B. Tech. Project Report Phase II

Deciphering Breast Cancer Dynamics: HOXB2 and MMP11 Insights

Submitted in partial fulfilment of requirements for the award of the degree of
Bachelor of Technology from IIT Guwahati

Under the supervision of
Prof. Anil Mukund Limaye

Submitted by
Sarthak Ray
200106059

April 22, 2024

Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati 781039, Assam, INDIA

Certificate

This is to certify that the work presented in the report entitled **“Deciphering Breast Cancer Dynamics: HOXB2 and MMP11 Insights”** by Sarthak Ray **(200106059)**, represents an original work under the guidance of **Prof. Anil Mukund Limaye, Department of Biosciences and Bioengineering.** This study has not been submitted elsewhere for a degree.

Signature of student:

Sarthak Ray

Date: April 22, 2024
Place: IIT Guwahati

Sarthak Ray (200106059)

Signature of supervisor
Date: April 22, 2024
Place: IIT Guwahati

Prof. Anil Mukund Limaye
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati, India

Signature of HOD
Date: April 22, 2024
Place: IIT Guwahati

Head
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati, India

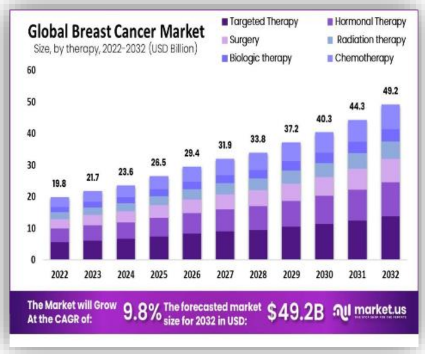
Table of Contents

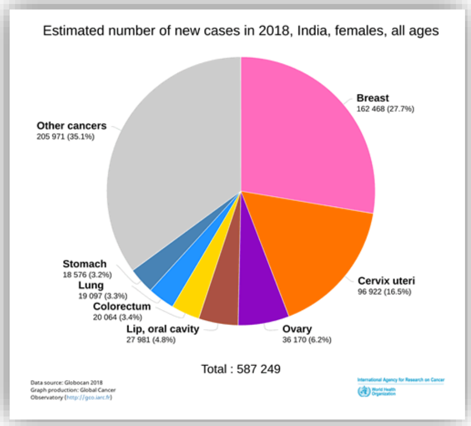
1 Abstract

We delve deeper into the intricate landscape of breast cancer (BRCA) treatment by comprehensively analysing the HOXB2 and MMP11 genes' influence on ER alpha expressions. This study focuses on elucidating the specific impact of HOXB2 and MMP11 on the expression patterns of ER alpha. Through rigorous differential gene expression analysis utilising TCGA data, we aim to uncover novel insights that further refine our understanding of molecular signals in BRCA, ultimately enhancing the prospects for tailored therapeutic interventions and improved patient outcomes.

2 Introduction

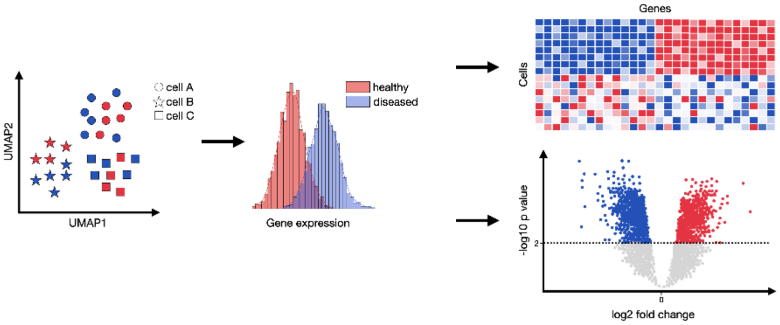
Breast cancer poses a serious threat to women's health around the world and is considered a powerful enemy in the field of global health. It is causing increasing concern due to its alarming 25% of all female malignancies. It is essential to understand the extent of breast cancer's impact on public health before diving into the disease's molecular complexity. Setting the stage for this investigation requires understanding the data and classifications.





Navigating the complexities of breast cancer involves understanding the intricate interplay between hormone expression and tumour behaviour, mainly influenced by estrogen and progesterone levels. These hormones, acting through their receptors, delineate distinct breast cancer subtypes, each presenting unique characteristics and therapeutic considerations. At the molecular level, unravelling the intricate connections between hormone expression and tumour dynamics lays the groundwork for comprehending breast cancer pathogenesis.

Differential gene expression (DGE) analysis is a potent tool that sheds light on the molecular subtleties of the disease within this complex landscape. By closely examining the transcriptome landscape, DGE studies reveal minute changes in gene expression patterns, making it possible to identify unique genetic signatures linked to different breast cancer subtypes. Examining important genes like HOXB2 and MMP11 is essential to this effort because it provides information about how these genes affect the course of breast cancer and how well a treatment works. By carefully examining the dynamics of gene expression, we hope to expand our knowledge of the biology of breast cancer and open the door to more focused and efficient treatment approaches.



