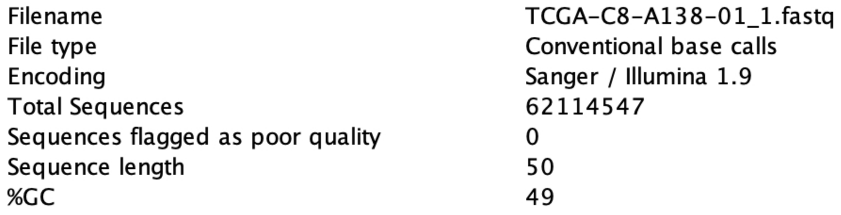
LAB#6 – OMICS APPLICATIONS

Group I

a) For the quality control of the raw data contained in the two FASTQ files the FASTQC tool was utilized. First, basic statistics data was obtained for the first file, with 62114547 total sequences detected. This is illustrated in Figure 1, having obtained the same results for the second file.



Filename	TCGA-C8-A138-01_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	62114547
Sequences flagged as poor quality	0
Sequence length	50
%GC	49

Figure 1 - Basic statistics data

Next, per base sequence quality analysis results were acquired. These are box-and-whisker plots that show aggregated quality score statistics at each position along the reads in the file. The quality score is a prediction of the probability of an error in base calling. A higher value indicates a smaller probability of error, with a lower score resulting in a significant portion of the reads being unusable. Figure 2 illustrates the plots obtained for each sequence. The red line within each yellow box represents the median quality score at the respective position. While the yellow box is the inner-quartile range for 25th and 75th percentile. The upper and lower whiskers correspond to the 10th and 90th percentile scores. For both file the results obtained are very good and all the scores are within the green range of the plot. The second file's results are slightly worse in the second half of the plot, but still in the desired range.

Finally, per base sequence content results were also derived. These plots report the percent of bases called for each of the four nucleotides at each position across all reads in the file. As is visible in Figure 2, for both files there is a clear non-uniform distribution for the first 15 nucleotides. This is due to a bias in the selection of the fragments and not in the sequencing itself, which is normal and expected for RNA-Seq data.

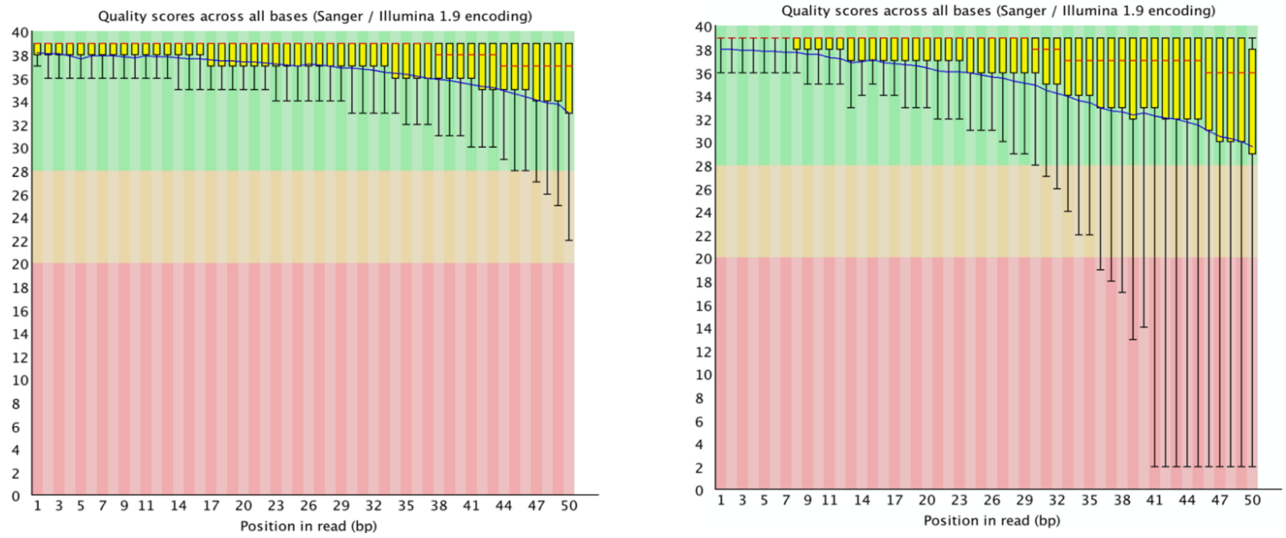


Figure 2 - Per base sequence quality plots for file 1 (left) and file 2 (right).

b) RNA-Seq analyzer Kallisto was used to estimate gene expression in the breast tumor sample `TCGA-C8-A138-01` contained in the FASTQ files provided. As reference transcriptome, the annotated human transcriptome available in https://github.com/pachterlab/kallisto-transcriptome-indices/ releases was used, having no need to build an index. The analysis was run and `abundance.tsv` file was generated. Resorting to the R programming environment, the results of the counts obtained by RNA-Seq were correlated against those in the read count table contained in the `TCGA_BRCA_Gene_ReadCounts.txt` file that was provided. The plot obtained is represented in Figure 3.

An analysis of the plot shows that for higher values of the counts, the points are more adjusted to the linear fit, as for lower values they aren't as much. Despite this, it can be concluded that a good correlation is obtained, with the dispersion in the lower range being a probable effect of technical noise, which is more visible in the lower counts.

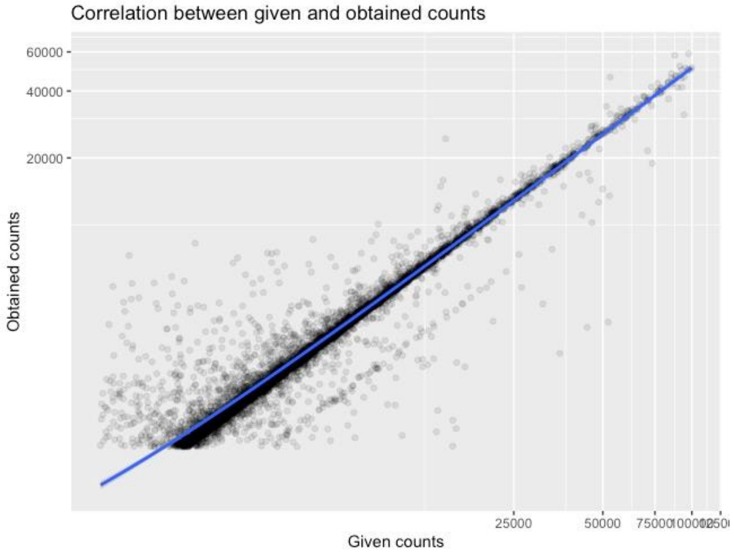


Figure 3 - Correlation between counts in data file and calculated (values in ln).

