Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)
A Subsidiary organ of the Organization of Islamic Cooperation (OIC)
Dhaka,Bangladesh

Bioinformatics Analysis of Deferentially Expressed Genes in Breast Cancer Using DESeq2

Authors
Sow Bocar Amadou Malick (170041073)
Fatoumatta Conteh (170041082)
Muhammed Sawo (170041083)

Supervisor
Tareque Mohmud Chowdhury
Assistant Professor
Department of Computer Science and Engineering

A Thesis submitted to the Department of Computer Science and Engineering in Partial fulfillment of Bachelor of Science in Computer Science and Engineering (CSE)

April, 2022

DECLARATION

We hereby declare that this thesis titled "Bioinformatics Analysis of Differentially Expressed Genes in Breast Cancer Using DESeq2" is an authentic report our study carried out as requirement for the award of degree Bachelor of Science in Computer Science and Engineering at the Islamic University of Technology, Gazipur, Dhaka, under the supervision of Tareque Mohmud Chowdhury, Assistant Professor, CSE, IUT in the year 2022.
The matter embodied in this thesis has not been submitted in part or full to any other institute for award of any degree.

Authors

14-5-22
....
Muhammed Sawo (170041083)

14-05-22
....
Fatoumatta Conteh (170041082)

MBSow 14-5-22
....
Sow Bocar Amadou Malick (170041073)

CERTIFICATE OF RESEARCH

The thesis titled, "BIOINFORMATICS ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES IN BREAST CANCER USING DESEQ2" submitted by Muhammed Sawo (170041083), Fatoumatta Conteh (170041082), Sow Bocar Amadou Malick (170041073) has been accepted as satisfactory in partial fulfillment of the requirement for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervisor

13.05.2022

. ...

Tareque Mohmud Chowdhury
Assistant Professor
Department of Computer Science and Engineering

April, 2022

Acknowledgment

We would like to express our gratitude to God Almighty for guiding us through all of our challenges. Day by day, we have felt your guiding. You are the one who allowed us to complete our degree. We will continue to put our faith in you for our future.

We would also like to express our gratitude to our supervisor Tareque Mohmud Chowdhury, who enabled us to complete this work. His instruction and counsel guided us through all stages of this essay writing. We also like to thank the committee members for making our defense a pleasant experience and for their insightful comments and ideas.

Finally, we would like to thank our parents and families as a whole whom the Almighty chooses for their moral support, endless prayers and motivation they have rendered throughout the course of our research. We are always grateful to our families and friends for their support.

Contents

Chapter 1

1.1 Introduction

Differential Gene Expression Analysis is a strong tool for determining if genes in two or more sample groups are expressed at significantly different levels. To estimate gene counts and identify deferentially expressed genes, we'll utilize the DESeq2 software. Also, while determining whether genes are deferentially expressed, we must account for variation in the data. The purpose is to see if differences between groups are substantial for each gene, given the biological differences between biological replicates. Using Normalized to Read Count Data (NRCD) and statistical analysis, DEG analysis was used to find quantitative differences in expression levels between experimental groups. For example; statistical testing is used to decide whether for a given gene and observed difference in read counts is significant. I.e., whether it is greater than what would be expected just due to natural random variation. The analysis requires gene expression values to be compared between sample group types. The goal is to determine which genes are expressed at different levels between conditions. It has become a widely used technology that allows for effective genome-wide relative gene expression quantification, and it is the method of choice for identifying deferentially expressed genes between two or more biological situations of interest. The primary challenges surrounding such DE analysis have been highlighted from the start, and several methodologies and tools have been offered in the relevant literature. One of the most difficult aspects of this study, as with any other statistical research, has been determining the probabilistic model that best fits the data, as well as the model's optimal parameter estimates. Another significant challenge was the requirement for data normalization in order to appropriately compare two biological situations by analyzing and removing any potential technological and/or biological biases. Last but not least, several research have emphasized the practical requirement to determine the ideal number of biological replicates per condition and the optimal library size. We'll go over the use of DeSeq2 method as a utilized methodology and tools for DE analysis in this article.

The gene outcomes can offer biological insights into processes affected by the conditions. greater than what would be expected just due to natural random variation.