3. Literature Review:

A summary of the genes we are concerned with is as follows:

- **Homeobox (HOX) Genes:**

  A family of transcription factors known as HOX genes is essential for tissue patterning, cell differentiation, and embryonic development. In breast cancer (BRCA), among other cancers, aberrant expression of HOX genes has been linked to tumour initiation, progression, and metastasis. HOXB2 is one of the HOX genes that has become important in the pathophysiology of breast cancer. According to research, HOXB2 expression is dysregulated in breast cancer tissues when compared to healthy breast tissue, and an aggressive tumour behaviour and unfavourable prognosis are linked to its overexpression. By modifying critical signalling pathways connected to cell proliferation, apoptosis evasion, and the epithelial-mesenchymal transition (EMT), HOXB2 mechanistically stimulates tumour growth and invasion. Additionally, HOXB2 has been implicated in conferring resistance to endocrine therapy in estrogen receptor-positive (ER+) breast cancers, highlighting its clinical relevance as a therapeutic target.

- **Matrix Metalloproteinase (MMP) Genes:**
  The zinc-dependent endopeptidases known as matrix metalloproteinases (MMPs) are essential modulators of the extracellular matrix (ECM) remodelling and are critical for the advancement and metastasis of cancer. MMP11, or stromelysin-3, is one of the MMP family members that has attracted much attention in breast cancer research because of its involvement in several tumour biology-related areas. When compared to normal breast tissue, MMP11 is frequently overexpressed in breast cancer tissues, and this overexpression is associated with advanced tumour stage, lymph node metastasis, and unfavourable patient outcomes. Through its ability to facilitate extracellular matrix degradation, enhance tumour cell migration and invasion, and modify the tumour microenvironment to support tumour growth and angiogenesis, MMP11 functionally promotes breast cancer invasion and metastasis. Moreover, MMP11 has been implicated in mediating resistance to chemotherapy and targeted therapies in breast cancer, underscoring its potential as a therapeutic target for overcoming treatment resistance and improving patient outcomes.

4. Objectives for Phase II:

To conduct Gene Expression analysis on BRCA data based on ER status and two specific genes, namely, HOXB2 and MMP11, elucidate the most correlated genes to the above, and document the entire analytical pipeline for comprehensive insights into targeted breast cancer treatments.

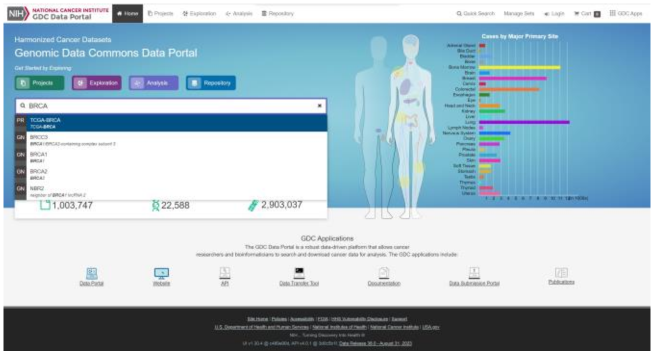
5. Materials and Methods

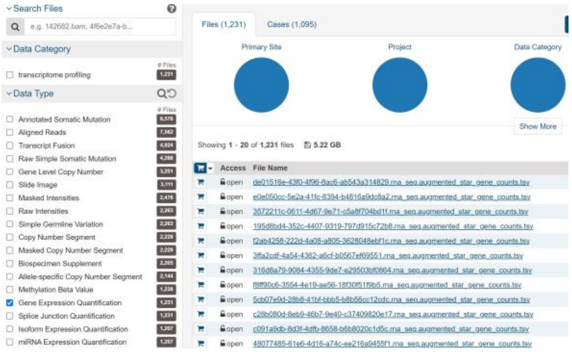
The entire code for the following pipeline can be found here.

5.1 Dataset for Analysis (TCGA)

The Cancer Genome Atlas (TCGA) project catalogues the genetic mutations responsible for cancer using genome sequencing and bioinformatics. This joint effort between NCI and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions. The steps to obtain datasets are as follows:

- Visit https://portal.gdc.cancer.gov/ and look for TCGA-BRCA in the search bar.
- Go to Explore Project Data.
- Select **Gene Expression Quantification** in the filter and download all the tsv files.


NIH NATIONAL CANCER INSTITUTE GDC Data Portal
Home
Projects
Exploration
Analysis
Repository
Quick Search
Manage Sets
Login
Cart
GDC Apps
Harmonized Cancer Datasets
Genomic Data Commons Data Portal
Get Started by Exploring:
Projects
Exploration
Analysis
Repository
BRCA
PR TCGA-BRCA
GN BRCC3
GN BRCA1
GN BRCA2
GN NBR2
1,003,747
22,588
2,903,037
Cases by Major Primary Site
GDC Applications
The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:
Data Portal
Website
API
Data Transfer Tool
Documentation
Data Submission Portal
Publications