Group II

a) Library preparation can seriously affect the outcome of the sequencing in terms of quality. For the 878 samples provided in the read count table, the evaluation of their quality was achieved through an estimating analysis of the read coverage and library complexity.

For the analysis of read coverage per sample, it was calculated the total number of reads counts per sample (sum of counts across all 20502 genes), posteriorly obtaining a histogram of the distribution of the total estimated counts (Figure 4). From the histogram (Figure 4), it can be inferred that the distribution appears to be unimodal, being possible to discriminate a peak within the interval [6 x 10^7, 1 x 10^8]. Moreover, it is also inferred from the distribution width (it ranges from 4.0 x 10^7 to 1.8 x 10^8) that there is a considerable number of samples with a higher number of reads. Indeed, samples contain a certain number of mRNA molecules and, if in a particular sample there are a few genes that are highly expressed, it is going to be underestimate the expression of lowly expressed genes, since they will be underrepresented in our library. To identify sequencing errors, every base should be covered more than once, and a normalization process should be performed to provide an analysis of these samples with some significance and avoid misleading results when it comes to differential expression.

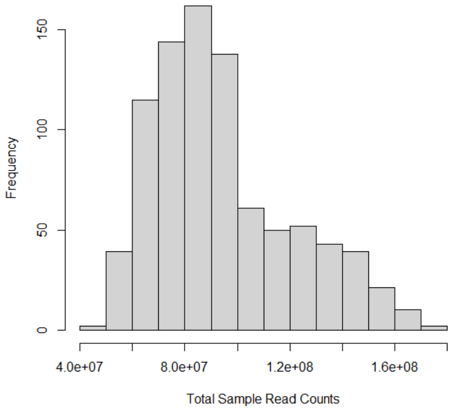


Figure 4 - Histogram illustrating the read coverage per sample, with each bin enclosing a range of 1x10^7 total read counts.

In order to evaluate how the distribution of expression across genes influences the profile with RNA-seq, it was assessed the library complexity. For this, there were randomly selected 1000 genes in each sample, having been normalized the read counts according to the total number of reads. Then, in order to compute a cumulative distribution (Figure 5-A) of read counts for each sample, the genes were sorted by their expression in decreasing order, allowing to construct a measure of the proportion of reads taken by a certain number of genes in each sample. By analysing the Figure 5-A, it is concluded that the library complexity distributions assessed for 1000 random genes resemble a similar evolution across all samples, although corresponding to differently ordered genes, being concluded that the most expressed genes are different. Moreover, only 200 genes result in more than 60% of the total number of counts for each sample, leading to the conclusion that there is a small number of overly expressed genes which takes a great majority of the total number of counts.

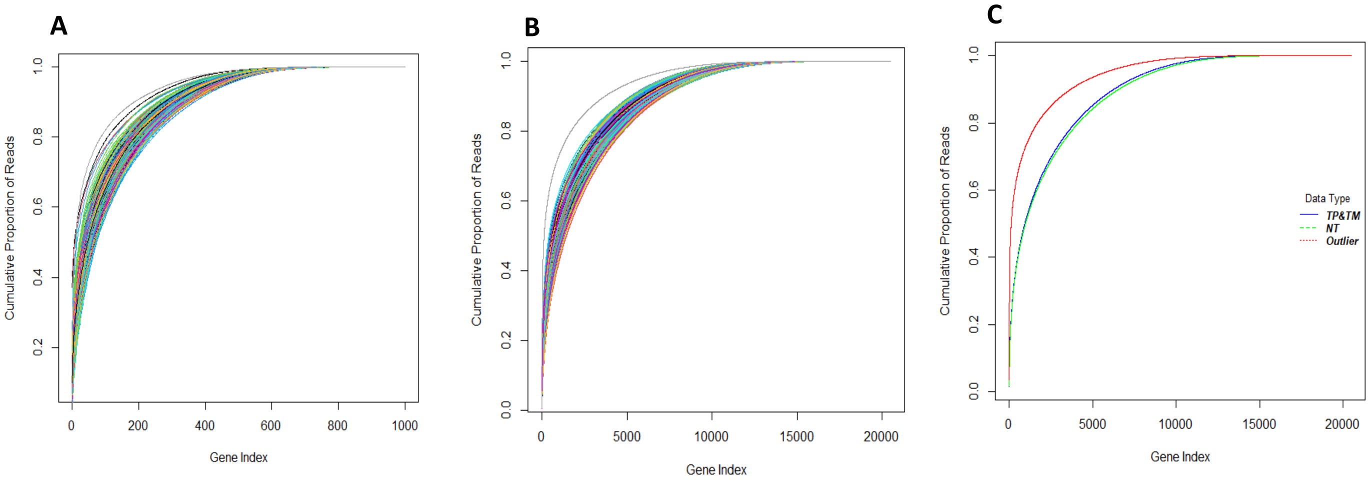


Figure 5 - (A) Normalized cumulative distributions of read counts as a function of 1000 random genes, for all samples; (B) Normalized cumulative distributions of read counts as a function of all genes, for all samples. (C) Mean of the normalized cumulative distributions for the cancerous (blue) and non-cancerous (green) samples and the outlier (red). TP - Primary Solid Tumor; NT - Solid Tissue Normal.

In order to broaden the analysis, the previously described process was repeated for the whole dataset (20502 genes) (Figure 5-B), allowing the identification of an outlier (sample TCGA.GI.A2C8.11), whose genetic complexity was much lower. To capture the divergence of this outlier sample from the remaining ones, it was plotted (Figure 5-C) the mean of the cumulative distributions for the different patient classifications (samples from Metastasic and Primary Solid Tumor were merged). For the outlier sample, only 121 genes accounted for approximately 50% of the read counts. For the samples regarding non-cancerous and cancerous patient classifications, a mean of approximately 50% of the read counts was achieved for 978 and 935 genes, respectively. This also allowed to conclude that the difference between cancerous and non-cancerous samples is negligible.

b) The purpose of normalization is to allow a proper analysis of gene expression profiles between samples, being the gene expression rarely considered at the level of raw counts, since libraries sequenced at a greater depth will result in higher counts. The distribution of the logarithm of read counts for 20 randomly selected raw samples through boxplots is displayed in Figure 7-A.

