Report for LAB6: Omics Applications
Bioinformatics

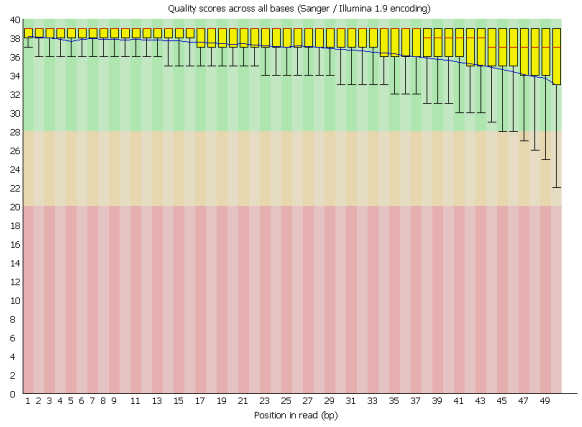
Ricardo Brancas 83557 | Margarida Ferreira 80832 | Felipe Gorostiaga 95383 | Benedict Schubert 95034

27th December 2019

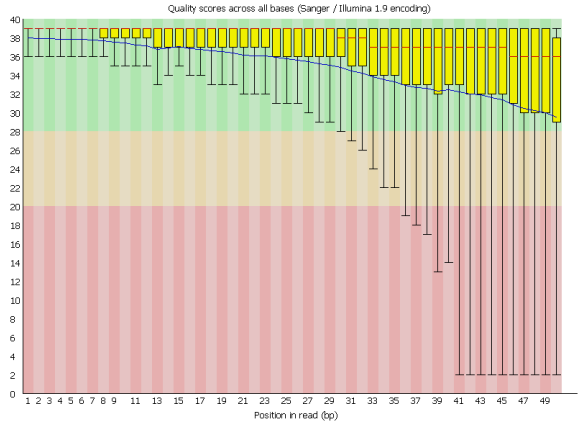
Group I.

I. a) Quality assessment of FASTQ

In figure 1 we present the per base sequence quality graphs obtained using the tool FastQC. The sequences have reasonably good quality scores, although there is some disparity between them. In particular the first sequence has slightly better quality than the second one, which is something we should be aware of.



(a) Quality graph for raw sequence 1.



(b) Quality graph for raw sequence 2.

Figure 1: Quality graphs for the raw sequences, as obtained in FastQC.

I. b) Estimated gene expression vs. provided read counts

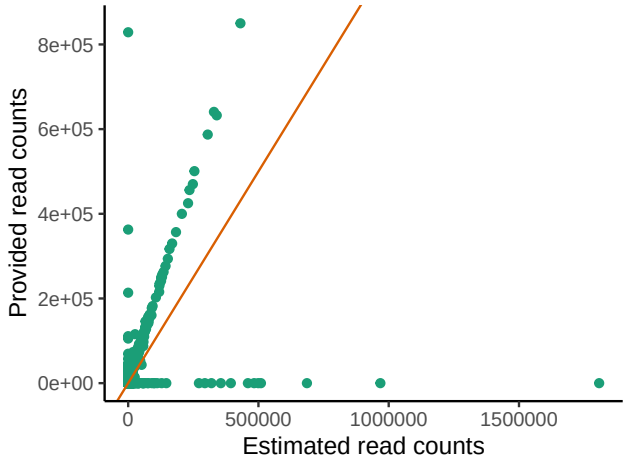
We chose to use the aligner Kallisto, together with the annotated human transcriptome from Ensembl [1]. We chose the cDNA sequences, as advised in the Kallisto FAQ [2].

To compare the transcript expression with the gene expression counts given, we downloaded a mapping from the Ensembl transcript identifiers to gene names. We used Ensembl Biomart [3] to get this mapping. We then took the transcript count estimates and the mapping and created a summarised table containing the total estimated number of reads for each gene. Finally, comparing this table with the read count table we were given, results in the graph in figure 2. Figure 2a shows all the estimated and provided read counts plotted against each other: each black mark represents a gene, the value on the y-axis is the provided read count, while the value on the x axis is the estimated read count; therefore, the closer a mark is to the $y = x$ line, the better the estimation. Analysing this graph, we can see that most genes are correctly estimated, as most marks lie within the $y = x$ line. Most outliers are from read counts of 0 which most likely represents a mismatch between the transcript identifiers and the gene names. In Figure 2b we show the same data, this time plotted in logarithmic scale so the values are more easily seen. Note that, in order to plot in a logarithmic scale, all the zero-valued read counts must be removed. In this plot we can again see that most of the estimates are accurate, even though there are a few outliers. We can also see that it appears that the higher the read count the better the estimation. This however is no true: what we are seeing is a result of the logarithmic scale: the errors are of the same order of magnitude, but since the scale is much larger, they seem smaller when compared to the actual read count values.