Data Category

	# Files
□ transcriptome profiling	1,231

Data Type

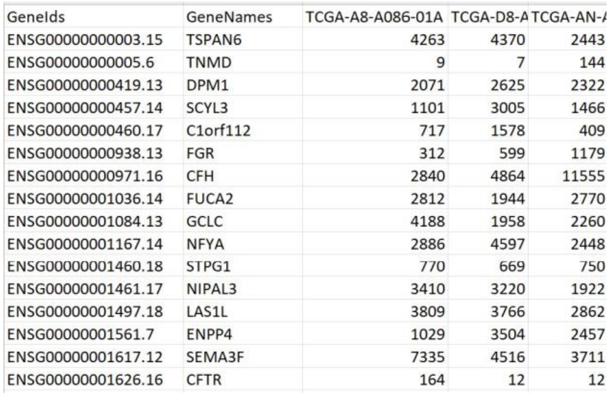
	# Files
□ Annotated Somatic Mutation	8,578
□ Aligned Reads	7,362
□ Transcript Fusion	4,824
□ Raw Simple Somatic Mutation	4,288
□ Gene Level Copy Number	3,291
□ Slide Image	3,111
□ Masked Intensities	2,478
□ Raw Intensities	2,263
□ Simple Germline Variation	2,263
□ Copy Number Segment	2,229
□ Masked Copy Number Segment	2,229
□ Biospecimen Supplement	2,205
□ Allele-specific Copy Number Segment	2,144
□ Methylation Beta Value	1,238
☑ Gene Expression Quantification	1,231
□ Splice Junction Quantification	1,231
□ Isoform Expression Quantification	1,207
□ miRNA Expression Quantification	1,207

Showing **1 - 20** of **1,231** files **5.22 GB**

Access	File Name
open	de01516e-43f0-4f96-8ac6-ab543a314829.rna_seq.augmented_star_gene_counts.tsv
open	e0e050cc-5e2a-41fc-8394-b4816a9dc8a2.rna_seq.augmented_star_gene_counts.tsv
open	3572211c-0611-4d67-9e71-c5a8f704bd1f.rna_seq.augmented_star_gene_counts.tsv
open	195d8bd4-352c-4407-9319-797d915c72b8.rna_seq.augmented_star_gene_counts.tsv
open	f2ab4258-222d-4a08-a805-3628048ebf1c.rna_seq.augmented_star_gene_counts.tsv
open	3ffa2cdf-4a54-4362-a6cf-b0567ef69551.rna_seq.augmented_star_gene_counts.tsv
open	316d8a79-9084-4355-9de7-e29503bf0864.rna_seq.augmented_star_gene_counts.tsv
open	f8ff90c6-3554-4e19-ae56-18f30f51f9b5.rna_seq.augmented_star_gene_counts.tsv
open	5cb07e9d-28b8-41bf-bbb5-b8b56cc12cdc.rna_seq.augmented_star_gene_counts.tsv
open	c28b080d-8eb9-46b7-9e40-c37409820e17.rna_seq.augmented_star_gene_counts.tsv
open	c091a9db-8d3f-4dfb-8658-b6b8020c1d5c.rna_seq.augmented_star_gene_counts.tsv
open	48077485-61e6-4d16-a74c-ee216a9455f1.rna_seq.augmented_star_gene_counts.tsv

5.1.1 Count Data

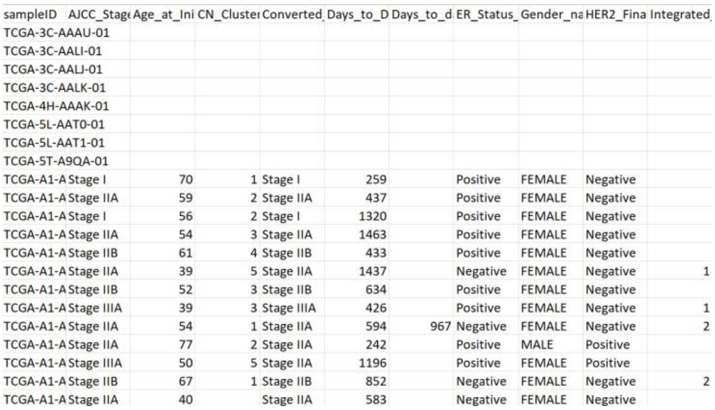
The count data in the TCGA BRCA dataset represent the quantification of gene expression levels through RNA sequencing. This dataset is organised with genes as rows and samples as columns. The basis of the DGE analysis is these counts. This results from creating a data frame from all the tsv files' unstranded sequence counts. The resulting data frame looks as follows:



GeneIds	GeneNames	TCGA-A8-A086-01A	TCGA-D8-A	TCGA-AN-A
ENSG00000000003.15	TSPAN6	4263	4370	2443
ENSG00000000005.6	TNMD	9	7	144
ENSG00000000419.13	DPM1	2071	2625	2322
ENSG00000000457.14	SCYL3	1101	3005	1466
ENSG00000000460.17	C1orf112	717	1578	409
ENSG00000000938.13	FGR	312	599	1179
ENSG00000000971.16	CFH	2840	4864	11555
ENSG00000001036.14	FUCA2	2812	1944	2770
ENSG00000001084.13	GCLC	4188	1958	2260
ENSG00000001167.14	NFYA	2886	4597	2448
ENSG00000001460.18	STPG1	770	669	750
ENSG00000001461.17	NIPAL3	3410	3220	1922
ENSG00000001497.18	LAS1L	3809	3766	2862
ENSG00000001561.7	ENPP4	1029	3504	2457
ENSG00000001617.12	SEMA3F	7335	4516	3711
ENSG00000001626.16	CFTR	164	12	12