While the majority of normalization methods work well, RPKM, FPKM etc. are only needed if expression values need to be compared between different genes within the same sample, for which the different gene lengths must be taken into consideration, since longer genes yield more RNA, meaning more fragments and more reads [1]. The normalization method selected was the trimmed mean of M (TMM).

The first step in the normalization process implemented consisted in the removal of genes whose read counts were zero across all samples (8 genes). The subsequent normalization is performed using R library edgeR, sequentially applying the DGEList(), calcNormFactors() and voom() methods.

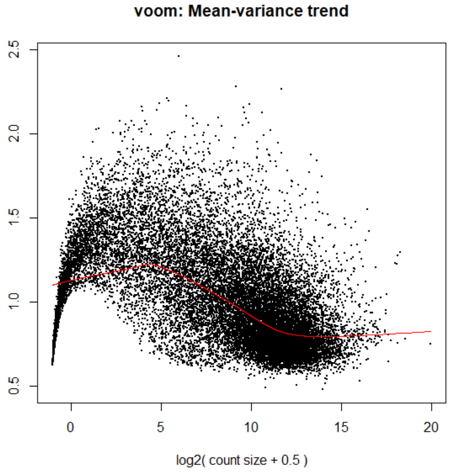


Figure 6 – Mean variance trend plot, obtained from voom analysis.

calcNormFactors calculates the normalization factors to scale the raw library sizes, with the normalized read counts obtained by dividing raw read counts by the TMM-adjusted library sizes [1]. As the final step, *voom* calculates, for every sample and gene, the counts per million reads (CPM) and log2-transform these. A linear model is fitted to the log2-CPM values [1], having considered the default design matrix, which assumes that all samples are essentially replicas, being reasonably suited in the case of the dataset provided. Finally, the resulting residual standard deviations for every gene are used to fit a global mean-variance trend across all genes and samples (Figure 6) [1].

The resulting distributions were once again displayed as boxplots (Figure 7-B), being relative to the same 20 random samples as before. From the side-comparison with the raw data distributions, the median of normalized counts between samples is more constant after normalization, allowing their gene expression profiles to be comparable. Running the code for different randomly chosen samples, no samples seemed to be raising concerns. As for the outlier sample previously identified (TCGA.GI.A2C8.11), the boxplot of the sample (870 sample) after the normalization process was analysed, having a comparable size and median in comparison to the remaining samples, not impairing the normalization process.

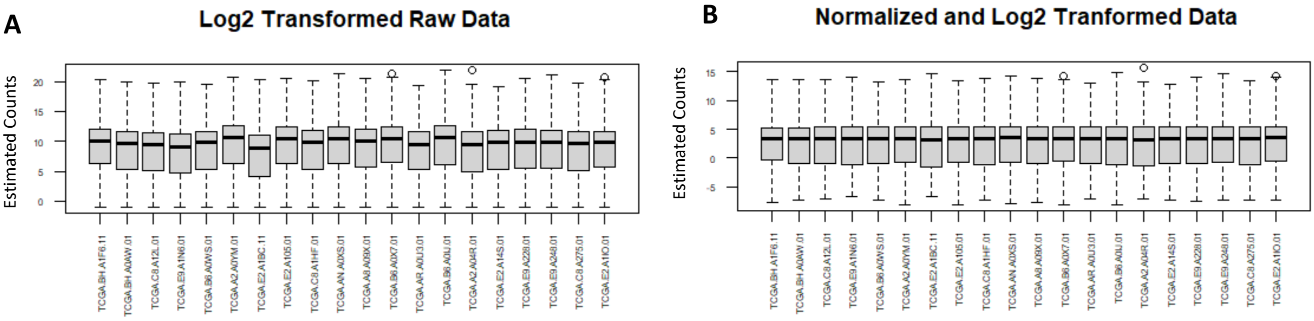


Figure 7 – (A) Boxplots of the logarithm of read counts regarding 20 random samples, prior to normalization. (B) Boxplots of normalized and log-transformed read counts regarding 20 random samples.

c) A principal component analysis was performed to understand what the main causes of gene variance were and if there were no batch effects. This analysis is a useful technique for exploratory data analysis, allowing a better visualization of the data shape and the identification of variations present in a dataset with many variables. Principal components, also called PCs are as many as samples and represent a linear combination of original predictor variables which capture the variance in the dataset and are sorted by decreasing variance – making PC1 the component which determines the direction of highest variability in the data. All succeeding principal components capture the remaining variation without being correlated with the previous component.

The data for the PCA analysis was centered and the first 20 PCs with the higher percentage of variation were plotted – Figure 8.

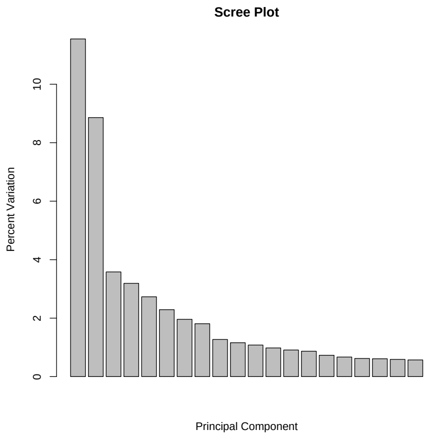


Figure 8 - Scree plot of the first 20 principal components, resulting from the PCA analysis, and respective percent variation (calculated by squaring the "sdev" element of the PCA

The first step was to create a matrix which contained all the variables present in the `TCGA_BRCA_ClinicalAnnotation.txt` table and the sample type (normal, tumor or metastasis). Using the function `prcromp` in R, a PCA was conducted.

In Figure 9 it can be seen that the main axis of variance, PC1, separates samples according to their cell type. On the left of the graph we have breast cancer samples while on the right we have normal tissue samples from healthy cells of breast cancer patients.

