

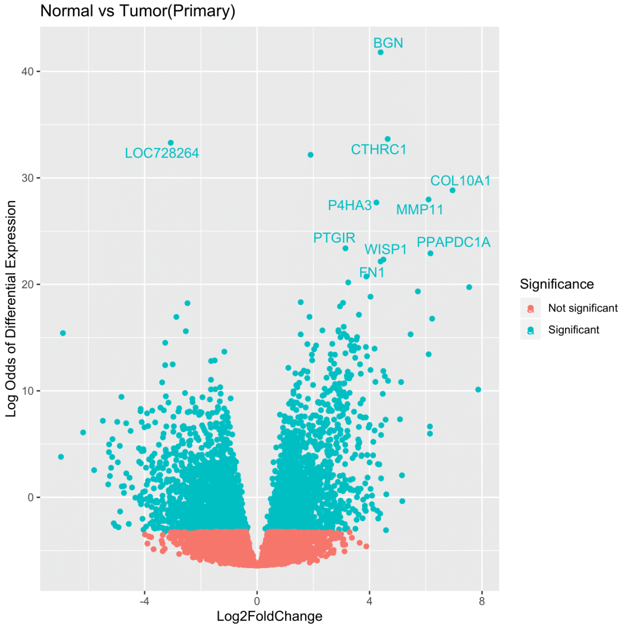


Figure 2.16: Volcano plot for the Normal *vs* Tumour

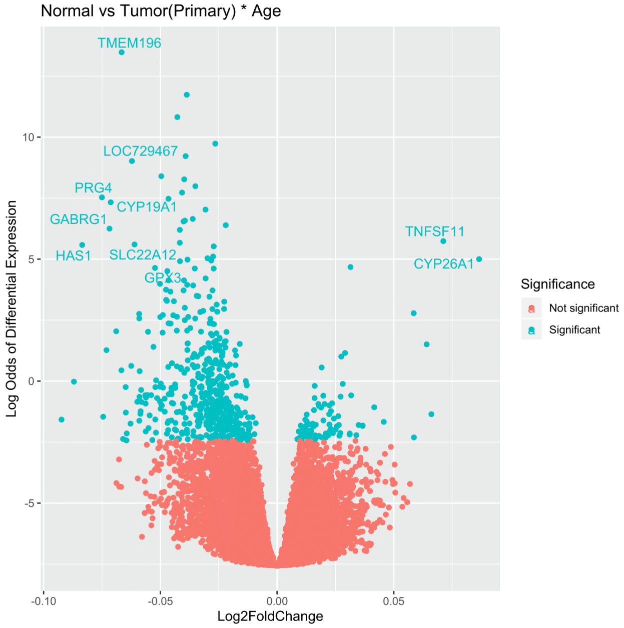


Figure 2.17: Volcano plot for the Normal *vs* Tumour

In Fig. 2.16, the red points refer to p-values higher than 0.05 (chosen arbitrarily), while the remaining one were colored with blue. In fact, this procedure was only carried out in order to allow the understanding that just as p-value, the B-statistic (the quantity represented in the y-axis) is also a measure of statistical significance, and so, a linear separation in the p-value also lead to a linear separation in the B-statistics.

As in this exercise we're looking for differentially expressed genes in normal *vs* tumour situation, the genes written in Fig. 2.16 are the ones that show the top ten t-statistics (in absolute-value), but that verify a minimimal Log$_2$ Fold Change bigger than 3 (in absolute value) - to ensure the variation of expression in both situations - and a maximum p-value of 0.05 - to ensure statistical significance. The t-statistics were chosen as the metric to choose the top 10 differentially expressed genes in this case as it is a good compromise between statistical relevance and differential significance.

For the interaction Fig. 2.17, the approach was the same: t-statistic as metric, respecting a maximal pvalue and minimal Log$_2$ Fold Change (in absolute value). It is possible to observe that the top 10 genes are now completely different form the previous situation. These genes are in fact genes that don't present a very high Log$_2$ Fold Change between the normal and tumoral situations, but in which that quantity is strongly enhanced or decreased (as we're in a positive or negative interaction case) with the age of the sample. In fact, in a plot of all the samples, with the age in the x-axis and the differential expression in y-axis for a given gene, in one of the top 10 genes mentioned in Fig. 2.16 we'd expect to see a big difference between the normal and tumoral samples constant along the age of the sample (*e.g* two parallel lines), while in the one of the top 10 genes mentioned in Fig. 2.17 it'd be expected to see that difference varying significantly with the age along the x-axis.