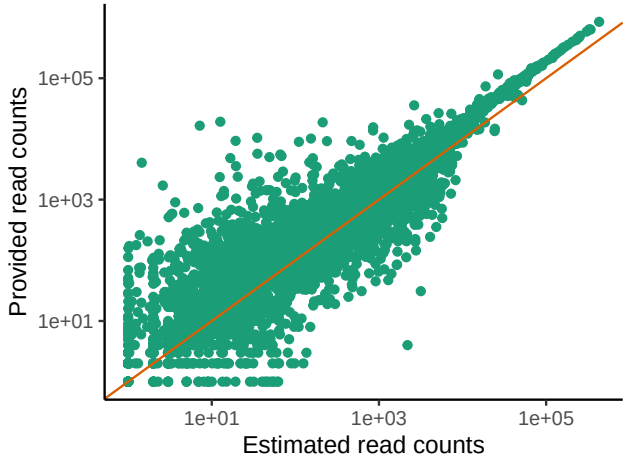
[1] `ftp://ftp.ensembl.org/pub/release-98/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz`

[2] `https://pachterlab.github.io/kallisto/faq`

[3] `http://www.ensembl.org/biomart/martview/a797838aa8255de1efa6fb6d11322eb5`



(a) Scatter plot of estimated vs. provided read counts.



(b) Log-log scatter plot of estimated vs. provided read counts after removing 0 values.

Figure 2: Comparison between the estimated counts and the provided values.

Group II.

II. a) Read coverage and library complexity

To analyse the read coverage, we created the histogram in figure 3. We can see that the count values range from roughly 40 million to 180 million. We can also see that most samples have more than 60 million reads, although a very small number has only around 40 million. This might mean that in these samples lowly expressed genes are below the detection threshold. We can also see that there are a lot more genes with read counts in the 60 to 100 million range than in the rest of the domain, which indicates that we will probably need some kind of normalisation of the data.

Library complexity refers to the number of unique DNA fragments present in a given library. To compute library complexity, we used a sampling-based approach.

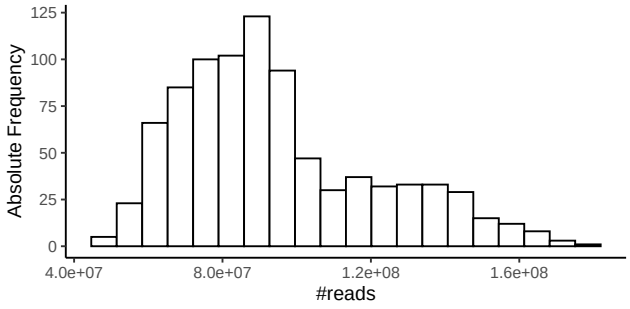


Figure 3: Read counts histogram.

For each tissue sample, we sampled the genes with probability proportional to the read count of each gene for that patient. In figure 4 we plot the library complexities for all samples. Analysing the figure we can see that, in general, in samples from normal tissues we hit more genes with fewer samples, which means that the genes are more uniformly represented. On the other hand, in deseased tissues, we get a lower percentage of all the detected genes with the same number of samples, showing that the read counts for the genes in deseased tissues is farther from a uniform distribution: there are a few genes with a much higher read count than the others. Finally, there is one outlier, a normal sample, for whom the gene coverage is a lot lower than that of the other samples, for the same number of random samples measured.

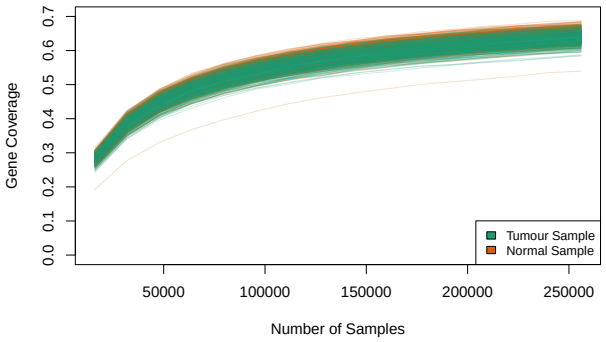


Figure 4: Library complexity.

II. b) Normalisation of the data

In order to make gene expression profiles comparable between samples, we normalised the data using the Trimmed Mean of M (TMM) method. Then, we produced a series of plots which help us visualise how the data changed after the normalisation.

First, we plotted the histogram of the normalised read counts, which can be seen in figure 5. We can see a few

differences in the distribution of normalised read counts when compared with that of the raw read counts (in figure 3). We can see that there is no longer such an accentuated "tail" on the right-side of the histogram (the higher read count values). While in the raw read counts the last bin was accounting for genes with a read count of around 180 million, in the normalised version we no longer have such high values: the last bin has read count of about 170 million read counts. Furthermore, we can identify in the normalised read counts a bimodal distribution: there are 2 peaks in the histogram, at around 75 million and 90 million normalised reads. These could not be seen in the original read count histogram where only one peak was discernible at around 85 million read counts.