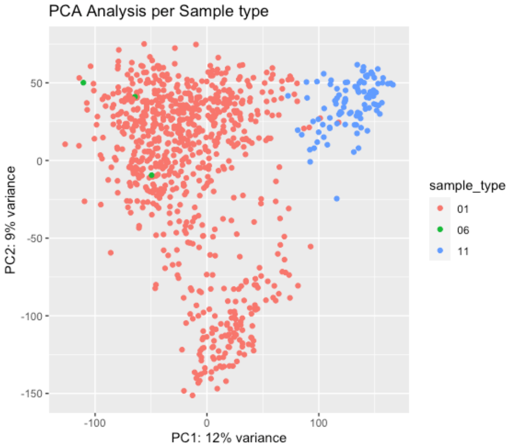


Figure 9 - PCA plot separated according to sample type. 01- tumour samples; 06 – metastatic samples; 11 – normal samples.

Other characteristics were evaluated, as it can be seen in Figure 10. The second axis of variance, PC2, separates samples according to the expression of Estrogen and Progesterone receptors. PAM50 (Prediction Analysis of Microarray 50 and tests a sample of the tumor for a group of 50 genes), which is a classifier used for breast tumor subtyping, also contributes to this variance. Transcriptome analyses of human breast tumors have helped defining robust molecular subtypes of cancer: Luminal A, Luminal B, HER2-enriched, Basal-like or Normal- like through PAM50. We can verify that Basal-like samples seem to cluster at the bottom, Luminal A and B samples cluster mostly at the top and HER2-enriched samples at the centre, between both top and bottom clusters. Although the separation is evident for the Estrogen and Progesterone receptors, it is not possible to affirm that they account for most of the variance, since PC1 and PC2 explain respectively 12% and 9% of the total variance, as it was possible to see in the scree plot. However, this allowed us to verify that there were no batch effects, which are non-biological factors that influence the distribution of data, due for example to the year of diagnosis.

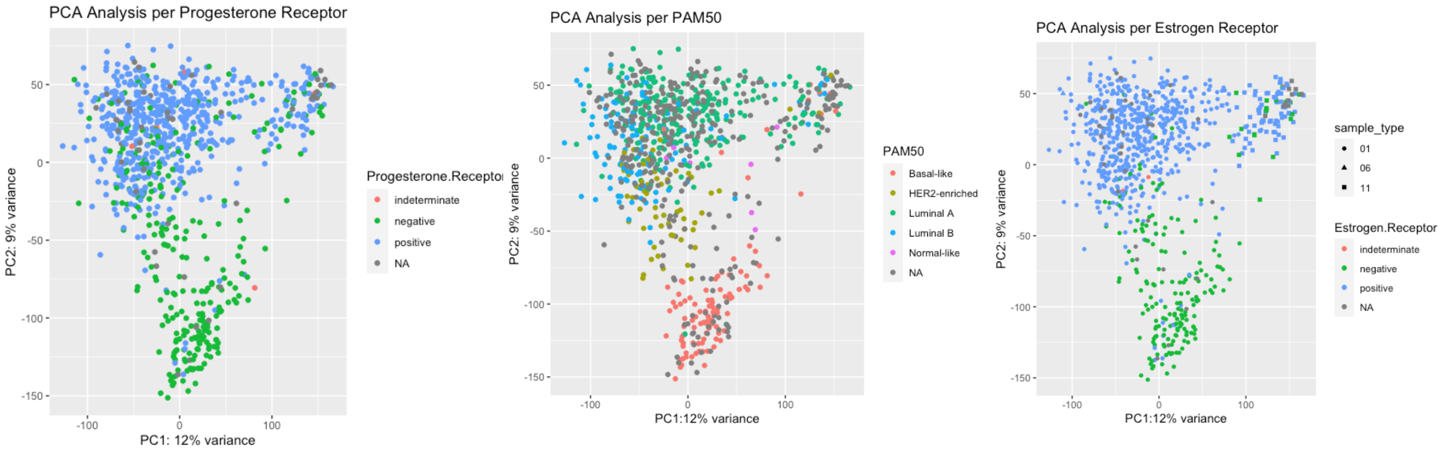


Figure 10 - From left to right: PCA plot separated by Progesterone receptor, expression; PCA plot separated according to PAM50 subtypes; PCA plot separated by Estrogen receptor expression.

Later, a rotation variable returned by the R function `prcomp` was used to get the loading scores of the genes and thus identify the genes that were associated with the first two axis of variance (PC1 and PC2). Genes with high positive or negative scores are those that most influence the position of samples in the plot. Table 1 show the first 10 genes which contribute the most to PC1 and PC2.

Table 1 - First 10 most contributively genes to PC1 and PC2.

PC1	ADH1B	SCARA5	SOX10	TUSC5	WIF1	SFRP1	CA4	FGFBP1	GLYAT	HEPACAM
PC2	C1orf64	TFF1	AGR3	ANKRD30A	CST9	CPB1	SERPINA11	ESR1	KCNJ3	TFF3

Some of the genes that are mainly associated with PC1, such as SCARA5 and TUSC5, are putative tumor suppressors; ADH1B plays a major role in ethanol catabolism, while other, like SOX10 WIF1, play a role in decisive cell pathway. Therefore, it makes sense that these genes are the ones that most contribute to the variance in PC1. Regarding the genes associated with PC2, we searched on the KEGG database and found out that TFF1, ESR1 and KCNJ3 are involved in the Estrogen signalling pathway, therefore is coherent that they highly contribute to PC2.

d) With the aim of comparing the differential gene expression between Normal and Tumour samples, some functions in R such as `linearfit`, `eBayes` and `topTable` were used. A linear analysis of the most expressed genes gives a certain condition was performed.

The baseline was defined as being the Normal condition, i.e, not having the tumour. Then the gene expression between this baseline, tumour samples, and the age influence were compared. The mean of every column was not subtracted to the column itself, and therefore the variables may not be totally independent from each other. However, for this particular analysis, it won't greatly change the results.

A volcano plot is usually used to identify meaningful changes in large data sets composed of replicate data, arranging genes along dimensions of biological and statistical significance [2]. The vertical axis specifies the statistical significance of those genes (in this case, we will be using the B value) and it is associated with the odds of differential expression and statistical evidence, or reliability of the change. Lower p-values are associated with a higher significance, and that's why genes with low p values (highly significant) will appear toward the top of the plot. The horizontal axis is the fold change and indicates the biological impact or the magnitude of the change. To be noticed that a logarithmic scale was used, so that up and down regulation appear symmetric. (log2 Fold-change, the 'biological' part of the plot).

The resulting differentially expressed genes in both conditions under analysis can be seen in Figure 11. Note that many more genes can have a significant impact in these conditions, but under certain criteria explicitly described, some genes were chosen to be analysed.