5.1.2 Clinical Data

The clinical data in TCGA BRCA complements genomic information by providing essential insights into patients' demographic, clinical, and pathological characteristics. This dataset encompasses a range of variables, including patient age, gender, tumour stage, hormone receptor status (ER/PR), HER2 status, survival status, and other relevant clinical annotations. Clinical data is indispensable in classifying patient samples into meaningful groups for DGE analysis, allowing researchers to investigate how gene expression patterns correlate with specific clinical features. The data frame looks as follows:

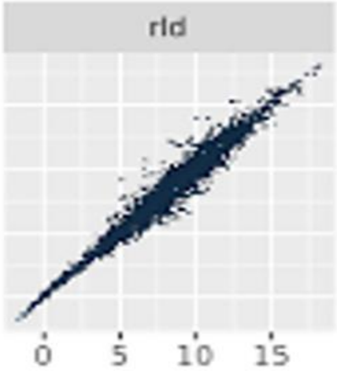


sampleID	AJCC_Stag	Age_at_Ini	CN_Cluste	Converted_	Days_to_D	Days_to_d	ER_Status_	Gender_na	HER2_Fina	Integrated_
TCGA-3C-AAAU-01										
TCGA-3C-AALI-01										
TCGA-3C-AALJ-01										
TCGA-3C-AALK-01										
TCGA-4H-AAAK-01										
TCGA-5L-AAT0-01										
TCGA-5L-AAT1-01										
TCGA-5T-A9QA-01										
TCGA-A1-A	Stage I	70	1	Stage I	259		Positive	FEMALE	Negative	
TCGA-A1-A	Stage IIA	59	2	Stage IIA	437		Positive	FEMALE	Negative	
TCGA-A1-A	Stage I	56	2	Stage I	1320		Positive	FEMALE	Negative	
TCGA-A1-A	Stage IIA	54	3	Stage IIA	1463		Positive	FEMALE	Negative	
TCGA-A1-A	Stage IIB	61	4	Stage IIB	433		Positive	FEMALE	Negative	
TCGA-A1-A	Stage IIA	39	5	Stage IIA	1437		Negative	FEMALE	Negative	1
TCGA-A1-A	Stage IIB	52	3	Stage IIB	634		Positive	FEMALE	Negative	
TCGA-A1-A	Stage IIIA	39	3	Stage IIIA	426		Positive	FEMALE	Negative	1
TCGA-A1-A	Stage IIA	54	1	Stage IIA	594	967	Negative	FEMALE	Negative	2
TCGA-A1-A	Stage IIA	77	2	Stage IIA	242		Positive	MALE	Positive	
TCGA-A1-A	Stage IIIA	50	5	Stage IIA	1196		Positive	FEMALE	Positive	
TCGA-A1-A	Stage IIB	67	1	Stage IIB	852		Negative	FEMALE	Negative	2
TCGA-A1-A	Stage IIA	40		Stage IIA	583		Negative	FEMALE	Negative	

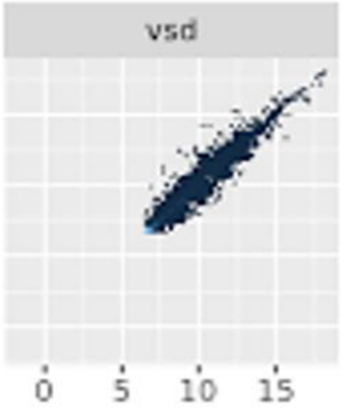
5.2 Transformations on Variance

Following data preprocessing to address sequencing depth, gene length, and RNA composition biases, transformations on variance are applied to ensure robust and accurate evaluation of gene expression changes across samples. These transformations are essential because raw RNA-seq counts typically exhibit non-constant variance across the range of expression levels, violating the assumptions of many statistical methods. Three commonly used methods are:

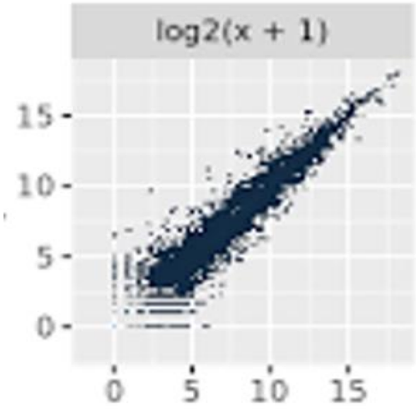
- One commonly used method for variance stabilisation is the **regularised log transformation (rlog)** provided by the DESeq2 package. The rlog transformation stabilises the variance across the mean expression level, effectively normalising the data and making it more amenable to downstream analyses such as clustering and differential expression testing. It also has a regularised component that helps with small sample sizes.



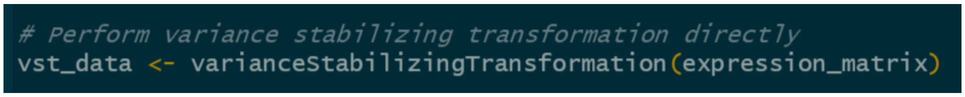
- Another widely used transformation is the **variance stabilising transformation (VST)**, also available in DESeq2. VST is based on a similar principle of stabilising the variance across the mean expression level but may be preferred in certain scenarios where the data exhibit specific characteristics. It models the relationship between the mean and variance of gene expression.