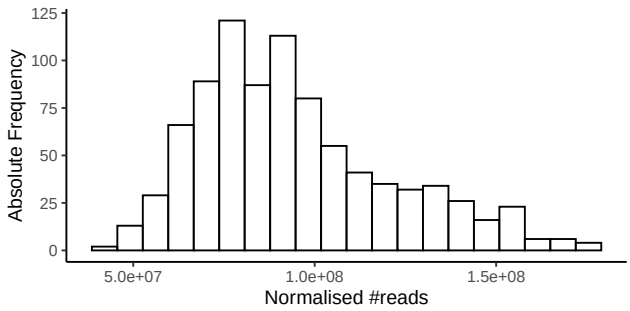
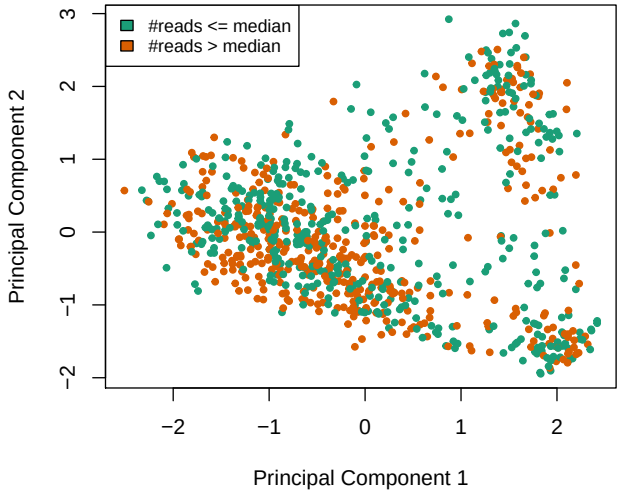


Figure 5: Normalised read counts histogram.

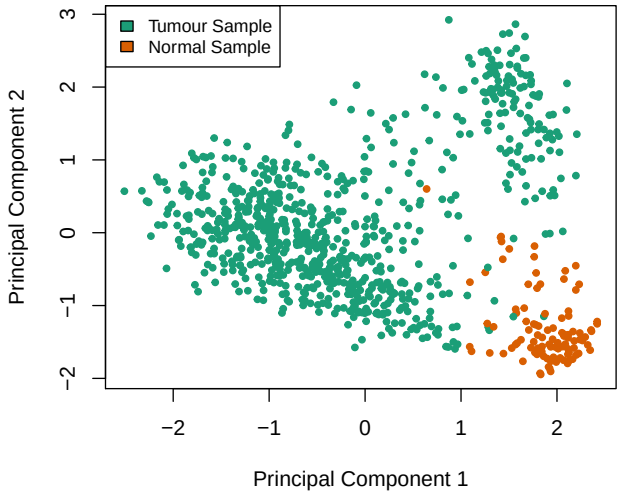
We wanted to ensure that after normalisation the number of read counts we had for a sample was not a factor when analysing how the read counts of each gene affect the samples. We want to consider only the ratio between genes' read counts, not the number of read counts per sample. To that end, we produced a Multi-Dimensional Scaling (MDS) plot where each mark corresponds to a sample, and the distance between samples corresponds to the leading Biological Coefficient of Variation (BCV), using the function `plotMDS`. In this plot, we coloured each mark according to the normalised read count of the respective sample, separating those that have a read count above the median (orange marks) from those that have it below the median (green marks). The result is shown in figure 6a. We can see that there is no apparent distinction between the green and orange marks, which shows our normalisation is successful: the read counts do not contribute to the main axes of variance.

Finally, we produced another MDS plot, this time to show the separation between normal and diseased tissue: we coloured the marks according to that classification. The result can be seen in figure 6b. There is a clear separation between the samples of normal tissues and the samples taken from tumours. It also appears that there are two distinct groups of samples from tumours: one more to the upper right corner of the plot, and another that takes over the bottom left half. We will analyse these two clusters in more detail on exercise II. c).

We can also see that there are a few normal-sample outliers: some orange marks are detached from the orange cluster. These are a cause for concern. A naive classifier might label them as diseased, causing a false positive. There are also a few green marks within the orange cluster, and these are even more preoccupying: they may be classified as healthy when they are in fact diseased. A false negative such as this may result in a patient declared healthy when they, in fact, have cancer, which means they will not be getting the treatment they need.



(a) MDS plot of samples coloured according to normalised read counts.



(b) MDS plot with samples coloured according to whether they are from tumour or normal tissue.

Figure 6: Multidimensional Scaling (MDS) plots of the data after normalizing.

II. c) Phenotypic traits and genes that dominate variance

In an effort to separate the 2 different clusters formed by diseased tissue samples in the MDS plot in figure 6b, we produced a new MDS plot evidencing the PAM50 classification of each sample, by colouring the respective mark with a different colour for each class. We can see that there is a clear distinction between the clusters. The upper right cluster corresponds to tissue samples with Basal classification, while the group on the bottom left is comprised of the Luminal type samples (both A and B). We can also see that the HER2-enriched samples are positioned "in the middle" of these two groups. Finally, there are too few normal-like samples to discern any pattern. We conclude that PAM50 phenotype dominates data variance, and that the PAM50 genes are associated with the main axes of variance. Since we can explain all data variance using biological factors, we do not think there is any non-biological effect worth acting on.

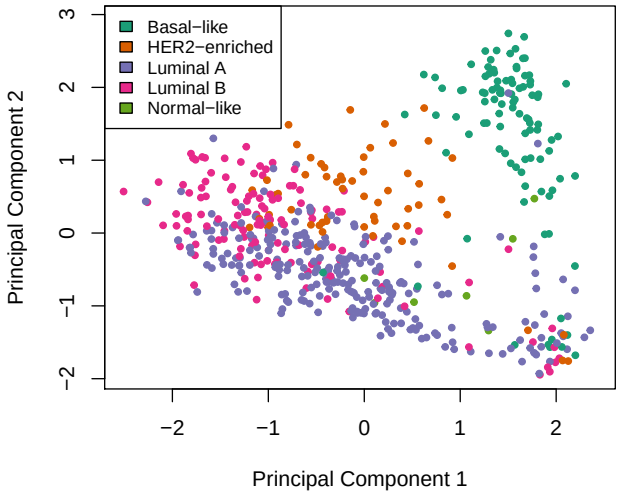


Figure 7: MDS plot of the data after normalizing, grouped by PAM50 classification.

