B. Tech. Project Report Phase I

Precision Treatment in BRCA: Unraveling MMP and HOX Gene Signatures for Targeted Therapy Decisions

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

November 14, 2023
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 **"Precision Treatment in BRCA: Unraveling MMP and HOX Gene Signatures for Targeted Therapy Decisions"** 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: **<Signature>**

Date: Nov. 14, 2023
Place: IIT Guwahati

Sarthak Ray (200106059)

Signature of supervisor **<Signature>**

Date: Nov. 14, 2023
Place: IIT Guwahati

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

Signature of HOD
Date: Nove 14, 2023
Place: IIT Guwahati

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

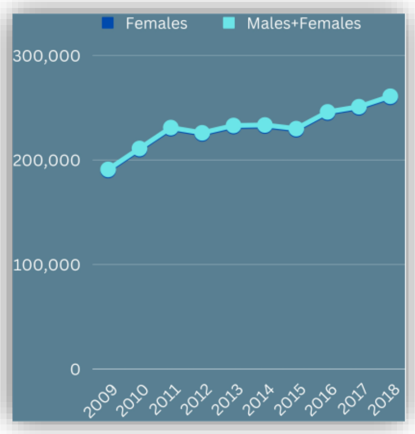
Table of Contents

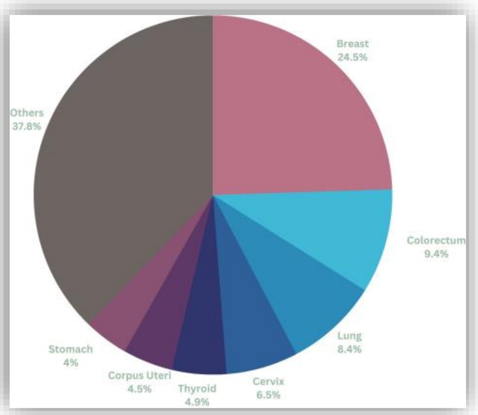
1 Abstract

This study delves into precision medicine for breast cancer (BRCA), adopting a focused assessment of the MMP and HOX genes by differential gene expression analysis using TCGA data. The primary goal is to identify discrete molecular signals that can guide nuanced decisions about endocrine therapy over chemotherapy in the treatment of BRCA. Our study contributes to a more refined and individualized approach to treatment by revealing these crucial genetic markers, improving the possibilities for improved therapeutic results in breast cancer patients.

2 Introduction

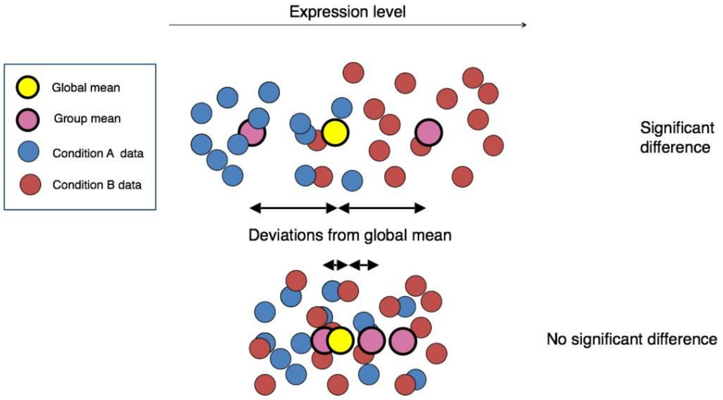
Breast cancer is a strong foe in the area of global health, posing a significant burden on women's health worldwide. It accounts for a startling 25% of all female malignancies, and its frequency is causing growing alarm. Before delving into the molecular complexities of breast cancer, it is crucial to recognize the scale of its public health impact. Understanding the data and classifications is essential to frame this investigation's context.





The interaction between hormone expression and tumour growth is one crucial factor that adds complexity to the breast cancer landscape. Hormones, particularly estrogen and progesterone, have a significant impact on the behaviour of breast cancer cells. The various hormone receptor statuses define separate subtypes of breast cancer, each with its own set of features and therapeutic implications. Understanding the subtle molecular foundations that drive breast cancer pathogenesis begins with unravelling the numerous linkages between hormone expression and tumour dynamics.

Amidst the complexities of breast cancer, the use of cutting-edge technology such as differential gene expression (DGE) analysis emerges as a strong tool for unravelling the disease's molecular complexities. DGE research reveals small changes in gene expression patterns by studying the transcriptome landscape, providing a lens through which we can distinguish the distinct genetic signatures associated with breast cancer subtypes.