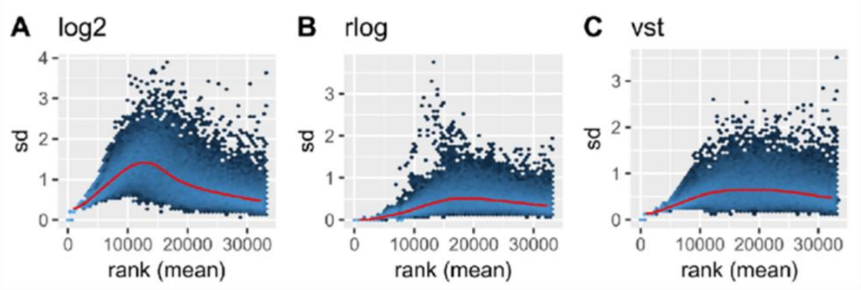
- Additionally, a common gene expression analysis practice is applying a **log2 transformation** to the normalised counts. This transformation helps in data visualisation and interpretation by compressing the dynamic range of expression values, making fold changes more intuitive and facilitating sample comparisons. The log2 transformation is beneficial for identifying differentially expressed genes and visualising expression patterns across experimental conditions.



Here, we use Variance Stabilising transformation (VST) using a simple line of code in R.

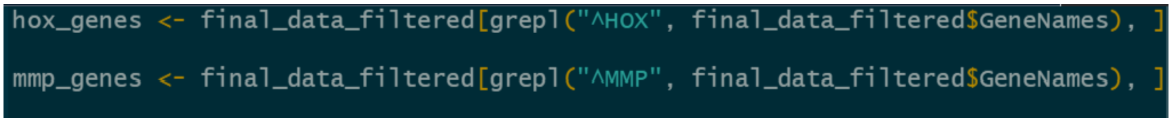


```
# Perform variance stabilizing transformation directly
vst_data <- varianceStabilizingTransformation(expression_matrix)
```

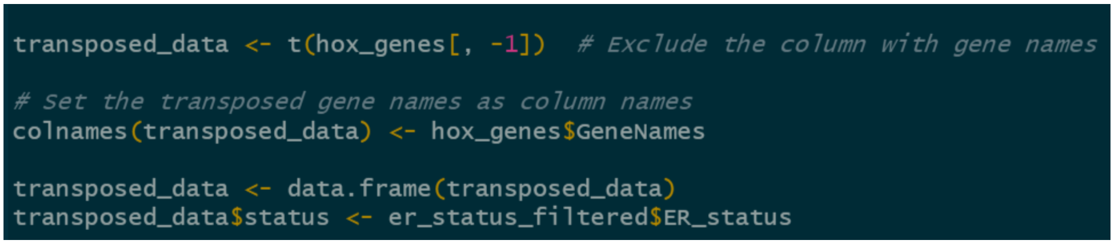


5.3 Data Analysis

- Extract MMP and HOX gene expression data from the dataset. Split the data into two groups based on ER status: ER-positive (ER+) and ER-negative (ER-) samples.

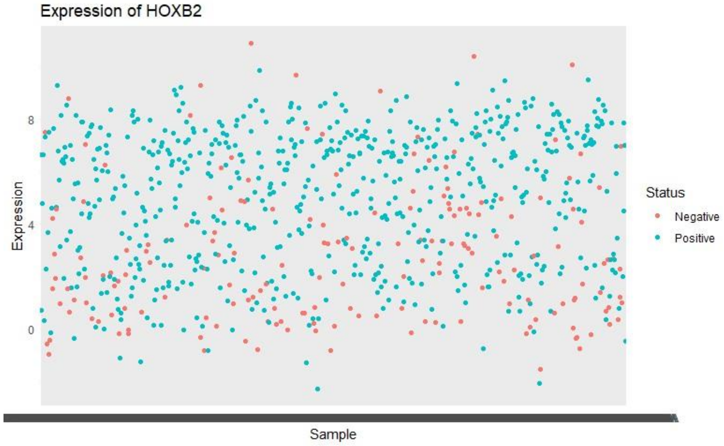


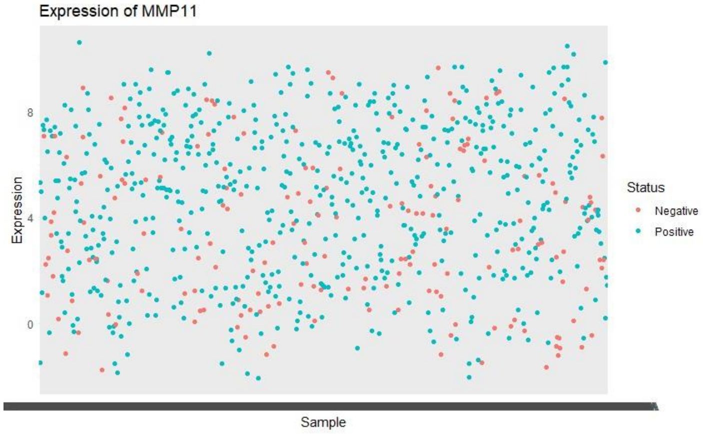
```
hox_genes <- final_data_filtered[grepl("^HOX", final_data_filtered$GeneNames), ]

mmp_genes <- final_data_filtered[grepl("^MMP", final_data_filtered$GeneNames), ]
```