II. d) Differentially expressed genes between tumour and normal samples

We used the `edgeR` package to compute the differential expression of each gene with respect to whether the sample is healthy or has a tumour. Then, we selected the top-5 most differentially expressed genes. The results are shown in table 1.

Next, we repeated te proccess, but this time the differential expression of each gene was computed with respect not only to whether the sample is healthy or has a tumour but also to the age of the patient. The results are shown in table 2.

We can see that the top 5 genes are exactly the same, from which we conclude that the age of the patient does not affect the differences in expressed genes between tumourtumours and normal breast samples.

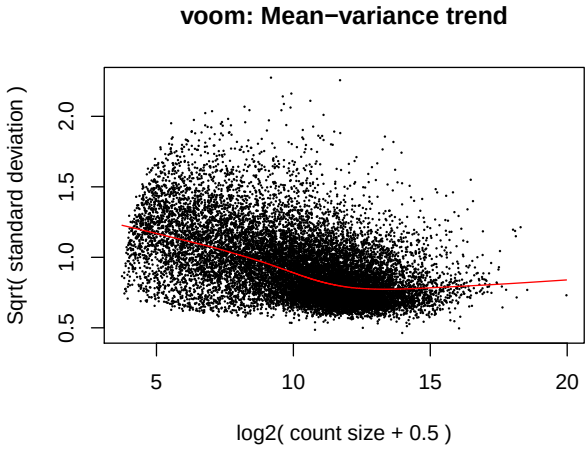


Figure 8: Vooms?

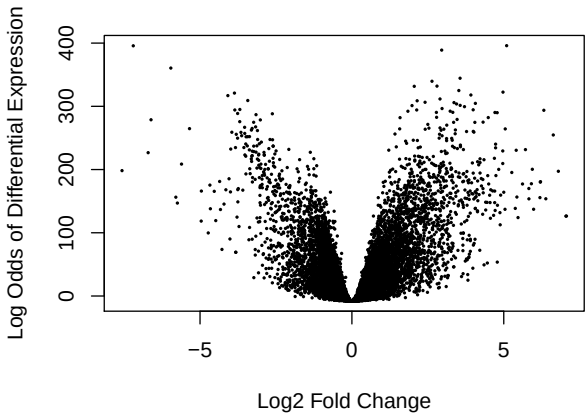


Figure 9: Vooms?

Group III.

III. a) How well can genes recapitulate related immunohistochemistry tests

We want to show how the gene expression for some genes relates to the binary classification for the related immunohistochemistry-based tests in the patient data. To do so, we produce 3 plots, showed in figures 12, 13 and 14.

First, in figure 12a we have a scatter plot where each mark is a tissue sample, and it is placed according to the read count of ESR1 on the x-axis and the read count of ERS2 on the y-axis. Each mark is colored according

Table 1: Top 5 most differentially expressed genes between normal and tumour breast samples

	logFC	AveExpr	t	P.Value	adj.P.Val	B
CD300LG	5.095	1.386	36.42	0	0	395.9
COL10A1	-7.198	5.751	-36.41	0	0	395.7
LOC728264	2.956	3.785	35.94	0	0	388.9
MMP11	-5.962	7.204	-33.99	0	0	360.5
ABCA10	3.561	1.039	32.90	0	0	344.4

Table 2: Top 5 most differentially expressed genes considering both the tumour classification of the tissue and the age of the patient

	logFC	AveExpr	t	P.Value	adj.P.Val	B
COL10A1	-7.202	5.751	-36.43	0	0	395.9
CD300LG	5.091	1.386	36.41	0	0	395.6
LOC728264	2.951	3.785	36.04	0	0	390.2
MMP11	-5.967	7.204	-34.04	0	0	361.1
ABCA10	3.553	1.039	33.11	0	0	347.5

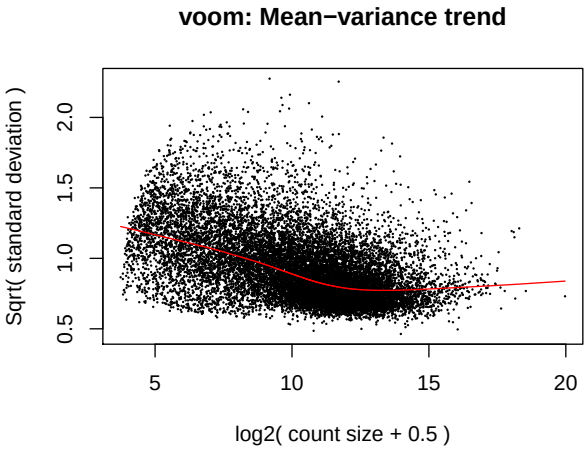


Figure 10: Vooms?