3. Literature Review:

A recent review summarized a few of the commonly analyzed gene expressions in Breast Cancer. They are as follows:

- **Estrogen Receptor (ESR), Progesterone Receptor (PR), and HER2:**
  The hormone receptors PR and ESR are important players in breast cancer. Endocrine therapies that target hormone receptors are effective in treating breast cancers that express these receptors. In particular breast cancers, the receptor HER2 is overexpressed. HER2-directed therapies can be used to target breast cancers that are HER2-positive, which are linked to aggressive behaviour.
- **Homeobox (HOX) Genes:**
  HOX genes are linked to cancer, particularly breast cancer, and play important roles in embryonic development. The development, spread, and metastasis of breast cancer may be influenced by particular HOX genes. The regulatory networks impacted by HOX

genes and their potential applications as therapeutic targets and diagnostic markers are investigated.

- **Matrix Metalloproteinase (MMP) Genes:**
  MMPs are connected to the invasion and metastasis of cancer and play a role in the remodelling of the extracellular matrix. There is evidence of MMP dysregulation in breast cancer. Research looks into the functions of particular MMPs in the development of breast cancer as well as their potential as therapeutic targets and prognostic indicators.

Targeting the estrogen receptor (ESR) and progesterone receptor (PR), **endocrine therapy** is a fundamental component of the treatment of hormone receptor-positive breast cancers. These treatments, which are widely used in clinical practice, effectively block hormone-driven tumour growth. These treatments include aromatase inhibitors and selective estrogen receptor modulators (SERMs). Despite their achievements, problems like resistance still exist.

Triple Negative Breast Cancer (TNBC) lacks expression of ESR, PR, and HER2, making it challenging to target with traditional hormone therapies or HER2-directed treatments. TNBC is heterogeneous, and ongoing research aims to identify molecular subtypes and potential therapeutic targets within this subgroup.

.

4. Objectives for Phase I:

To conduct Differential Gene Expression (DGE) analysis on BRCA data based on ER status, elucidating the most differentially expressed genes and documenting the entire analytical pipeline for comprehensive insights into breast cancer subtypes.

5. Materials and Methods

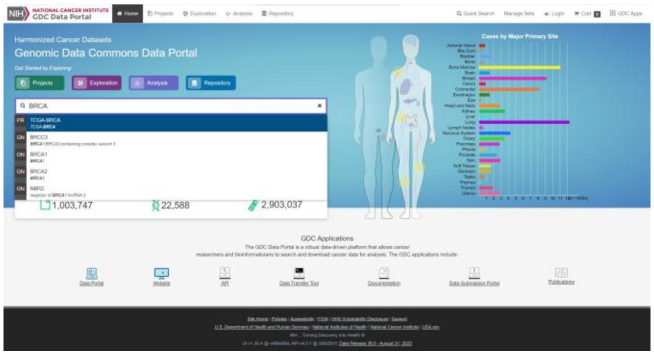
The entire code for the following pipeline can be found here.

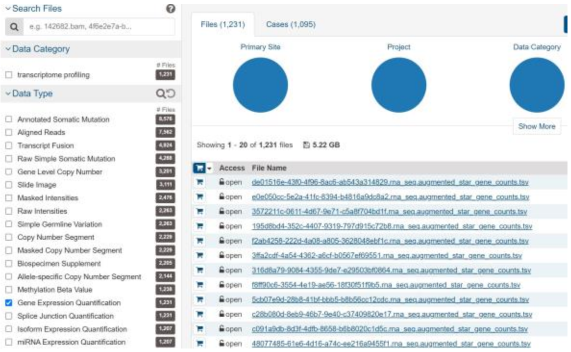
5.1 Dataset for Analysis (TCGA)

The Cancer Genome Atlas (TCGA) is a project to catalogue 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.





Data Category

	# Files
□ transcriptome profiling	1,231

Data Type

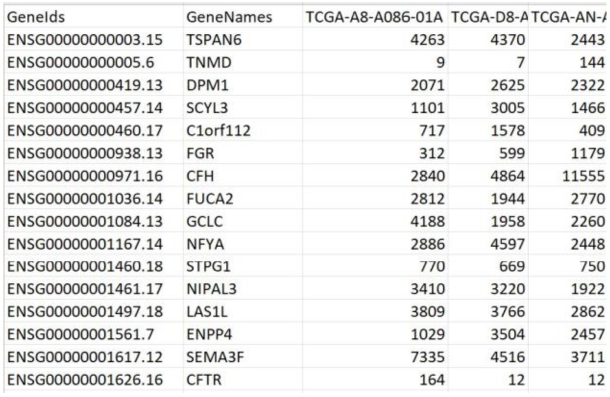
	# Files
□ Annotated Somatic Mutation	8,576
□ Aligned Reads	7,562
□ Transcript Fusion	4,924
□ Raw Simple Somatic Mutation	4,288
□ Gene Level Copy Number	3,201
□ Slide Image	3,111
□ Masked Intensities	2,476
□ Raw Intensities	2,263
□ Simple Germline Variation	2,263
□ Copy Number Segment	2,229
□ Masked Copy 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