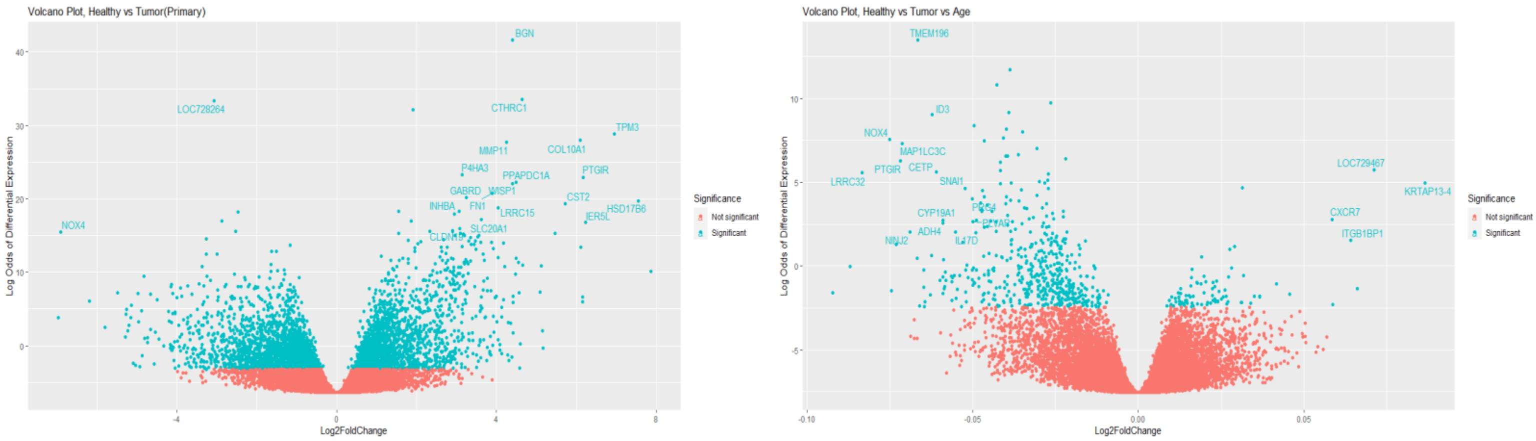


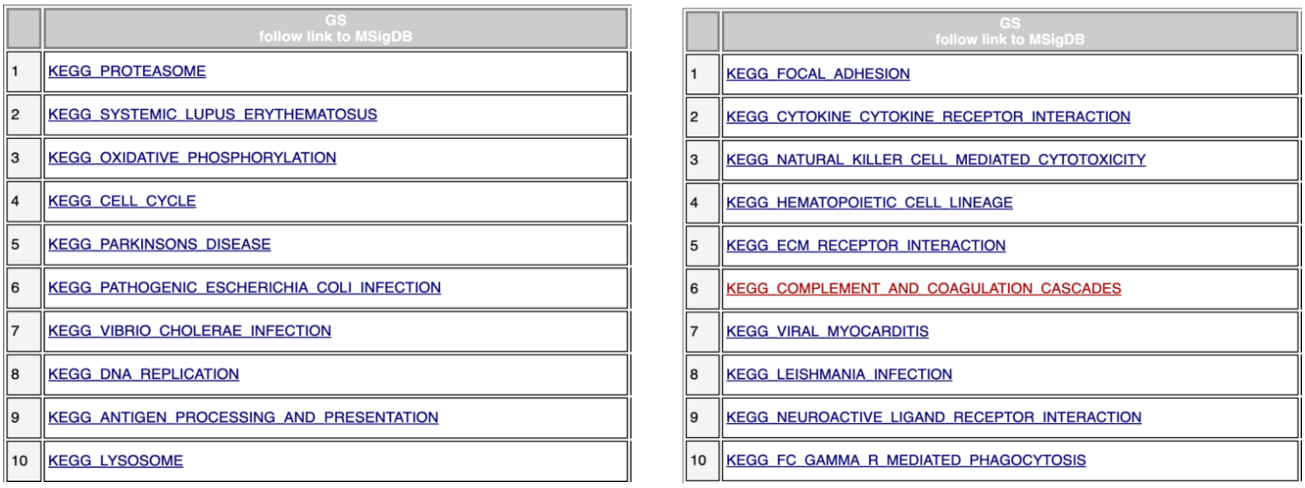
Figure 11 - Volcano plots illustrating the differential gene expression. On the left Volcano plot illustrating the most differential expressed genes in samples with tumour, being the baseline normal samples. Statistical relevant samples are coloured in blue for a threshold value of B statistics greater than 0. (B value>15 and log2(FC)<-2 or log2(FC)>4. On the right Volcano plot illustrating the most differential expressed genes in samples with tumour, under the influence of the age factor, being the baseline normal samples. Statistical relevant samples are coloured in blue for a threshold value of B statistics greater than 0. (B value>0 and log2(FC)>0.05.

Lastly, to analyse the differential expression in terms of pathways, a gene set enrichment analysis (GSEA) was performed, providing the software with the ranked gene lists obtained (see Table 2).

Table 2 - Top 10 ranked list of gene expression analysis.

Healthy vs. Primary Tumour	Healthy vs. Primary Tumour * Age
Genes	**Genes**
BGN	TNFS11
CTHRC1	CYP26A1
TPM3	LOC200726
COL10A1	PLK5P
MMP11	CNGB3
P4HA3	SLC01A2
PTGIR	C4orf37
PPAPDC1A	RSPH3
WISP1	KIAA1958
FN1	ZNF782

This allowed the evaluation of the differential expression of gene sets related to certain cellular pathways, KEGG pathways (see Figure 12).