Next, a GSEA (Gene Set Enrichment Analysis) was performed to identify pathways that are differentially expressed on tumor and healthy cells. For this analysis, a p-value of 5% was chosen as the cutoff.
For such a threshold, the group obtained 24 metabolic pathways overexpressed in the tumor cells in relation to healthy cells and 7 for underexpressed pathways. For the interaction between patient age and differential gene activation, 0 pathways were obtained for a positive interaction (i.e. differential expression increases with age) but 24 were found for the negative interaction. The lack of pathways for a positive interaction is not entirely unexpected given the corresponding Volcano plot (Fig 2.17). Indeed there is a very small amount of genes with both fold changes higher than 1 (0 in log scale) and a significant Log Odds of differential expression. This can be visually identified as a lack of points in the right upper quadrant of the plot.

Some identified pathways are worth mentioning since they are more commonly associated with tumor cells. The pathways related with cell cycle (KEGG_CELL_CYCLE) and DNA replication (KEGG_DNA_REPLICATION) are overexpressed, being evidence that the tumor cells replicate more than healthy cells. Oxydative phosphorylation is also overexpressed (KEGG_OXIDATIVE_PHOSPHORYLATION since the tumor cells' fast growth needs a larger amount of energy.

There are some pathways which were identified both in the list of metabolic pathways overexpressed in the tumor cells in relation to healthy cells and in the list containing a negative interaction between patient age and differential gene activation. In other words, these are pathways that are overexpressed in tumor cells but whose expression generally diminishes with age. They are present in Table 2.1.

The fact that only 5 in 24 genes belong to both lists supports the idea that the overexpression of certain pathways in tumor cells are not attributable to age differences alone, but rather to other biological effects.

Table 2.1: Pathways present in both cases

KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION
KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS
KEGG_GRAFT_VERSUS_HOST_DISEASE
KEGG_LEISHMANIA_INFECTION

3 Group III

3.1 Exercise (a)

To assess the ability of the cognate genes' mRNA expression at recapitulating the binary classifications that result from the immunohistochemistry-based tests, two measures were implemented: the Area Under the Curve (AUC) for the ROC plot and the minimal distance from the ROC curve to the point corresponding to the best combination of sensitivity and specificity (true positive rate = 1 and false positive rate = 0).

The four ROC curves corresponding to the four possible combinations (estrogen receptor immunohistochemistry test label and ESR1 gene expression level, estrogen receptor immunohistochemistry test label and ESR2 gene expression level, progesteron receptor immunohistochemistry test label and PGR gene expression level and HER2 immunohistochemistry test label and ERBB2 gene expression level) are present in Figure 3.1.