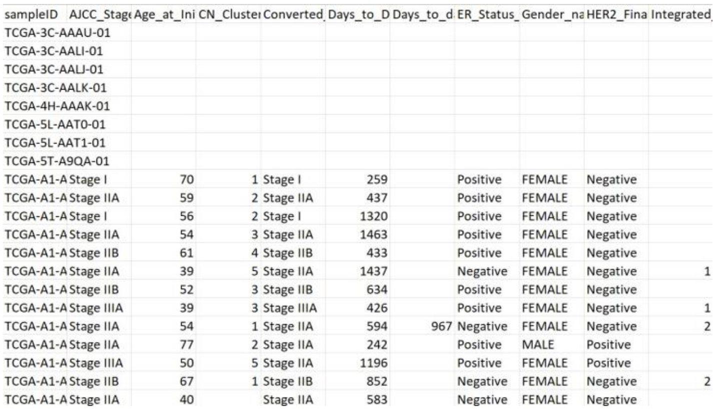
The count data in the TCGA BRCA dataset represents the quantification of gene expression levels through RNA sequencing. The structure of this dataset comprises genes as rows and samples as columns. These counts serve as the foundation for DGE analysis. This is obtained after combining the unstranded sequence counts from all the tsv files into a data frame. The resulting data frame looks as below:



GeneIds	GeneNames	TCGA-A8-A086-01A	TCGA-D8-A	TCGA-AN-/
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:

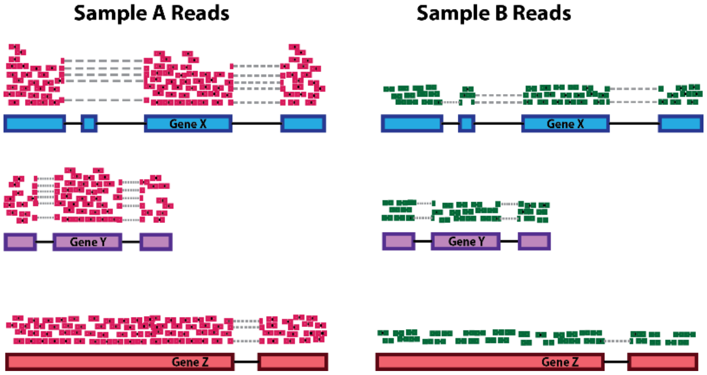


sampleID	AJCC_Stag	Age_at_Ini	CN_Cluste	Converted_	Days_to_D	Days_to_d	ER_Status_	Gender_na	HER2_Fina	Integrated_
TCGA-3C-AAAU-01										
TCGA-3C-AALI-01										
TCGA-3C-AALJ-01										
TCGA-3C-AALK-01										
TCGA-4H-AAAK-01										
TCGA-5L-AAT0-01										
TCGA-5L-AAT1-01										
TCGA-5T-A9QA-01										
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 Data Normalization

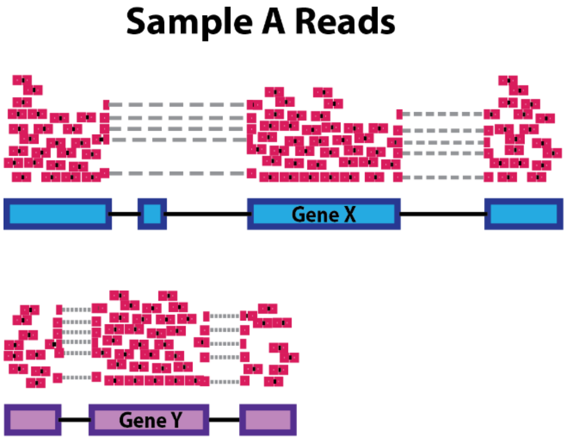
Data preprocessing is essential to tackle the following factors:

- **Sequencing depth:** The quantity of reads produced per sample, or sequencing depth, directly affects the measurement of gene expression. Consequently, to effectively evaluate changes in gene expression between samples, careful thought and normalization for sequencing depth variations are necessary.