	GS follow link to MSigDB
1	KEGG_PROTEASOME
2	KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS
3	KEGG_OXIDATIVE_PHOSPHORYLATION
4	KEGG_CELL_CYCLE
5	KEGG_PARKINSONS_DISEASE
6	KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION
7	KEGG_VIBRIO_CHOLERAE_INFECTION
8	KEGG_DNA_REPLICATION
9	KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION
10	KEGG_LYSOSOME

	GS follow link to MSigDB
1	KEGG_FOCAL_ADHESION
2	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
3	KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY
4	KEGG_HEMATOPOIETIC_CELL_LINEAGE
5	KEGG_ECM_RECEPTOR_INTERACTION
6	KEGG_COMPLEMENT_AND_COAGULATION_CASCADES
7	KEGG_VIRAL_MYOCARDITIS
8	KEGG_LEISHMANIA_INFECTION
9	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION
10	KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS

*Figure 12 - From left to right: Enriched pathways (top10) for Healthy vs. Primary Tumour (over rand under expressed) and for Healthy vs. Primary Tumour*Age (Negative Interaction).*

Knowing the phenotype of a tumour and that cancer cells typically have high replication rate and metabolic activity it can be concluded that the results go with the expectations. It is possible to observe an up-regulation in related pathways, such as DNA replication and oxidative phosphorylation. DNA replication is a biological process in which dysregulation can cause genome instability. This instability is one of the hallmarks of cancer and confers genetic diversity during de tumour development. [3]

The surrounding tissue of a tumour can react and trigger an immune response. Since the mRNA sequencing was performed in tissue samples containing both cancerous and non-cancerous cells, it is expected that we will also observe up-regulation of pathways related to this cellular response. Indeed, we find that both the proteasome and the antigen processing and presentation pathways are significantly enriched. These pathways play a role in immune responses, stress signalling, inflammatory responses, and apoptosis, which the organism utilizes to combat the tumors. Proteasome system is an important regulator of cell growth and apoptosis [4]. Specific proteasome inhibitors can act as anti-cancer agents, as they potently induce apoptosis in tumour cells. They are also able to prevent angiogenesis and metastasis and increase the sensitivity of cancer cells to apoptosis. Other enriched pathways are related to immune response and increased metabolism such as oxidative phosphorylation, hematopoietic cell lineage or natural killer cell mediated cytotoxicity.

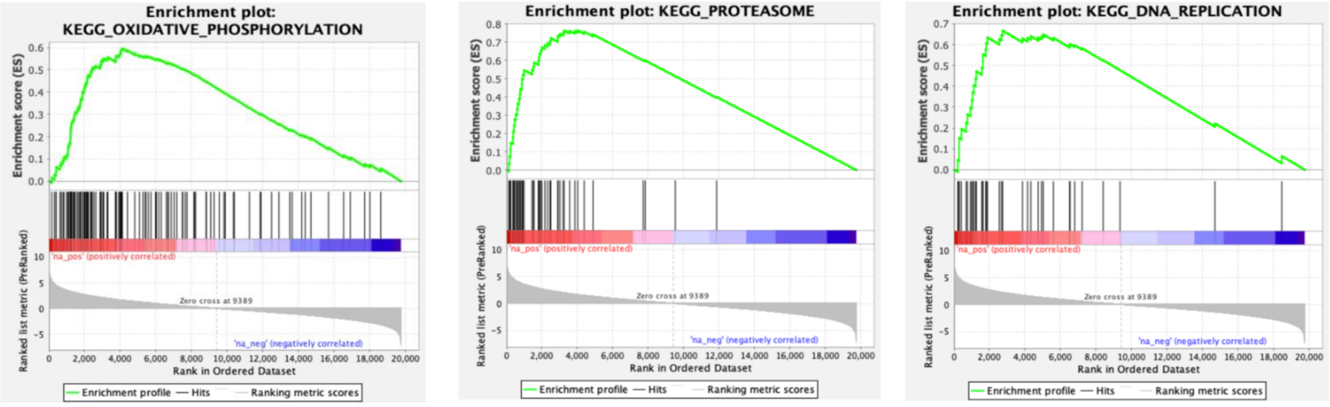


Figure 13 - Enrichment plots obtained with the GSEA software. Enrichment plots showing up-regulation of genes involved in oxidative phosphorylation, related with proteasome, and DNA replication, respectively (from left to right).

Group III

a) The goal of this exercise is to assess how good is the cognate genes' mRNA (ESR1, ESR2, PGR and ERBB2) expression at recapitulating the binary (positive and negative) classifications obtained from immunohistochemistry-based tests for the respective proteins (estrogen receptor, progesterone receptor and human epidermal growth factor receptor 2, respectively).

Initially, it was necessary to extract data so that it was study only the tumour samples (01 indicative), discarding the 'NA', 'equivocal', 'unknown' and 'indeterminate', to analyse the performance measurement of this classification problem.

The distribution of the normalized counts, for each gene, its represented in Figure 14. Analysing the plots, it is possible to settle that, overall, a smaller number of counts is associated with an absence of the protein. Though, for ESR2 gene, that doesn´t happen, being the number of positive and negative classifications, for the same range of normalized counts, approximately the same.

The Receiver Operating Characteristic (ROC) curve is used to evaluate the accuracy of a continuous measurement for predicting a binary outcome. ROC is a probability curve and the area under the curve (AUC) represents degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s – true positives and true negatives. By analogy, higher the AUC, better the model is at distinguishing between positives and negatives. [5]

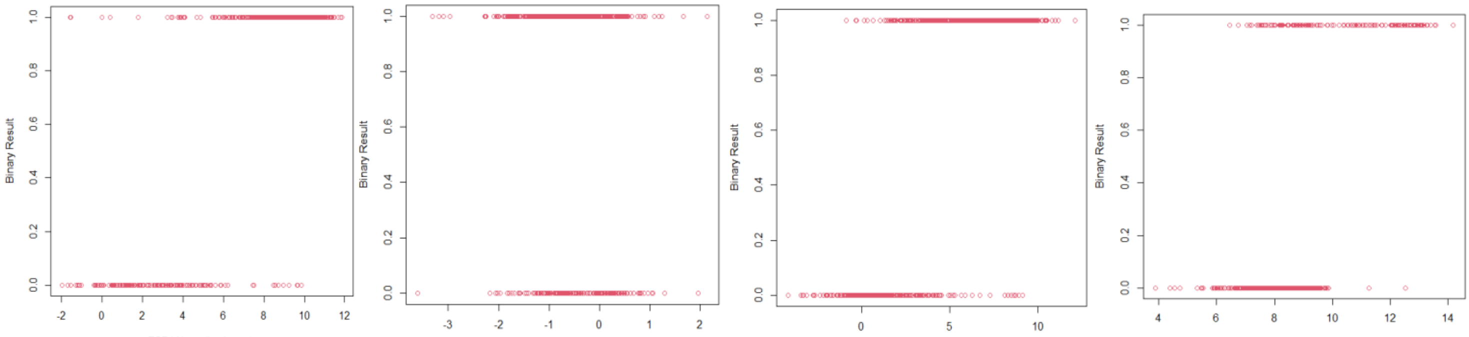


Figure 14 - Distribution of the normalized counts for the cognate genes: ESR1, ESR2, PGR and ERBB2 from left to right. The horizontal axis represents the normalized RNA-seq read counts and the vertical axis represents the respective binary classification outcome (1 - positive and 0 - negative).