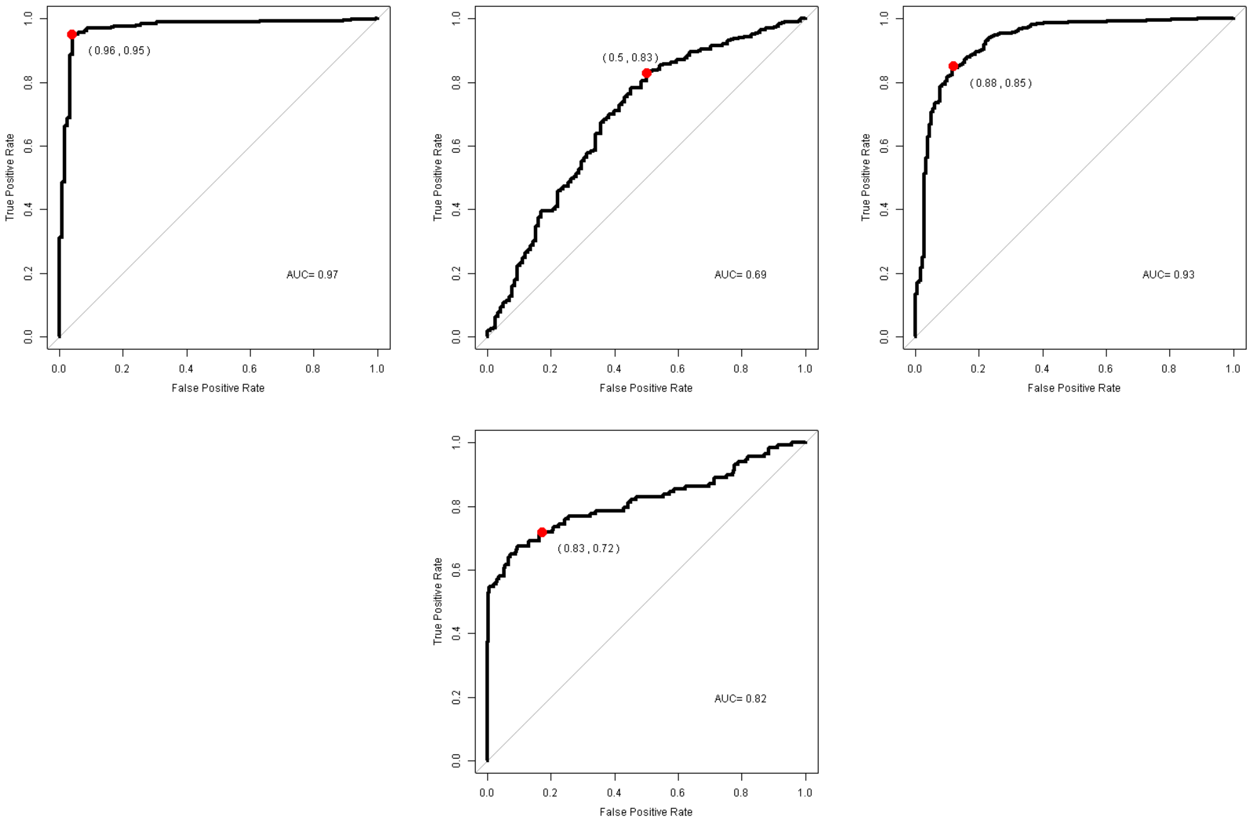


Figure 3.1: ROC curves, respective areas and chosen cutoff points. From left to right and top to bottom: Estrogen 1, Estrogen 2, Progesteron and HER2

The ROC curve represents the trade-off between sensitivity (or True Positive Rate) and specificity (or 1 – False Positive Rate). Classifiers with ROC curves closer to the top-left corner are usually indicative of a better performance. Quantitatively, better curves are associated with an AUC closer to 1 and a smaller minimal distance.

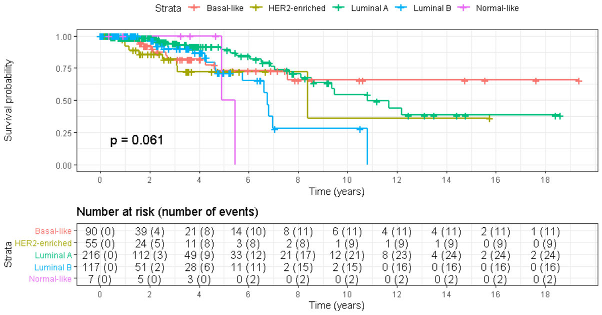
Comparing the two Estrogen ROC curves, it can be said that the ERS1 gene expression level seems much more accurate for estrogen receptor immunohistochemistry test label classification. This is quantitatively supported by the AUC values (AUC for ERS1 is 0.97 much higher than the AUC for ERS2 which is 0.69) and minimal distance from the ROC curve to the point corresponding to the best combination of sensitivity and specificity (0.064 for ERS1 which is much smaller than 0.528 for the ERS2).

In general, the ERS1 and PGR genes expression are good at recapitulating the result from the immunohistochemistry test since they have a AUC higher than 90%. The HER2 has a fairly reasonable AUC of 82% and ERS2 has a bad AUC of 69%.

3.2 Exercise (b)

Breast cancer may be classified into 5 different subtypes. This classification is done based on 50 different genes (PAM50 signature) and includes the classes: Basal-like, HER2-enriched, LuminalA, LuminalB and Normal-like. The patients are classified into one of these groups and then followed-up until they die or drop-out of the study.

The survival curves for the entire study duration are present in Figure 3.2.