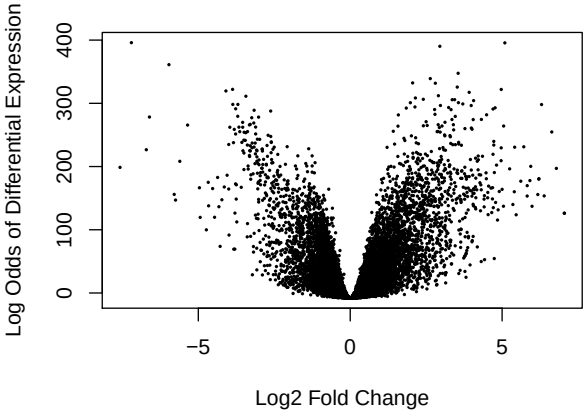


Figure 11: Vooms?

to the presence of estrogen receptors in the respective tissue sample. We can see by this graph that the expression of ESR1 is highly related to the presence of estrogen receptors: the tissue samples with estrogen receptors tend to have higher read counts of ESR1. On the other hand, ESR2 appears to have little influence on the presence of estrogen receptors. So we hypothesise that ERS1's expression is very good at recapitulating the presence of estrogen receptors in a tissue sample.

To confirm that hypothesis, we plotted a Receiver Operating Characteristic (ROC) curve. It plots the True Positive Rate (TPR) on the y-axis against the False Positive Rate (TNR) on the x-axis, and shows how the performance of a binary classifier varies as its discrimination threshold is varied. A model has better accuracy at classifying samples (fewer false positives and false negatives) when the Area Under the Curve (AUC) value is higher, i.e., the curve is above and as far as possible from the $y = x$ line. So, ideally, we want the curve to be as close as possible to upper left corner of the graph: All true positives, no false positives. The ROC curve for read counts of ESR1 gene is shown in figure 12b. We can see that it is possible to get a TPR $\approx$ 90% and a TNR $\approx$ 10% using this classifier, which adds up to a very good classification accuracy.

We conclude ERS1's expression is very good at recapitulating the presence of estrogen receptors in a tissue sample, but the same is not true of ESR2.

Secondly, in figure 13a we want to evaluate how PGR relates to the presence of progesterone receptors in tissue samples. Since we only have one gene to evaluate, we adopt a different approach: we used a histogram to

plot the number of tissue samples we have (in the y-axis) against the read count of PGR. As we can see, the trend is for the tissue samples that test positive for the presence of progesterone receptors to have a higher read count of PGR, and for the the negative samples to have lower values.

Again, we plotted a ROC curve shown in figure 13b, this time showing how well the read counts of PGR's expression classifies samples the presence of progesterone receptors in tissue samples. This time, the curve does not come as close to the upper left corner as with ESR1. However we can maintain the TPR as high as 90% maintaining the FPR at 25%. We still have AUC = 0.9, so we conclude PGR's expression is very good at recapitulating the presence of progesterone receptors in a tissue sample.

Finally, in figure 14 we use a similar histogram to show the relation between the presence of HER2 protein in the tissue sample to the read count of ERBB2 gene, as well as the usual ROC curve.

ERBB2's expressivity is not a very good indicator of the presence of HER2 protein in the sample. The AUC is now 0.77 wich is well below the other two values. We can see that we can keep a very low FPF, less than 5%, but we always have a very unsatisfactory TPR, around 50%. This means that using ERBB2's expressivity alone to classify, samples that test positive for HER2 protein will be classified as positive with the same probability that they will be classified negative: the classifier is wrong 50% of the time. However, samples that test negative for HER2 protein will be classified as negative 95% of the time.

We conclude that another measure must be used to complement a classifier for the presence of HER2 protein on tissue samples, perhaps the expressivity of another gene. If a ERBB2-based classifier is used, we should be careful when handling its results. If it classifies a sample as 'positive', we can safely trust that result, as it has only 5% chance of being wrong; on the other hand, if the classifier returns 'negative', we should not trust it because our classifier classifies half the positive samples as 'negative'.

III. b) Survival analysis

First, we produced a typical survival analysis plot, shown in figure 15.

This plot shows the probability of a patient surviving past a certain number of days depending on the sample's PAM50 classification. Analysing this plot, we conclude that Normal-like classification patients have the worst diagnosis: in this dataset, none survived past 2000 days. The second worst scenario is Luminal B: no patient suvived for longer than 4000 days after diagnosis, and only 25 % suvived past 2500 days. In 3rd place, we have HER2-enriched patients. Of these, around 70% live up to 3000 days, but only around 30% go past that date. Then, in 4th place, we have patients with Luminal A type samples. These show a linear survival probability decay from day 0 to day 4500. By day 4500, only around 30-35% are expected to have survived. Finally, the best prognostic is Basal-like. For these patients, we have again a approximately linear decay in survival until it hits 55% at around 2500 days after diagnosal. After that, our data shows a survival rate past 7000 days of around 55%.

However, these results should be taken with a grain of salt. We can see that we have very large confidence intervals for most classes, which tell us we do not have enough data to make conclusive predictions, possibily due to the large amount of censored data.

III. c) Gene signature that best classifies molecular subtypes

We compute once again the differencial gene expression, as in exercise II. d), only this time it is with respect to the PAM50 classidication. We pick the top 5 most differentially expressed genes, which are shown in table 3. Of these 5 genes, 3 (FOXA1, FOXC1, MLPH) are in PAM50 and the remaining 2 (SFRS13B, SIDT1) are not.