An excellent model has AUC near to 1 which means it has good measure of separability. A poor model has AUC near to 0 which means it has worst measure of separability, meaning it is reciprocating the result: it is predicting 0s as 1s and 1s as 0s. When AUC is 0.5, it means the model has no class separation capacity whatsoever. [5]

In this sense, AUC was used to estimate how well the expression of each cognate gene, at mRNA level, recapitulates the classification made based on the detection of the respective protein. The ROC curves obtained, with the respective AUC values, are represented in Figure 15.

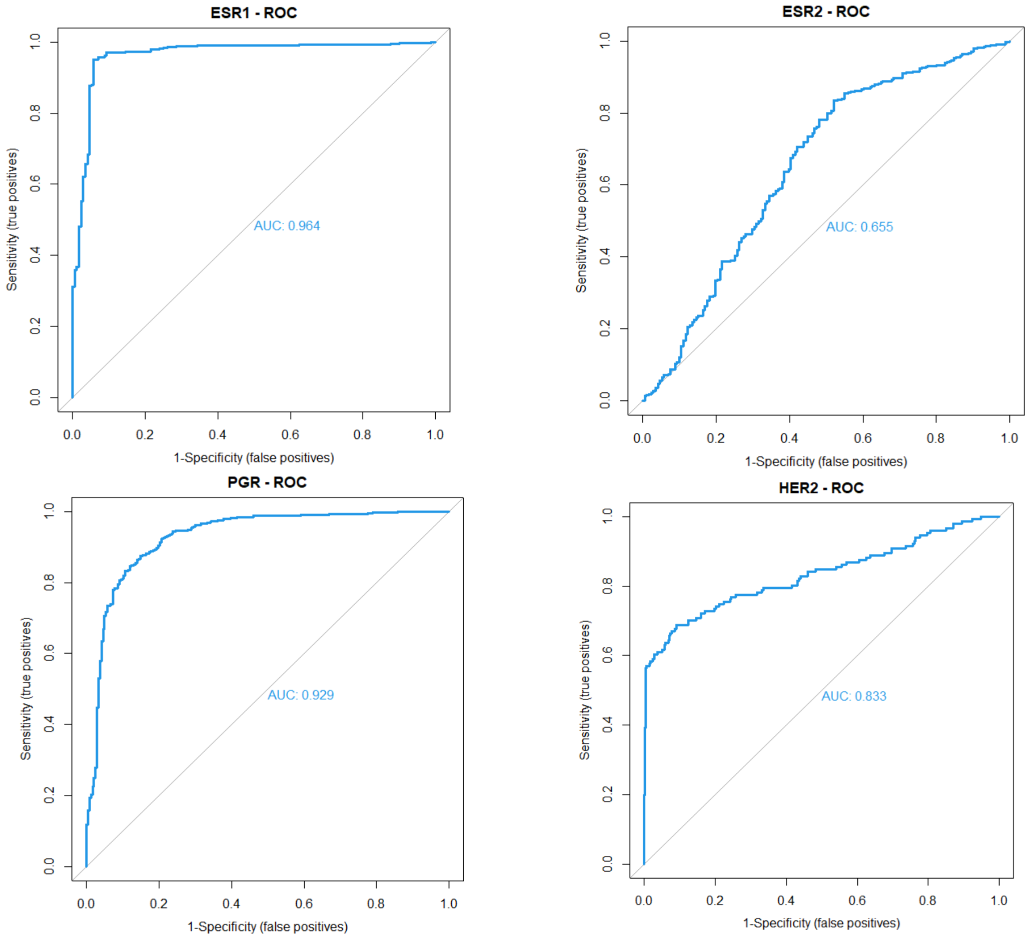


Figure 15 - ROC curves for the binary classifier for the) Estrogen Receptor (above), Progesterone Receptor (below left) and Human Epidermal Growth Factor Receptor 2 (below right).

Analysing the ROC curves for ESR1 and ESR2 genes, it can be concluded that ESR1 gene mRNA expression is better for predicting the binary outcome of the estrogen receptor, displaying a higher AUC. For the ESR2 ROC curve, AUC is near 0.5, which means that the classifier has no capability of separating positive from negative. This is consistent with the distribution of normalised counts obtained for this gene in Figure 14 and commented above. On the contrary, the ESR1 reads counts provide a good estimate of the presence of the respective receptor.

For the progesterone and HER2 genes ROC curves, both display a good AUC, being the first higher, which indicates that its expression is better for predicting the binary outcome of its respective protein (progesterone receptor).

It can also be established a threshold for gene expression that minimizes the false positives and negatives from these ROC curves. For ESR1, for example, it can be established a threshold of around 6 for the normalized counts, which can be confirmed by the first graphic in Figure 14.

To sum up, the ESR1, PGR and ERBB2 mRNA expressions are good at recapitulating the binary classifications obtained from immunohistochemistry-based tests for the respective proteins, being the first the more accurate.

b) The aim of this exercise is to compare the Kaplan-Meier curves of the different molecular subtypes and rank these in terms of prognosis.

The PAM50 is a gene expression assay to categorize breast tumours into five intrinsic subtypes: Luminal A, Luminal B, HER2 enriched, Basal-like and Normal-like, being particularly useful for epidemiologic studies where fresh tissue is typically not available [6]. In order to see which of the subtypes provide the worst and best prognostic, a survival analysis was performed, which consists in an analysis of time-to-event data. The Kaplan-Meier curves obtained are represented in Figure 16.