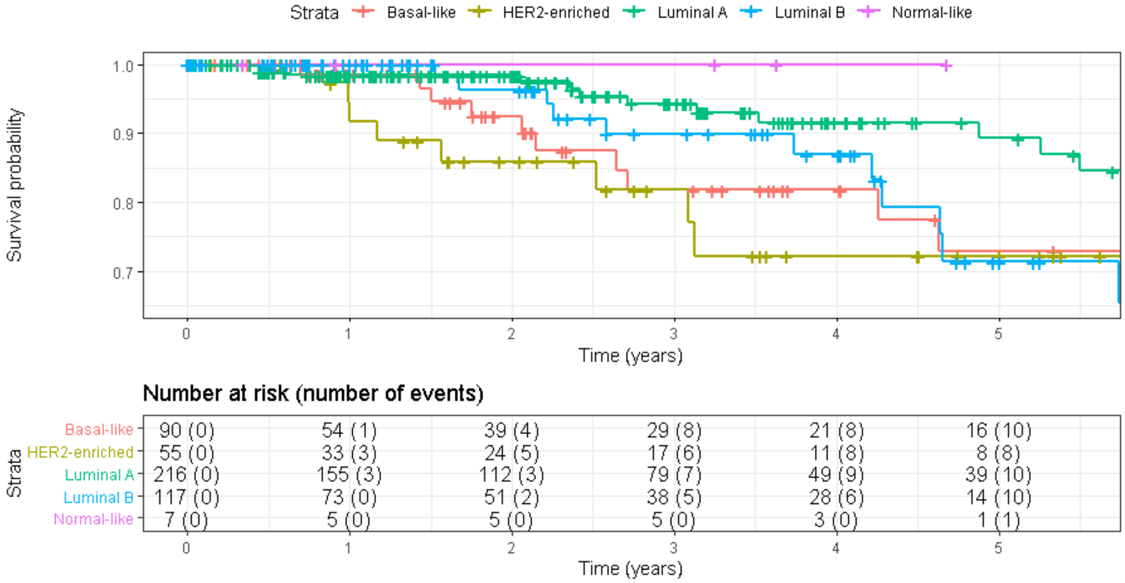


Strata	0	2	4	6	8	10	12	14	16	18
Basal-like	90 (0)	39 (4)	21 (8)	14 (10)	8 (11)	6 (11)	4 (11)	4 (11)	2 (11)	1 (11)
HER2-enriched	55 (0)	24 (5)	11 (8)	3 (8)	2 (8)	1 (9)	1 (9)	1 (9)	0 (9)	0 (9)
Luminal A	216 (0)	112 (3)	49 (9)	33 (12)	21 (17)	12 (21)	8 (23)	4 (24)	2 (24)	2 (24)
Luminal B	117 (0)	51 (2)	28 (6)	11 (11)	2 (15)	2 (15)	0 (16)	0 (16)	0 (16)	0 (16)
Normal-like	7 (0)	5 (0)	3 (0)	0 (2)	0 (2)	0 (2)	0 (2)	0 (2)	0 (2)	0 (2)

Figure 3.2: Survival analysis curves for all tumor subtypes for the entire study duration

The analysis will focus on the first 5 years since from that point onwards, the number of patients still on the study is not representative of the cancer patient population and therefore, results are not conclusive. Hence, the analysis will focus on the first 5 years of the study as shown in Figure 3.3.

In particular, the study starts with very few subjects classified as $Normal - like$. At year 5, most patients (5) are censored (drop-out of the study) leaving only two who die at around year 5, setting the survival probability at zero. Although being the first class (and one of the only two) to reach zero survival probability, it has, actually, the best survival rate until year 4 (zero deaths). This is obviously not representative of the whole $Normal - like$ breast cancer patients population due to the already mentioned low number of patients in this class.



Strata	0	1	2	3	4	5
Basal-like	90 (0)	54 (1)	39 (4)	29 (8)	21 (8)	16 (10)
HER2-enriched	55 (0)	33 (3)	24 (5)	17 (6)	11 (8)	8 (8)
Luminal A	216 (0)	155 (3)	112 (3)	79 (7)	49 (9)	39 (10)
Luminal B	117 (0)	73 (0)	51 (2)	38 (5)	28 (6)	14 (10)
Normal-like	7 (0)	5 (0)	5 (0)	5 (0)	3 (0)	1 (1)

Figure 3.3: Survival analysis curves for all tumor subtypes for a 5 year window

With a qualitative visual analysis, the remaining four classes can be ranked by prognosis during the first 4 years of the study. Listing the classes from the best to the worst prognosis we have: $LimunalA$, $LuminalB$, $Basal - like$, $HER2 - enriched$.

Table 3.1: Survival rate (in percentage) for each of the first 5 years and subtype of breast cancer

	Time(years)				
	1	**2**	**3**	**4**	**5**
Basal	98	91	78	72	62
HER2	92	83	74	58	50
LA	98	97	92	84	80
LB	100	96	88	82	58
Normal	100	100	100	100	50

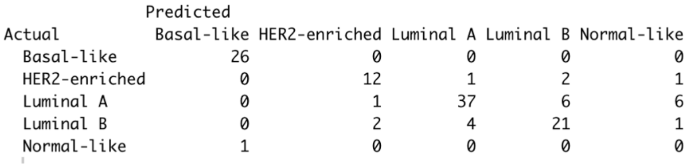
As a more quantitative measure, the survival rates were calculated (Table 3.1), allowing us to compare the prognosis at certain time points in the study. Two chosen time points were the year 2 and 4.