We use the read counts of these 5 genes as training data for a K-Nearest Neighbours (KNN) classifier, with k=5. This classifier, when given the read counts of a new sample, computes the distance for then new sample to the training samples (the distance is smaller when the read counts for the same genes are closer). Then, it picks the 5 training smaples that are closest to the new sample, i.e., the 5 neares neighbours. It will then classify the new sample with the most common class among those 5 neighbours.

To test this classifier, we performed leave-one-out cross-validation: from our N classified samples, we used N−1 as training data, and the remaining one as test data. We show the results of this in the form of a confusion matrix, in figure 18.

[TODO: explain confusion]

We tried using different classifiers besides KNN: elastic nets (before we realised these are only fit of regression problems) and naive bayes. We tried using more and fewer genes in our gene expression (always the most differentially expressed): 1, 2, 3, 4, 5, 10, 20, 50 and 100. We also tried different values for the number of neighbours (k) in KNN: 1, 3, 5, 7, 9, 11, 13 and 15. In the end, we concluded the best behaving model to be the KNN with k=5 and using only the top 5 most differentially expressed genes' read counts as training data.

After, we evaluated the performance of a KNN classifier with k=5 that uses the PAM50 genes' read expressions as training data using the same cross validation method. Despite all our efforts, the results are disheartening (and shown in figure 19).

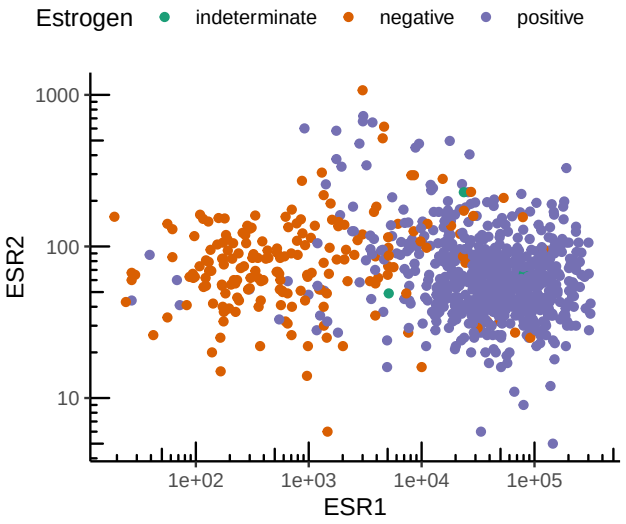
Table 3: My table

	logFC	AveExpr	t	P.Value	adj.P.Val	B
FOXC1	-3.982	3.321	-15.80	0	0	101.94
SFRS13B	-1.941	0.854	-15.29	0	0	95.74
MLPH	4.168	7.127	15.03	0	0	92.64
SIDT1	3.183	4.731	14.59	0	0	87.45
FOXA1	5.231	7.212	14.48	0	0	86.20

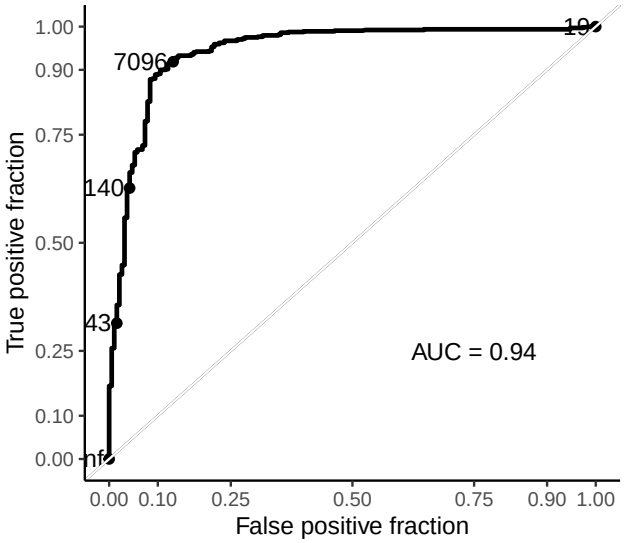
[**TODO: explain and compare confusion**]

III. d) Classifying molecular subtype

We used our KNN classifier from exercise III. c). To visualise the results we produced the MDS plot that can be seen in figure 20