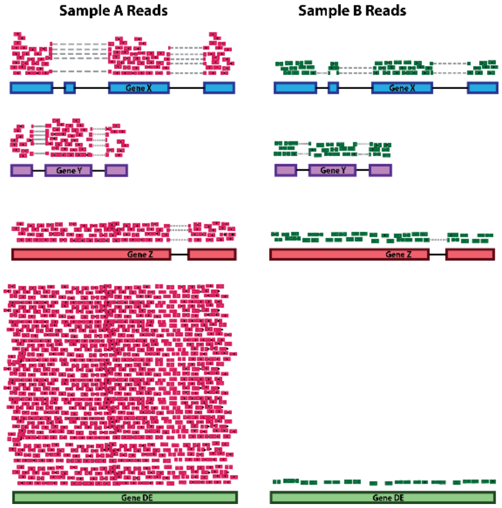


In the above image, we can observe that each gene appears to exhibit a twofold increase in expression in Sample A compared to Sample B. However, this apparent fold change is attributed to Sample A having double the sequencing depth of Sample B.

- **Gene length:** A potential confounding factor resulting from variations in gene length is revealed by comparing the levels of gene expression between Gene X and Gene Y. Because Gene X is longer than Gene Y, even though both have similar expression levels, the number of reads mapped to the former is much higher than the number mapped to the latter. This emphasizes the importance of considering gene length variations when interpreting data.

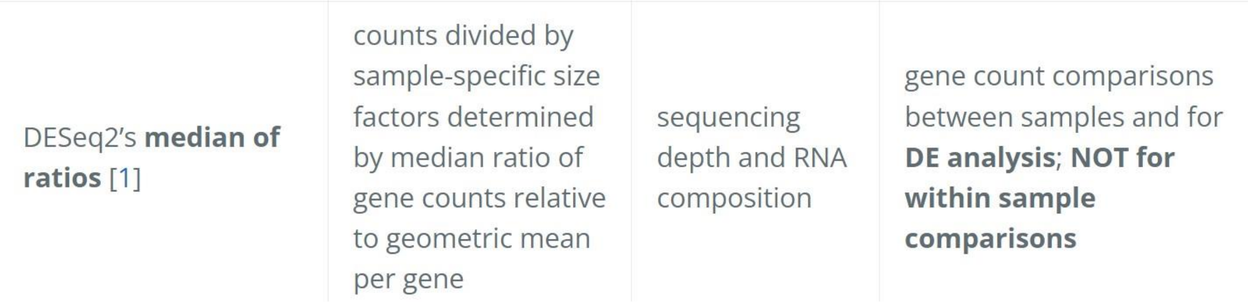


- **RNA composition:** If not adequately addressed, the presence of a differentially expressed (DE) gene in Sample A, which makes a significant contribution to the overall counts in the scenario described, can skew the normalized counts.



It can be difficult to normalize by the total number of counts when one sample has a disproportionate contribution from the DE gene to the total counts while the other does not. An accurate representation of gene expression profiles requires careful consideration of alternative normalization techniques, such as transcripts per million (TPM) or trimmed mean of M values (TMM), especially when genes with disproportionate contributions to the total RNA composition are present.

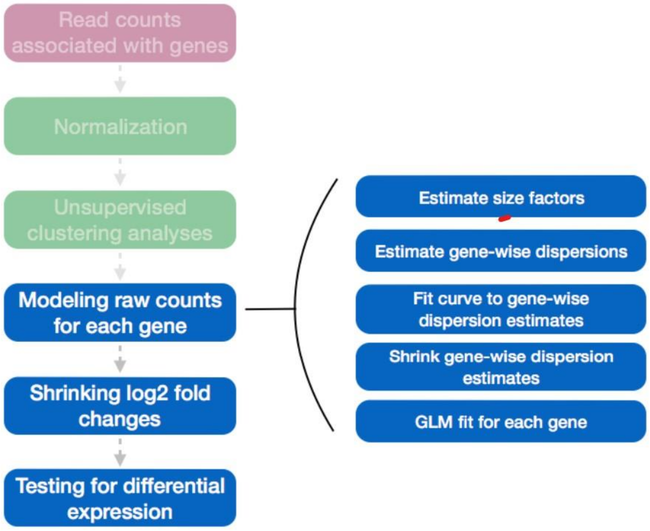
In R, data normalization can be achieved through DESeq2's **estimatesizefactors()** function. It uses a method called **median of ratios,** and it has the following features:



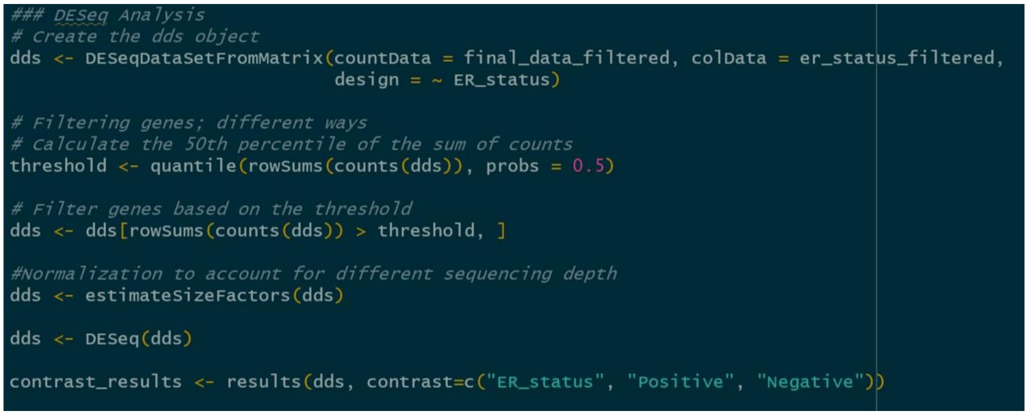
DESeq2’s **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**

5.3 DGE using DESeq2

- Differential expression analysis with DESeq2 involves multiple steps, as displayed in the flowchart, in blue.


Read counts associated with genes
Normalization
Unsupervised clustering analyses
Modeling raw counts for each gene
Shrinking log2 fold changes
Testing for differential expression
Estimate size factors
Estimate gene-wise dispersions
Fit curve to gene-wise dispersion estimates
Shrink gene-wise dispersion estimates
GLM fit for each gene

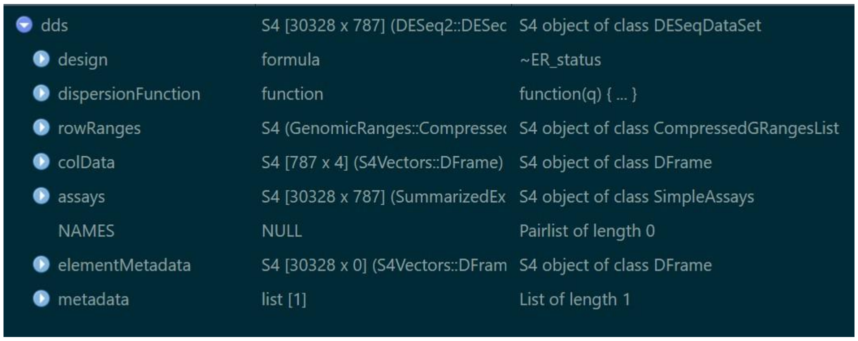
- All of this can be achieved by using just a few lines of code in DESEQ2 as follows:



```
### DESeq Analysis
# Create the dds object
dds <- DESeqDataSetFromMatrix(countData = final_data_filtered, colData = er_status_filtered,
                              design = ~ ER_status)

# Filtering genes; different ways
# Calculate the 50th percentile of the sum of counts
threshold <- quantile(rowSums(counts(dds)), probs = 0.5)

# Filter genes based on the threshold
dds <- dds[rowSums(counts(dds)) > threshold, ]

#Normalization to account for different sequencing depth
dds <- estimateSizeFactors(dds)

dds <- DESeq(dds)

contrast_results <- results(dds, contrast=c("ER_status", "Positive", "Negative"))
```

We ran the DGE analysis by contrasting ER+ve and ER-ve values from the clinical data.

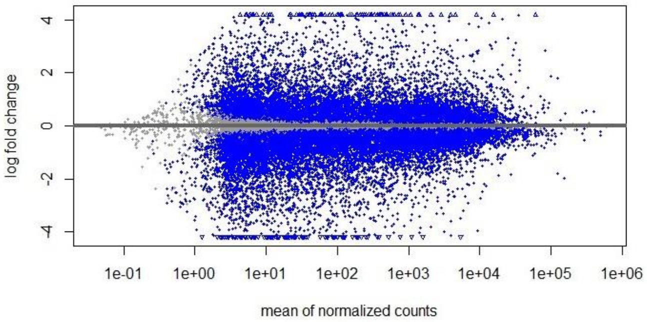
- The resultant object obtained upon running the code is as follows:



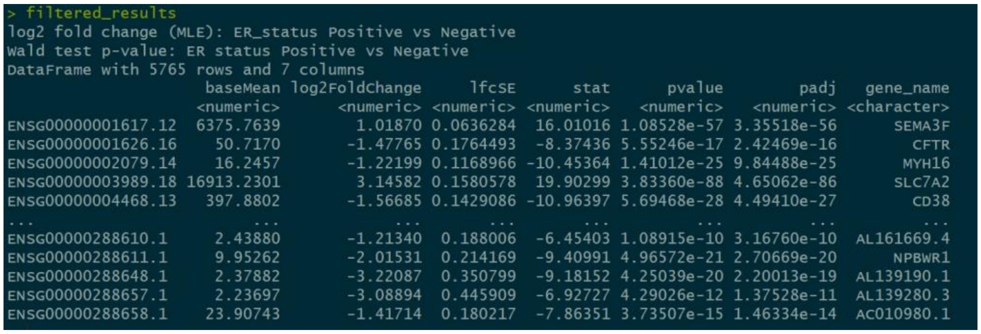
```
dds                   S4 [30328 x 787] (DESeq2::DESec   S4 object of class DESeqDataSet
  design              formula                           ~ER_status
  dispersionFunction  function                          function(q) { ... }
  rowRanges           S4 (GenomicRanges::Compressec     S4 object of class CompressedGRangesList
  colData             S4 [787 x 4] (S4Vectors::DFrame)  S4 object of class DFrame
  assays              S4 [30328 x 787] (SummarizedEx    S4 object of class SimpleAssays
  NAMES               NULL                              Pairlist of length 0
  elementMetadata     S4 [30328 x 0] (S4Vectors::DFram  S4 object of class DFrame
  metadata            list [1]                          List of length 1
```

Observe that we have cut short the number of genes to half (~30k) by removing genes whose row counts were below the 50th percentile mark so as to facilitate faster processing.

- An MA plot is a type of Bland-Altman plot used in computational biology to represent genomic data visually. The plot illustrates the variations between measurements made in two samples by converting the data onto the M (log ratio) and A (mean average) scales and then plotting these values. MA plot for the above object is as follows:



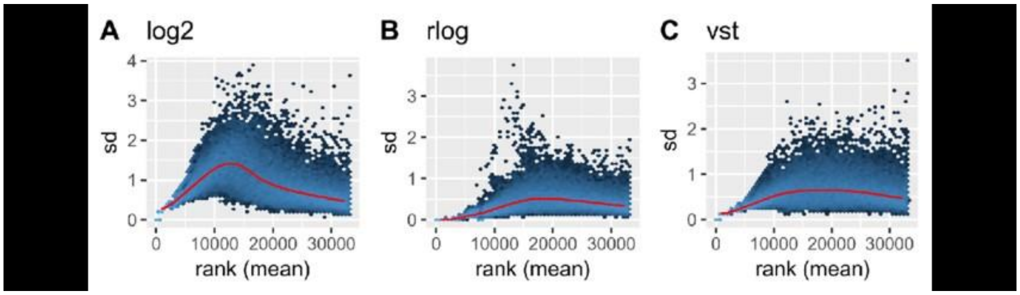
- Now, we need to extract the genes which have significant differential expression and are statistically sound. We set the |log2FC| at >= 1, which means more than or equal to a 2-fold differential expression. Additionally, we set P_{adj}<0.01. We obtained around 5765 genes. The table is as follows:



```
> filtered_results
log2 fold change (MLE): ER_status Positive vs Negative
Wald test p-value: ER status Positive vs Negative
DataFrame with 5765 rows and 7 columns
                     baseMean log2FoldChange     lfcSE      stat      pvalue        padj   gene_name
                    <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric> <character>
ENSG00000001617.12  6375.7639        1.01870 0.0636284  16.01016 1.08528e-57 3.35518e-56      SEMA3F
ENSG00000001626.16    50.7170       -1.47765 0.1764493  -8.37436 5.55246e-17 2.42469e-16        CFTR
ENSG00000002079.14    16.2457       -1.22199 0.1168966 -10.45364 1.41012e-25 9.84488e-25       MYH16
ENSG00000003989.18 16913.2301        3.14582 0.1580578  19.90299 3.83360e-88 4.65062e-86      SLC7A2
ENSG00000004468.13   397.8802       -1.56685 0.1429086 -10.96397 5.69468e-28 4.49410e-27        CD38
...                       ...            ...       ...       ...         ...         ...         ...
ENSG00000288610.1     2.43880       -1.21340  0.188006  -6.45403 1.08915e-10 3.16760e-10  AL161669.4
ENSG00000288611.1     9.95262       -2.01531  0.214169  -9.40991 4.96572e-21 2.70669e-20      NPBWR1
ENSG00000288648.1     2.37882       -3.22087  0.350799  -9.18152 4.25039e-20 2.20013e-19  AL139190.1
ENSG00000288657.1     2.23697       -3.08894  0.445909  -6.92727 4.29026e-12 1.37528e-11  AL139280.3
ENSG00000288658.1    23.90743       -1.41714  0.180217  -7.86351 3.73507e-15 1.46334e-14  AC010980.1
```