Considering the survival analysis for year 2 the ordered classes by prognosis (from best to worst) are: $Normal-like(100)$, $LuminalA(97)$, $LuminalB(96)$, $Basal-like(91)$, $Her2-enriched(83)$.
For the 4th year of study the order is the same: $Normal-like(100)$, $LuminalA(84)$, $LuminalB(82)$, $Basal-like(72)$, $Her2-enriched(58)$.

Both year analysis agree with the simple visual analysis.

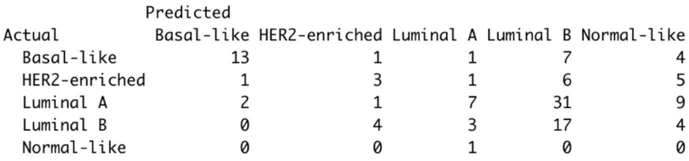
3.3 Exercise (c)

Firstly, we removed the samples without tumor. Then, we removed the samples that don't have a PAM50 result (NA value). Then, we selected the best genes using Lasso with a 10-fold cross validation. Our first resulting gene signature had 90 genes. We ran Lasso multiple times and saw that this process is stochastic, because the resulting gene signatures were different. In the first signature, none of the genes were in the PAM50 list. However, in one of the resulting gene signatures, there were four genes common among both lists. To compare the performance of our gene signature with the PAM50 genes, we ran a Naïve Bayes classifier on both. For this, we had to split the gene counts table in a training set (3/4 of data) and a testing set (1/4 of data). In order to evaluate the performance, we present the confusion matrices and the accuracies.



```
              Predicted
Actual         Basal-like HER2-enriched Luminal A Luminal B Normal-like
  Basal-like           26             0         0         0           0
  HER2-enriched         0            12         1         2           1
  Luminal A             0             1        37         6           6
  Luminal B             0             2         4        21           1
  Normal-like           1             0         0         0           0
```

Figure 3.4: Confusion Matrix for our gene signature



```
              Predicted
Actual         Basal-like HER2-enriched Luminal A Luminal B Normal-like
  Basal-like           13             1         1         7           4
  HER2-enriched         1             3         1         6           5
  Luminal A             2             1         7        31           9
  Luminal B             0             4         3        17           4
  Normal-like           0             0         1         0           0
```

Figure 3.5: Confusion Matrix for PAM50

PAM50: acc = 40/121 = 33%

Signature: acc = 96/121 = 79%

3.4 Exercise (d)

First, we had to remove the samples that already have a PAM50 molecular subtype assigned. Then, we used the Naïve Bayes model that was learned in the previous question by our gene signature to predict the PAM50 molecular subtype of each of these samples. In order to understand whether we were doing a good job we performed a survival analysis with the predicted data. As shown in Fig. 3.6, only 2 molecular subtypes were identified. Around 80% of them are "Basal-like", which is much higher than the results in b). However, as in b), "Basal-like" is still a better prognosis than "Luminal B".

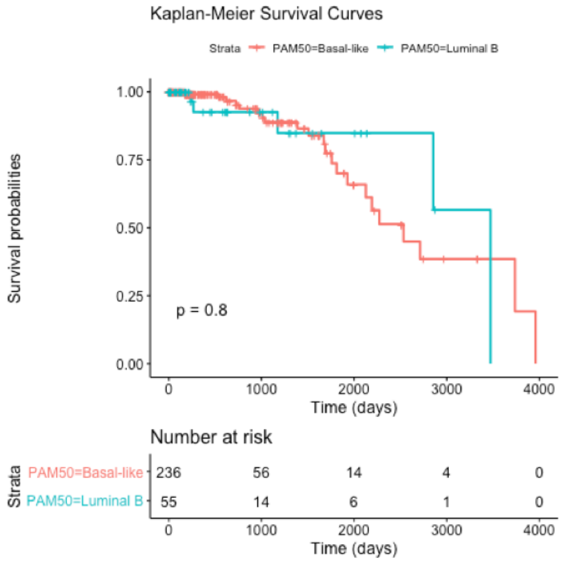


Figure 3.6: Survival analysis for data without PAM50 ground truth