1.2 Problem Statement

The discovery of differentially expressed (DE) genes is one important effort in this area. However, RNAseq data is highly diverse and contains a substantial number of zero counts, making DE gene detection difficult. To address these issues, new techniques beyond the traditional ones based on a nonzero difference in average expression are required. Several approaches for analyzing differential gene expression in RNAseq data have been developed. It is vital to analyze and compare the efficacy of differential gene expression analysis methods for RNAseq data in order to provide assistance on picking an appropriate tool or developing a new one.

The three components of differential gene expression analysis of RNA-seq data are normalization of counts, parameter estimates of the statistical model, and testing for differential expression. We've taken a dataset and want to apply the three (3) processes listed above to it. In this section, we give a basic summary of the approaches utilized by the numerous algorithms that carry out these three steps. Although some of the programs may test for multi-class differences or multi-factored investigations including a variety of biological settings and sequencing techniques, we will focus on the most common scenario of comparing differential expression between two cellular states or phenotypes.

1.3 Research Challenges

- **Read Mapping:** Finding original read sources in reference genomes is a necessary with the provided collection of data, but it comes with computational problems, as well as numerous trade-offs between speed and accuracy that result in multiple sub-optimal solutions.

- **Count Computation:** Estimated gene expression with count is a must, given the amount of reads mapped on each gene, but deciding on the appropriate approach to establish gene boundaries and handle reads mapping on multiple genes is indeed a problem.

- **Count Normalization:** It entails removing biases between and among samples, but gene-specific biases are difficult to eliminate, which can have an impact on downstream analyses.

- **Differential Expression Analysis:** This is our fundamental goal; real differential expression can only be found using accurate and powerful testing, as well as appropriately represented technological and biological variability.

1.4 Motivation

- **Novelty:** You might find that some other genes have an unanticipated function in the process we're looking into. A novel concept is one that is unique in the topic or scope being studied. It could be a new methodology or design that paves the way for new information to emerge. It could be a method of attempting to add to the present knowledge base with the intent of doing so. In general, it's a feature of research when it takes a topic that has already been the subject of previous studies and gives it a fresh and unique twist. Scholars can achieve this by altering elements such as the design of the study, the location of past studies, or the demography of previous studies, or by completely modifying the database. Doing in-depth exploratory research and comparing your idea to what is currently out there on the subject is the greatest method to tell if your idea is original or not.

- **Context:** When you can compare the activity of one gene to that of others, it becomes easier to interpret its significance. It provides the frames of reference through which the study, as well as its methodological techniques, arguments, findings, conclusions, and recommendations, can be viewed.

- **Systems:** That is to gain system levels understanding of the process by measuring all genes simultaneously so that we can identify which biological pathways are important.

1.5 Scope

A large number of computer techniques have been developed for analyzing differential gene expression in RNA-seq data. Our goal and focus with DESeq2 is only on differential gene expression. We analyze the stages below and report the differences between what we discover of what was not found in any other paper and that can help scientists with cancer therapy[1].