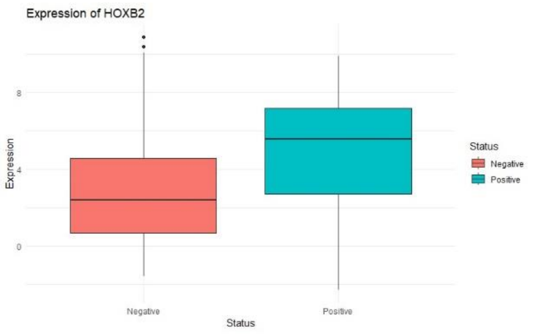
```
transposed_data <- t(hox_genes[, -1])  # Exclude the column with gene names

# Set the transposed gene names as column names
colnames(transposed_data) <- hox_genes$GeneNames

transposed_data <- data.frame(transposed_data)
transposed_data$status <- er_status_filtered$ER_status
```

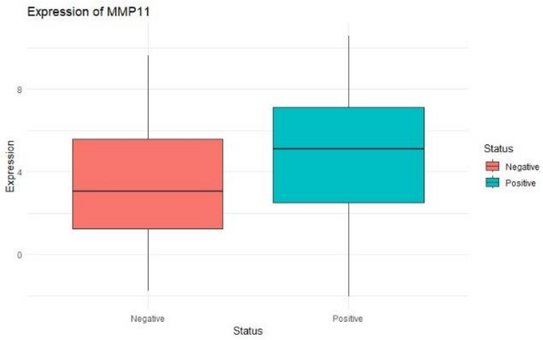
- Generate dot plots to visualise the expression levels of MMP and HOX genes in ER+ versus ER- samples. Each dot represents the expression level of a gene in a specific sample, with ER+ and ER- samples plotted separately. This visualisation method quickly assesses gene expression patterns and differences between the two ER status groups.



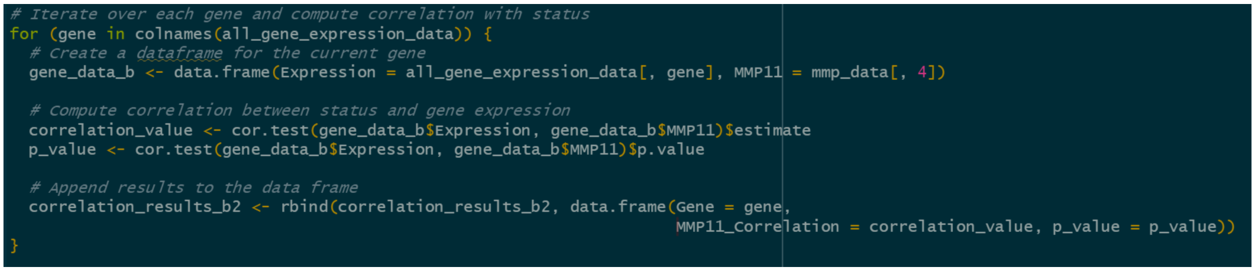


- Draw box plots for each MMP and HOX gene, comparing expression levels between ER+ and ER- samples. Box plots provide a visual summary of the distribution of expression levels within each ER status group and facilitate the identification of potential differences.



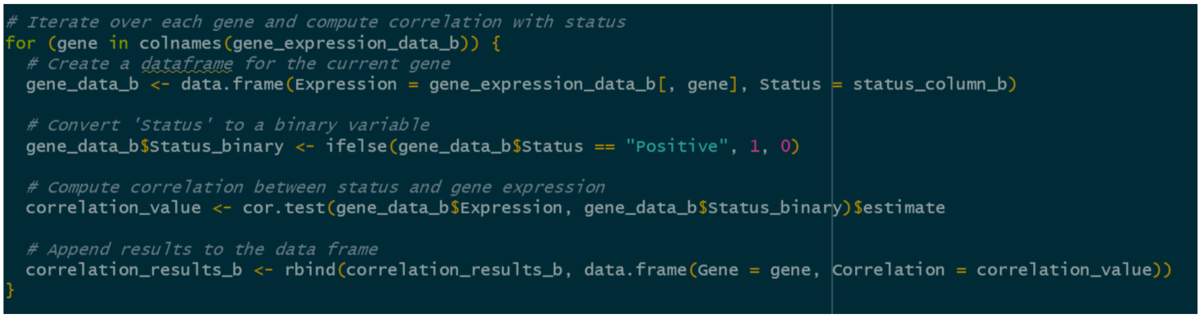


- Calculate the correlation coefficients between each MMP and HOX gene and all other genes in the dataset. Correlation analysis provides insights into potential co-expression patterns and regulatory relationships between MMP, HOX, and other genes. Examining the correlation matrix makes it possible to identify genes highly correlated with MMP and HOX genes, which may indicate functional associations or shared regulatory mechanisms.



```
# Iterate over each gene and compute correlation with status
for (gene in colnames(all_gene_expression_data)) {
  # Create a dataframe for the current gene
  gene_data_b <- data.frame(Expression = all_gene_expression_data[, gene], MMP11 = mmp_data[, 4])

  # Compute correlation between status and gene expression
  correlation_value <- cor.test(gene_data_b$Expression, gene_data_b$MMP11)$estimate
  p_value <- cor.test(gene_data_b$Expression, gene_data_b$MMP11)$p.value

  # Append results to the data frame
  correlation_results_b2 <- rbind(correlation_results_b2, data.frame(Gene = gene,
                                                                     MMP11_Correlation = correlation_value, p_value = p_value))
}
```