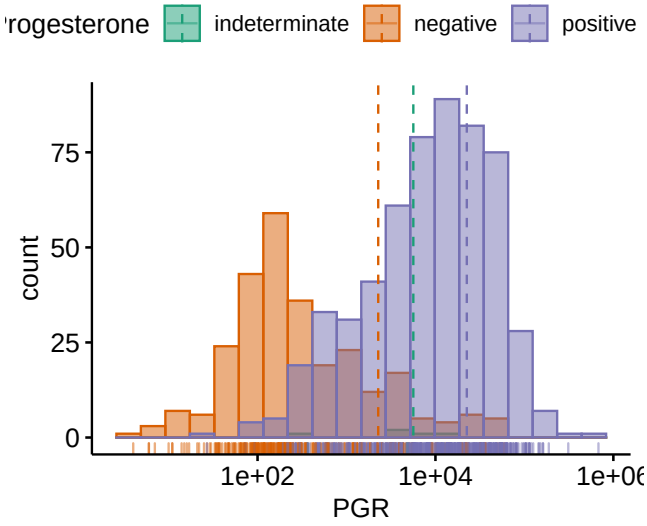


(a) Scatter plot of read counts of ESR1 and ERS2

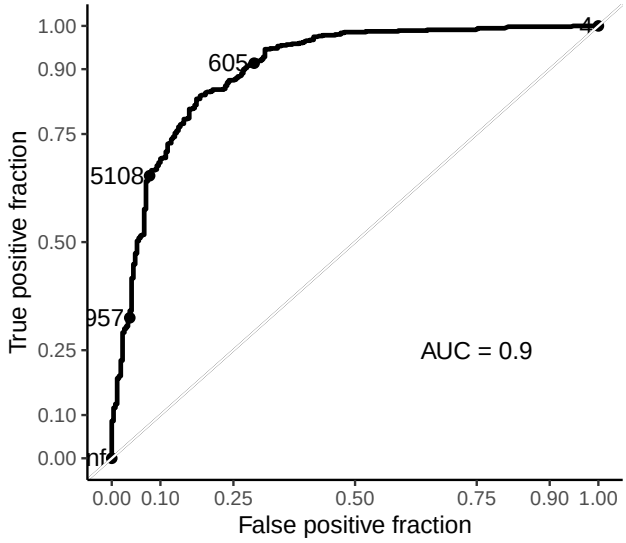


(b) ROC of ESR1 as a classifier of the presence of estrogen receptors.

Figure 12: Relation between ESR1 and ESR2's expression and the presence of estrogen receptors on a tissue sample.

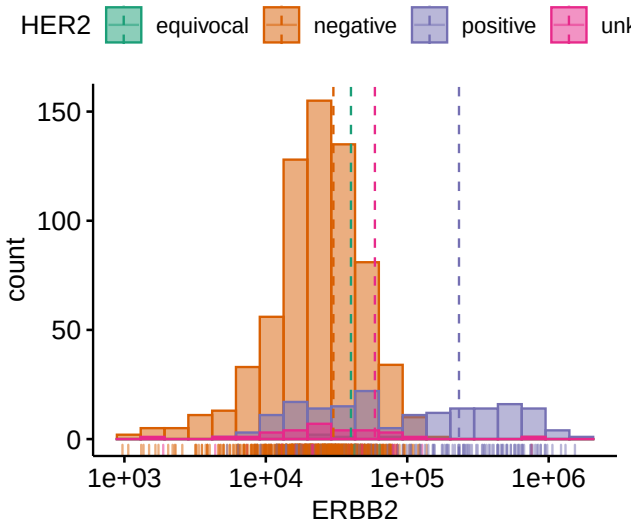


(a) histogram of tissue samples against the read count of PGR

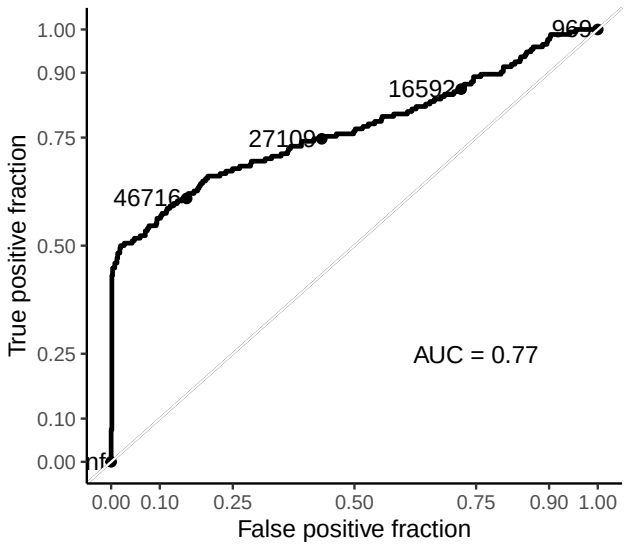


(b) ROC of PGR as a classifier of the presence of progesterone receptors.

Figure 13: Relation between PGR expression and the presence of progesterone receptors.



(a) histogram of tissue samples against the read count of ERBB2



(b) ROC of ERBB2 as a classifier of the presence of HER2 protein.

Figure 14: Relation between ERBB2 expression and the presence of HER2 protein.