5.4 Downstream Analyses

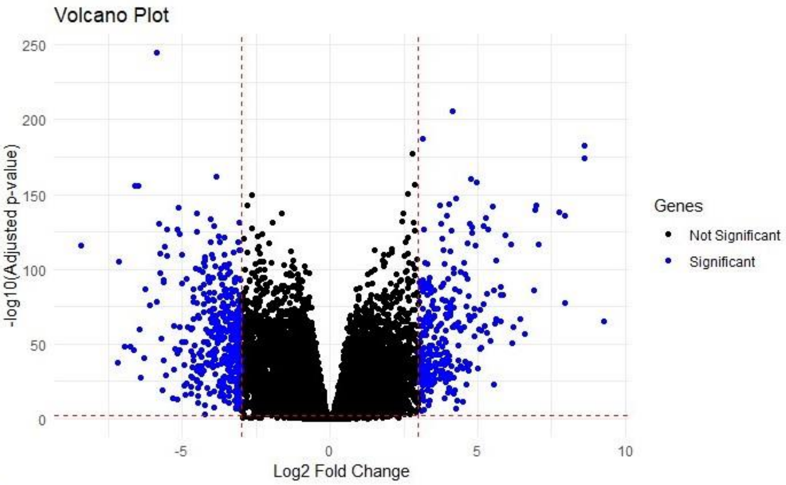
After running DESeq2 to identify differentially expressed genes (DEGs), several downstream analyses are typically performed to gain insights into the biological significance of the identified DEGs. However, before performing downstream analyses, variance stabilization must be done to remove the variance's dependence on the mean. Three commonly used methods are Variance Stabilizing Transformation(VST), rlog transformation and log2 transformation. We will be using VST for our purposes because it is faster and works on bigger datasets.



Here are some standard downstream analyses following DESeq2:

- **Volcano Plot**

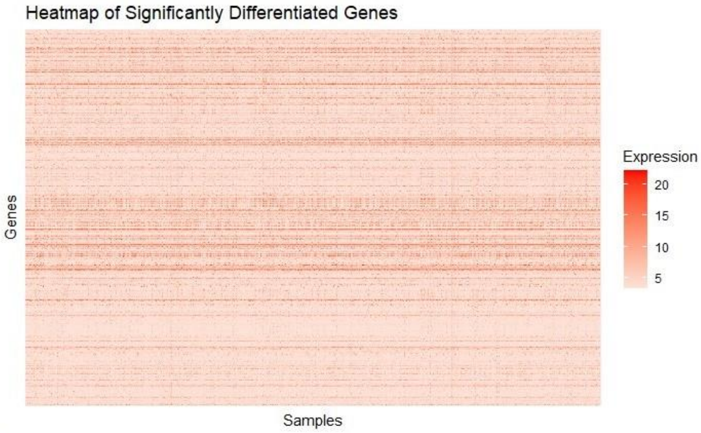
Folding change versus statistical significance is plotted in a volcano plot to visualize DEGs. Genes with notable changes in expression are identified with the aid of this plot.



Here, we set |log2FC| at >= 3 for better visualization purposes.

- **Heatmaps**

  Construction of heatmaps to visualize the expression patterns of DEGs across samples or conditions. Heatmaps facilitate the identification of gene clusters and patterns of co-expression.



Clustering can be done, and dendrograms can be plotted, but higher computing efficiency is required.

- **Additional analyses**

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

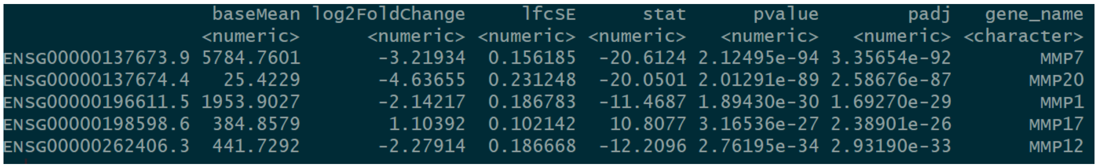
  - **Functional Enrichment Analysis:** Utilizing pathway enrichment analysis or gene ontology (GO) tools, one can determine the biological processes, pathways, and functions linked to the DEGs.

    - **Gene Set Enrichment Analysis (GSEA):** The ranked list of DEGs is subjected to GSEA analysis to determine if predefined gene sets are enriched, such as those linked to particular pathways or biological processes.
    - **Network Analysis:** Building gene interaction networks to find important gene modules or hubs that exhibit coordinated changes in expression. Functional modules and possible regulatory relationships can be found via network analysis.
    - **Survival Analysis:** Clinical and gene expression data are integrated to perform survival analysis, which evaluates the relationship between a given gene's expression and patient outcomes.