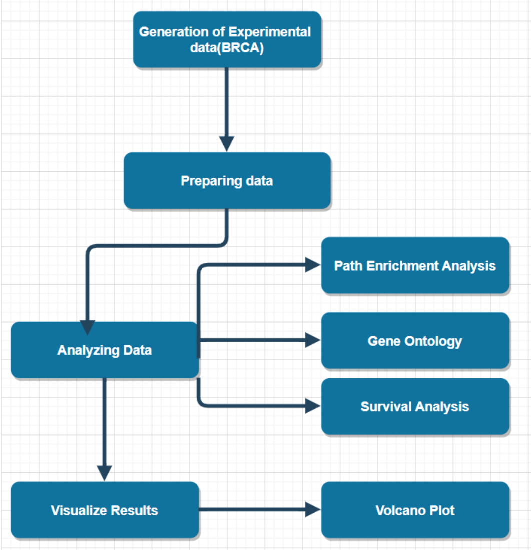


Figure 1.1: DEG steps

1.6 Research Contribution

When cells respond to diverse treatments/conditions, the discovery of DE genes aids in the study of gene function. DE gene detection can also be employed as a pre-processing step before grouping gene expression patterns or determining gene set enrichment. There are presently no standard approaches for finding DE genes using such data because to the limited history of RNA-seq and its continual development. Many statisticians have been working on this problem for a long time. Thus, statistical techniques for differential gene expression analysis, such as edgeR, DESeq, baySeq, and a two-stage Poisson model method, are available (TSPM)[3]. All of these statistical procedures are based on assumptions and conditions that are only partially met.Different methodologies rely on different assumptions, and as a result, different perceptions of reality may be captured.The various outcomes are not mutually exclusive. One of them, neither of them, or all of them could be correct at the same moment! Which is a little difficult to comprehend.As a result, we used DESeq2 for our differentially expressed genes. It helps to reduce the number of genes tested by removing genes that are unlikely to be DE before testing, such as those with low counts and outlier samples. A huge deal of comparisons were done for each of our analysis processes in order to achieve uniqueness among all the other analyses out there. Starting with the dataset used and performing of DESeq2 on it.

1.7 Thesis Outline

Our research is focused on identifying genes that are differently expressed. Techniques for testing for differential expression using negative binomial generalized linear models, such as data-driven prior distributions for dispersion and logarithmic fold changes, are included in the DESeq2 package. To begin, we used queries to download TCGA-BRCA (BReast CAncer gene) and relevant data from TCGA, then pre-processed the data with Bioconductor bioinformatics (tools) packages.Then, utilizing those packages, we ran downstream analyses such pathway enrichment, gene ontology, and survival analysis to determine the differentially expressed genes (DEG) of Breast Cancer genes. We evaluated and explored the relationship between samples using the DESeq2 program, did differential gene expression analysis, and visually explored the results.

1.8 Feature Selection

Gene expression data at the transcriptome level can show gene activity and physiological status in a biological system. In gene expression data, small samples with large dimensions and noise are prevalent. Even though a single gene chip or next-generation sequencing technology can detect tens of thousands of genes in a single sample, only a few gene groupings are connected to specific diseases or biological processes. Furthermore, due to the over-fitting problem, evaluating these duplicated genes not only consumes a lot of search space, but it also slows down data mining speed. Medicine has struggled to identify disease-related genes from the raw gene expression data. Additionally, discovering disease-related genes will aid in the development of relevant treatments.

Implementing knowledge-based interpretations to evaluate omics data can provide vital information about a variety of biological processes as well as represent the current physiological condition of cells and tissues. With a high number of genes and small samples, the most difficult part of analyzing gene expression data is extracting disease-related information from a tremendous amount of redundant data and noise. A significant step in addressing this problem has been gene selection[7], which involves removing duplicated and less significant genes.

1.9 Feature Selection Technique

DEGs are useful not only to doctors for diagnosing patients, but also to pharmaceutical companies for identifying genes that can be targeted for medications. During the previous few years, various strategies for feature or gene selection have been presented, including filter, wrapper, embedded, and, more recently, ensemble approaches.

We employ a filtering approach that evaluates the discriminative strength of features only on the data's inherent qualities[4].As a general rule, this technique calculates a relevance score and selects the highest-scoring features/genes using a threshold system. This category of strategies is independent of any categorization methodology, yet they may provide a more optimum set of features under certain scenarios. Using this method,let $S^{x*y} = \{S_{i,j}\}$ be a matrix comprising x genes and y samples from distinct groups specified by a target annotation, (Sx * y = [S 1x*y 1,S 2x*y 2,......S px*y p), where each matrix (S ix * y i) contains samples from the same group and (y 1+y 2 +......y p =y) The most informative genes are chosen from a subset of genes that are the most discriminative for the defined classes throughout the entire population of samples (Sk*nin Xx*y, kll m). The four steps of the ranking process of the filter approach cite6152088 are as follows:

1. We use the scoring function X(s) to calculate the difference in expression between different sets of samples and rank features/genes in decreasing order of calculated scores as fold change ratio. A high score is thought to indicate a DEG. The log2 of the ratio between the means of groups of dataset of genes is being calculated using DESeq2 tool. The numbers of which some are negative and some positive, maintaining the expression change's directionality.

2. The p-value, or statistical significance of the estimated scores, was determined. In statistics, the p-value denotes the probability of receiving a test statistic that is at least as bad as the one that was observed. The higher the p-value, the more significant the result (in

the sense of statistical significance). As a result, our cutoff levels were set at 0.05, which corresponds to a 5% chance of the tested hypothesis being accepted by chance.