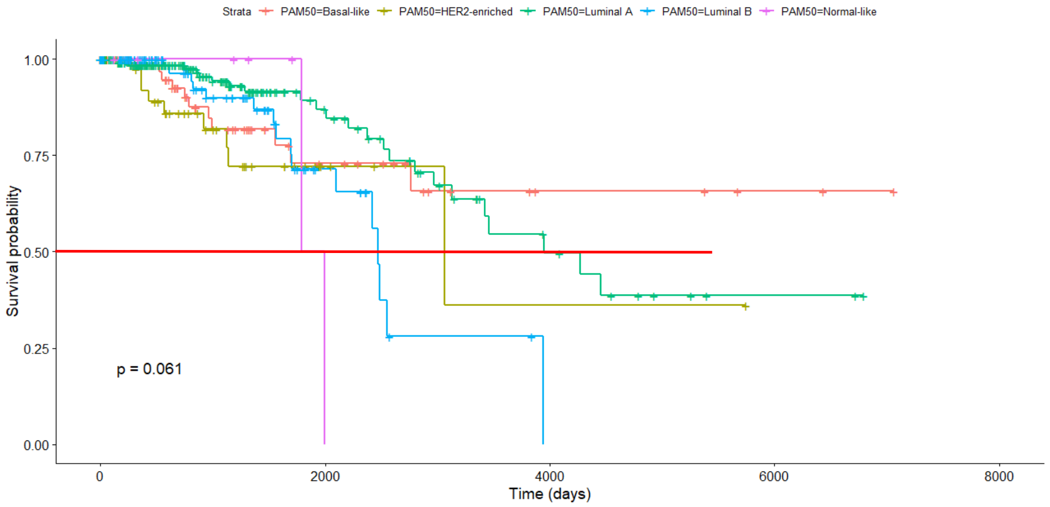


Figure 16 - Kaplan-Meier survival curve for each breast cancer subtype accordingly to the PAM50 classifier. The horizontal scale represents the time in days after patients' diagnosis and the vertical scale represents the cumulative survival probabilities. The p-value is equal to 0.061. Each line cut represents a censored sample.

Prognosis is a medical term for predicting the likely or expected development of a disease, including whether the signs and symptoms will improve or worsen (and how quickly) or remain stable over time [7].

Analysing the Kaplan-Meier survival curve for the five subtypes groups of breast cancer, it is possible to infer that some are more harmful than others. Comparing the median survival times[1], it is possible to establish an order from the worst prognostic to the best, regarding each respective subtype of breast cancer: Basal-like (worst prognostic), Luminal A, HER2 enriched, Luminal B and Normal-like (best prognostic).

To note that, although the Normal-like has a survival probability equal to 1 during the first 1980 days, in the instant 2000 days, it decays to 0 (only 2 deaths), so a correct prognostic is difficult to apply in this case. This can be explained by the fact that a small set of samples was used (7 patients normal like). A larger number would be required to perform a more reliable survival analysis.

Finally, it must be mentioned that the log-rank p-value for this analysis is greater than 0.05, which is usually considered the threshold for data to be statistically relevant.

[1] The median survival time is calculated as the smallest survival time for which the survivor function is less than or equal to 0.5 (represented by a red line in Figure 16).

c) The goal of this exercise is to find, from the expression data in primary tumours, the set of genes that best classify the 5 subtypes groups of breast cancer. Then compare this classification with the performance of the 50 PAM50 genes.

Class predicting with gene expression is widely used to generate diagnostic or prognostic models. The literature reveals that classification functions perform differently across gene expression datasets, being the most used in the literature the following: discriminant analyses or Bayes classifiers, tree-based, regularization and shrinkage, nearest neighbours and neural networks methods. [8]

Initially, to perform both variable selection and regularization it was applied the lasso (least absolute shrinkage and selection operator) regression. It was concluded that **9** genes of the selected signature, which contains 134 genes, are also present in the PAM50 signature.

For the comparison of the two signatures, the chosen classification function was Bayes Classifier. In this sense, two Naïve Bayes were trained using 3/4 of the total samples (362 samples). Then, the performance was evaluated in the testing set (121 samples) for the two classifiers. The resulted confusion matrices are represented in Table 3.

Table 3 - Confusion matrices after applying Naïve Bayes Classifier for PAM50\selected Signatures to the testing set.

PAM50\Signature	*Basal-like*	*HER2 enriched*	*Luminal A*	*Luminal B*	*Normal-like*
Basal-like	**13\12**	0\0	1\1	1\0	0\1
HER2 enriched	0\1	**13\12**	0\0	1\2	0\0
Luminal A	1\0	1\0	**51\53**	6\4	1\2
Luminal B	0\0	3\1	1\0	**28\31**	0\0
Normal-like	0\0	0\0	0\0	0\0	**0\0**

The testing set didn't consider any samples of the normal-like subtype, because, as referred in the previous exercise, this group has a very small set of samples (7). For the remaining classifications, it can be concluded that both classifiers have a great performance, with an accuracy of 87% for the PAM50 signature and 89% for the selected genes signature. However, PAM50 only requires 50 genes while the selected signature requires 134, making the PAM50 a better signature for classification of the five subtypes groups of breast cancer.

References

[1] F. Dündar, L. Skrabanek, and P. Zumbo, “Introduction to differential gene expression analysis using List of Figures,” *Appl. Bioinforma. Core*, no. September 2015, pp. 1–86, 2018.

[2] X. Cui and G. Churchill, “ Statistical tests for differential expression in cDNA microarray experiments,” *Genome Biology ,* vol. 4, no. 210, 2003.

[3] W. R. Hanahan D, “The hallmarks of cancer,” *Cell Press,* vol. 100, pp. 57-70, 2000.

[4] D. H. Lee and A. L. Goldberg, “Proteasome inhibitors: valuable new tools for cell biologist,” *CELL BIOLOGY,* vol. 8, 1998.

[5] S. Narkhede, “Understanding AUC - ROC Curve,” 26 janeiro 2018. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. [Accessed 15 12 2020].

[6] B. J. Caan, C. Sweeney, L. A. Habel, M. L. Kwan, C. H. Kroenke, E. K. Weltzien and P. S. Bernard, “Intrinsic Subtypes from the PAM50 Gene Expression Assay in a Population-Based Breast Cancer Survivor Cohort: Prognostication of Short- and Long-term Outc,” 2014.

[7] “What is the prognosis of a genetic condition?,” National Institutes of Helth (NIH), [Online]. Available: https://medlineplus.gov/genetics/understanding/consult/prognosis/. [Accessed 17 12 2020].

[8] V. L. Jong, P. W. Novianti, K. C. Roes and M. J. Eijkemans, “Selecting a classification function for class prediction with gene expression data,” *Bioinformatics,* vol. 32, no. 12, p. 1814–1822, 15 June 2016.