- Calculate the correlation coefficients between each MMP and HOX gene against the ER-alpha status. This will help us to find the relationship between the gene expression and whether or not the ER-alpha status is positive or negative.



```
# Iterate over each gene and compute correlation with status
for (gene in colnames(gene_expression_data_b)) {
  # Create a dataframe for the current gene
  gene_data_b <- data.frame(Expression = gene_expression_data_b[, gene], Status = status_column_b)

  # Convert 'Status' to a binary variable
  gene_data_b$Status_binary <- ifelse(gene_data_b$Status == "Positive", 1, 0)

  # Compute correlation between status and gene expression
  correlation_value <- cor.test(gene_data_b$Expression, gene_data_b$Status_binary)$estimate

  # Append results to the data frame
  correlation_results_b <- rbind(correlation_results_b, data.frame(Gene = gene, Correlation = correlation_value))
}
```

5.4 Additional analyses

Some other analyses can be further done on this data. They are as follows:

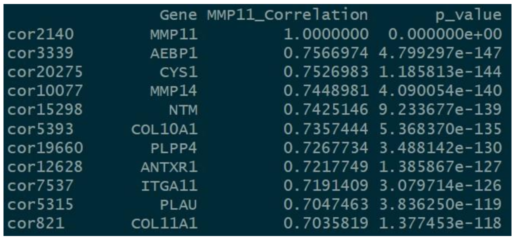
- **Pathway Enrichment Analysis:** Perform pathway enrichment analysis to identify biological pathways enriched among differentially expressed genes. This analysis can reveal the functional roles of MMP and HOX genes in breast cancer and provide insights into the underlying biological processes driving differential expression.
- **Machine Learning Classification:** Employ machine learning algorithms to build predictive models for breast cancer subtype classification based on MMP and HOX gene expression profiles. Evaluate the performance of these models using cross-validation techniques and assess the predictive power of MMP and HOX genes compared to other clinical and molecular features.
- **Network Analysis:** Employ machine learning algorithms to build predictive breast cancer subtype classification models based on MMP and HOX gene expression profiles. Evaluate the performance of these models using cross-validation techniques and assess the predictive power of MMP and HOX genes compared to other clinical and molecular features.

6.0 Results

We obtain lists of the most correlated genes with HOXB2 and MMP11. We also find the correlation between ER-alpha status and all the HOX and MMP genes.

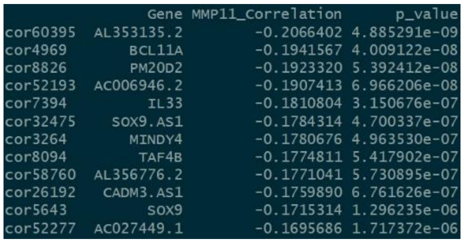
- Most correlated genes with **MMP11.**

Positive correlation



```
                Gene MMP11_Correlation       p_value
cor2140        MMP11         1.0000000  0.000000e+00
cor3339        AEBP1         0.7566974 4.799297e-147
cor20275        CYS1         0.7526983 1.185813e-144
cor10077       MMP14         0.7448981 4.090054e-140
cor15298         NTM         0.7425146 9.233677e-139
cor5393      COL10A1         0.7357444 5.368370e-135
cor19660       PLPP4         0.7267734 3.488142e-130
cor12628      ANTXR1         0.7217749 1.385867e-127
cor7537       ITGA11         0.7191409 3.079714e-126
cor5315         PLAU         0.7047463 3.836250e-119
cor821       COL11A1         0.7035819 1.377453e-118
```

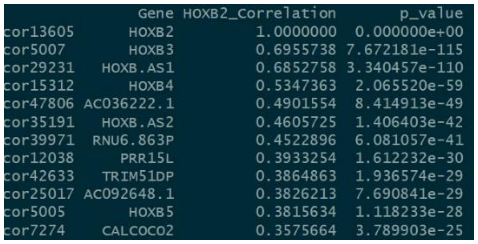
Negative correlation



```
                Gene MMP11_Correlation      p_value
cor60395  AL353135.2        -0.2066402 4.885291e-09
cor4969       BCL11A        -0.1941567 4.009122e-08
cor8826       PM20D2        -0.1923320 5.392412e-08
cor52193  AC006946.2        -0.1907413 6.966206e-08
cor7394         IL33        -0.1810804 3.150676e-07
cor32475    SOX9.AS1        -0.1784314 4.700337e-07
cor3264       MINDY4        -0.1780676 4.963530e-07
cor8094        TAF4B        -0.1774811 5.417902e-07
cor58760  AL356776.2        -0.1771041 5.730895e-07
cor26192   CADM3.AS1        -0.1759890 6.761626e-07
cor5643         SOX9        -0.1715314 1.296235e-06
cor52277  AC027449.1        -0.1695686 1.717372e-06
```