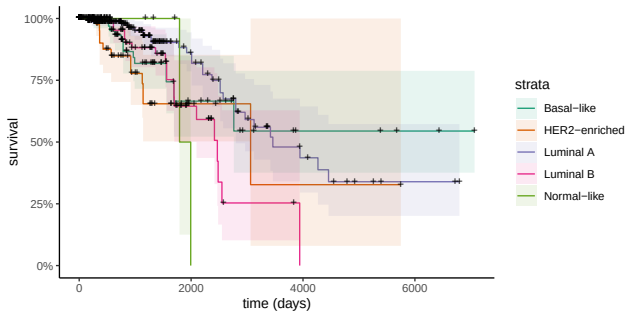


Figure 15: Hello

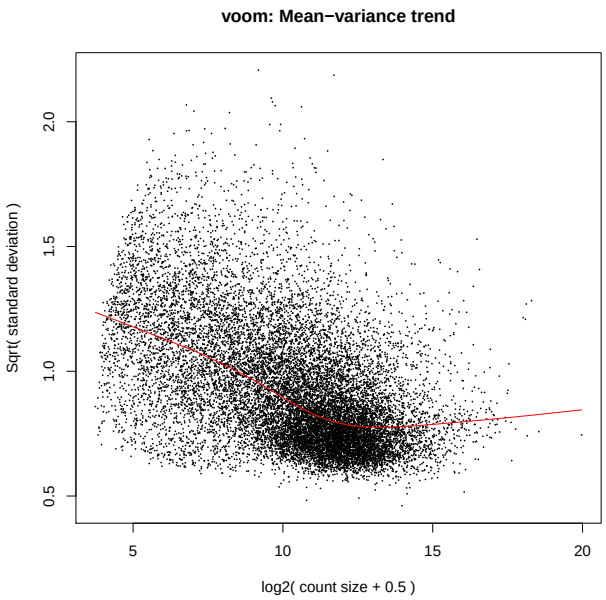


Figure 16: Hello

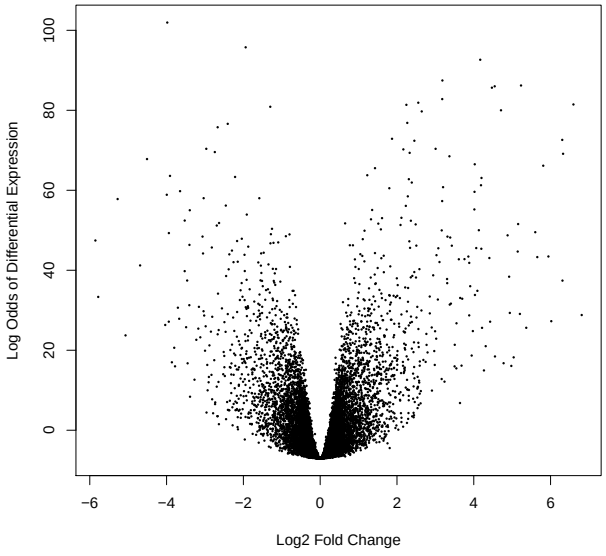


Figure 17: Hello

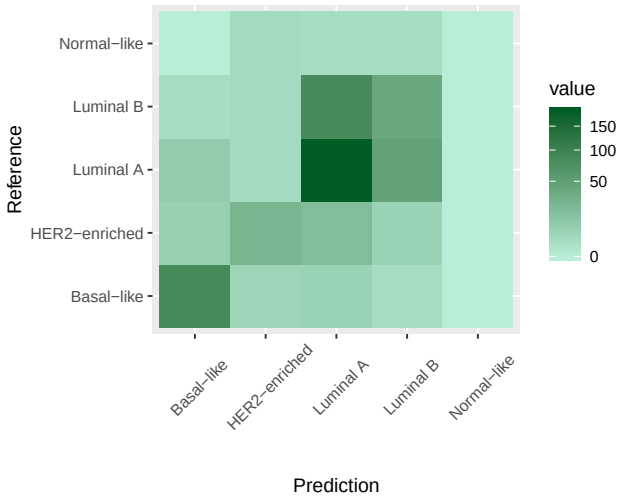


Figure 18: Confusion matrix of leave-one-out cross-validation 5-Nearest Neighbour classifier

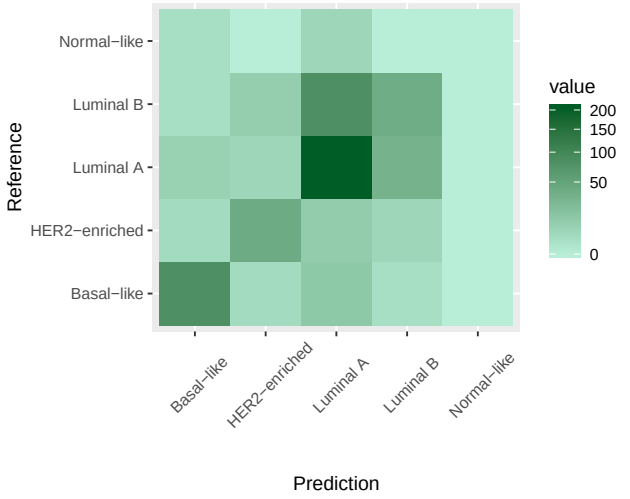


Figure 19: a

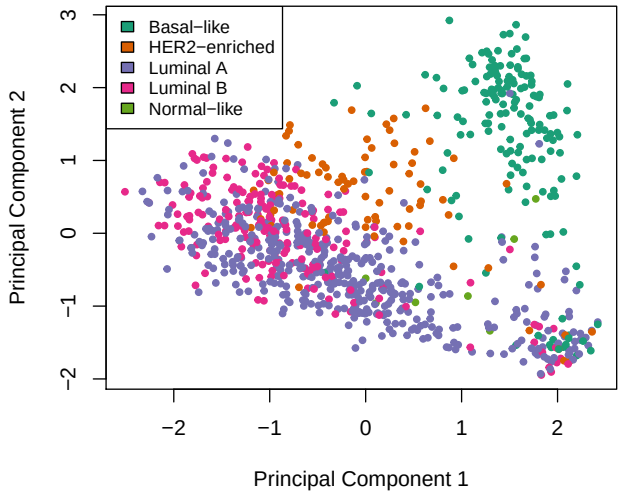


Figure 20: a