6.0 Results

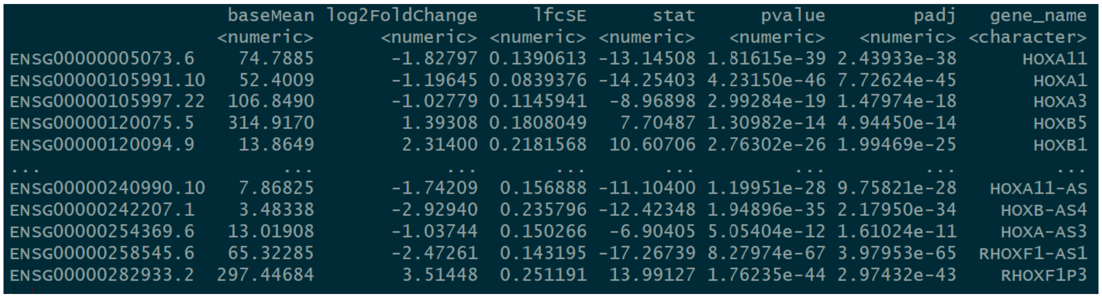
We obtain a list of the most prominently differentially expressed (|log2FC|>=1 & P_{adj}<0.05) MMPs and HOX genes.

- We get 5 MMPs, as listed in the table below.



```
                    baseMean log2FoldChange     lfcSE      stat      pvalue        padj   gene_name
                   <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric> <character>
ENSG00000137673.9  5784.7601       -3.21934  0.156185  -20.6124 2.12495e-94 3.35654e-92        MMP7
ENSG00000137674.4    25.4229       -4.63655  0.231248  -20.0501 2.01291e-89 2.58676e-87       MMP20
ENSG00000196611.5  1953.9027       -2.14217  0.186783  -11.4687 1.89430e-30 1.69270e-29        MMP1
ENSG00000198598.6   384.8579        1.10392  0.102142   10.8077 3.16536e-27 2.38901e-26       MMP17
ENSG00000262406.3   441.7292       -2.27914  0.186668  -12.2096 2.76195e-34 2.93190e-33       MMP12
```

- We get 18 HOX genes, as listed in the table below.



```
                    baseMean log2FoldChange     lfcSE      stat      pvalue        padj   gene_name
                   <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric> <character>
ENSG00000005073.6    74.7885       -1.82797 0.1390613 -13.14508 1.81615e-39 2.43933e-38      HOXA11
ENSG00000105991.10   52.4009       -1.19645 0.0839376 -14.25403 4.23150e-46 7.72624e-45       HOXA1
ENSG00000105997.22  106.8490       -1.02779 0.1145941  -8.96898 2.99284e-19 1.47974e-18       HOXA3
ENSG00000120075.5   314.9170        1.39308 0.1808049   7.70487 1.30982e-14 4.94450e-14       HOXB5
ENSG00000120094.9    13.8649        2.31400 0.2181568  10.60706 2.76302e-26 1.99469e-25       HOXB1
...                      ...            ...       ...       ...         ...         ...         ...
ENSG00000240990.10   7.86825       -1.74209  0.156888 -11.10400 1.19951e-28 9.75821e-28   HOXA11-AS
ENSG00000242207.1    3.48338       -2.92940  0.235796 -12.42348 1.94896e-35 2.17950e-34    HOXB-AS4
ENSG00000254369.6   13.01908       -1.03744  0.150266  -6.90405 5.05404e-12 1.61024e-11    HOXA-AS3
ENSG00000258545.6   65.32285       -2.47261  0.143195 -17.26739 8.27974e-67 3.97953e-65  RHOXF1-AS1
ENSG00000282933.2  297.44684        3.51448  0.251191  13.99127 1.76235e-44 2.97432e-43    RHOXF1P3
```

7.0 Conclusion & Future work

Finally, by identifying five Matrix Metalloproteinase (MMP) genes and eighteen Homeobox (HOX) genes as significantly differentially expressed, this thesis project provides new insights into Breast Cancer (BRCA). With MMP genes emphasizing their role in tumour microenvironment remodelling and cancer invasion and HOX genes revealing their involvement in regulating crucial cellular processes, these findings shed light on potential key players in BRCA progression.

Furthermore, we intend to study the roles of these genes more closely by observing their behaviour in samples in our own lab obtained from local hospitals. This work establishes the foundation for future studies into the functional roles of MMP and HOX genes, providing prospective biomarkers and therapeutic targets for a more complex understanding of BRCA and the creation of targeted treatments. It goes beyond simply listing differentially expressed genes.

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://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html
- https://rdrr.io/bioc/DESeq2/man/varianceStabilizingTransformation.html