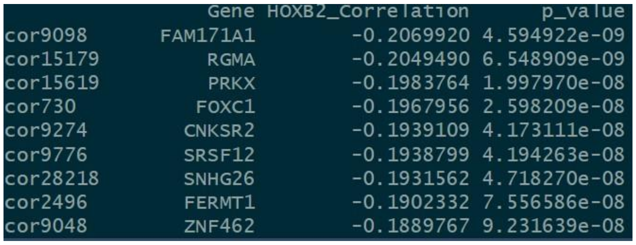
- Most correlated genes with **HOXB2**.

Positive correlation



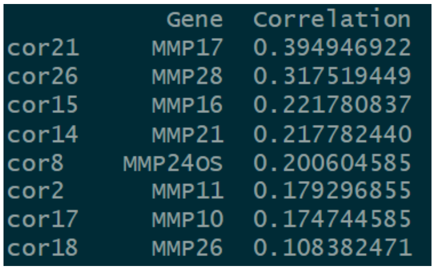
```
                Gene HOXB2_Correlation       p_value
cor13605       HOXB2         1.0000000  0.000000e+00
cor5007        HOXB3         0.6955738 7.672181e-115
cor29231   HOXB.AS1          0.6852758 3.340457e-110
cor15312       HOXB4         0.5347363  2.065520e-59
cor47806 AC036222.1          0.4901554  8.414913e-49
cor35191   HOXB.AS2          0.4605725  1.406403e-42
cor39971  RNU6.863P          0.4522896  6.081057e-41
cor12038      PRR15L         0.3933254  1.612232e-30
cor42633   TRIM51DP          0.3864863  1.936574e-29
cor25017 AC092648.1          0.3826213  7.690841e-29
cor5005        HOXB5         0.3815634  1.118233e-28
cor7274    CALCOCO2          0.3575664  3.789903e-25
```

Negative correlation



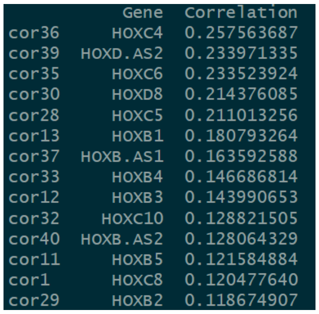
```
              Gene HOXB2_Correlation      p_value
cor9098   FAM171A1        -0.2069920 4.594922e-09
cor15179      RGMA        -0.2049490 6.548909e-09
cor15619      PRKX        -0.1983764 1.997970e-08
cor730       FOXC1        -0.1967956 2.598209e-08
cor9274     CNKSR2        -0.1939109 4.173111e-08
cor9776     SRSF12        -0.1938799 4.194263e-08
cor28218    SNHG26        -0.1931562 4.718270e-08
cor2496     FERMT1        -0.1902332 7.556586e-08
cor9048     ZNF462        -0.1889767 9.231639e-08
```

- Most correlated **MMP** genes with ER-alpha.



```
          Gene Correlation
cor21    MMP17 0.394946922
cor26    MMP28 0.317519449
cor15    MMP16 0.221780837
cor14    MMP21 0.217782440
cor8   MMP24OS 0.200604585
cor2     MMP11 0.179296855
cor17    MMP10 0.174744585
cor18    MMP26 0.108382471
```

- Most correlated **HOX** genes with ER-alpha.



```
          Gene Correlation
cor36    HOXC4 0.257563687
cor39 HOXD.AS2 0.233971335
cor35    HOXC6 0.233523924
cor30    HOXD8 0.214376085
cor28    HOXC5 0.211013256
cor13    HOXB1 0.180793264
cor37 HOXB.AS1 0.163592588
cor33    HOXB4 0.146686814
cor12    HOXB3 0.143990653
cor32   HOXC10 0.128821505
cor40 HOXB.AS2 0.128064329
cor11    HOXB5 0.121584884
cor1     HOXC8 0.120477640
cor29    HOXB2 0.118674907
```

7.0 Conclusion & Future work

To sum up, our examination of the HOXB2 and MMP11 genes in breast cancer has revealed their complex relationships with other genes. We have also shed light on the relation between HOX and MMP genes to ER status, highlighting their importance in developing the disease and hormone receptor status. These results deepen our understanding of the biology of breast cancer and could lead to customised treatment plans depending on ER status. To optimise therapeutic interventions and enhance patient outcomes, these results must be validated in larger cohorts, mechanistic studies must be carried out to clarify underlying molecular mechanisms, and their prognostic and predictive value must be investigated.

8.0 References

- https://portal.gdc.cancer.gov/projects/TCGA-BRCA
- https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html
- https://lashlock.github.io/compbio/R_presentation.html
- http://www.sthda.com/english/wiki/rna-seq-differential-expression-work-flow-using-deseq2
- https://www.blog.trainindata.com/variance-stabilizing-transformations-in-machine-learning/
- https://www.ncbi.nlm.nih.gov/gene/3212
- https://www.ncbi.nlm.nih.gov/gene/4320