3. We picked the some of the most informative genes among the top ranked genes that are statistically significant and alternatively, select some of the low ranked genes.

4. Finally confirm the genes selected and perform the rest of the DGE steps on them.

Chapter 2

Proposed Methodology

2.1 Our Proposed Approach

2.1.1 DESeq2

Below are the most basic steps for a differential expression analysis. A number of processes downstream of DESeq2 lead to the generation of counts or estimated counts for each sample. DESeqDataSet was used to import quantification data, and tximeta was used to build a SummarizedExperiment with additional metadata ().

- Count data in the form of a matrix of integer values, such as from RNA-seq or another high-throughput sequencing experiment, is expected as input by the DESeq2 package. The number in the matrix's i-th row and j-th column reflects how many readings from sample j may be assigned to gene i. The matrix should contain unnormalized or approximated numbers of sequencing reads (for single-end RNA-seq) or fragments (for multi-end RNA-seq).

- There are various ways to make count matrices. Because only count values can reliably estimate measurement precision, it's necessary to provide count matrices as input to DESeq2's statistical model (Love, Huber, and Anders 2014). Because the DESeq2 model internally corrects for library size, transformed or normalized values, like as counts scaled by library size, should not be used as input.

- The DESeqDataSet class is a technical detail that extends the RangedSummarizedExperiment class from the SummarizedExperiment package. The word "Ranged" refers to how assay data rows (in this case, integers) can be connected to chromosomal ranges (the exons of genes). This link allows you to use range-based capabilities provided by other Bioconductor tools to further analyze the data (e.g., discover the nearest ChIP-seq peaks to the differentially expressed genes).

- A DESeqDataSet object must be linked to a design formula. The design formula expresses the variables that will be used in modeling. The variables in the formula with plus signs between them should be preceded by a tilde (). (it will be coerced into an formula if it is not already). Although the design formula is used to estimate the model's dispersions and log2 fold changes, all differential analysis processes must be completed because the design formula is used to estimate the model's dispersions and log2 fold changes..

To determine if the sample order is consistent, we examine the count matrix and column data.

- The columns of the count matrix and the rows of the column data (information about samples) must be in the same order. DESeq2 will not guess which column of the count matrix corresponds to which row of column data; this information must be provided in a consistent order to DESeq2.
- We'll have to change one or the other to make the sample order consistent, as they're not in the correct order as supplied (if we do not, later functions would produce an error). The "fb" must also be removed from the row.

2.1.2 Input from the SummarizedExperiment

If you already have a SummarizedExperiment, you can quickly import it into DESeq2 by following the steps below. We begin by loading the package that contains the TCGA-BRCA dataset.

2.1.3 Pre-Filtering

- While pre-filtering low-count genes before executing the DESeq2 algorithms is not required, it is useful for two reasons: by deleting rows with very few reads, the dds data object's memory size is reduced, and the transformation and testing operations within DESeq2 are faster. Here, we perform a short pre-filtering to keep only rows with at least 10 total reads. It's worth mentioning that more stricter filtering is conducted automatically to enhance power via independent filtering on the mean of normalized counts within the dataset.
- The results function of the DESeq2 package conducts independent filtering by default, utilizing the mean of normalized counts as a filter statistic. A filter statistic threshold is used to maximize the number of adjusted p values lower than a significance level alpha (we use the conventional variable name for significance level, despite the fact that it has nothing to do with the dispersion parameter).
- The genefilter package's filtered p function is used to do the default independent filtering, and all of filtered p's parameters can be passed to the results function. The filter threshold value and the number of rejections at each quantile of the filter statistic are exposed as metadata of the object returned by results.

2.1.4 The Model DESeq2

- In our work (Love, Huber, and Anders 2014), we go over the DESeq2 model and all of the software methods in detail, and we also include the formula and descriptions in this part. The differential expression analysis in DESeq2 uses an expanded linear model of the form:

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = x_{j.}\beta_i$$

- where the gene counts $K_{ij}K_{ij}$ A negative binomial distribution with fitted mean ijij and a gene-specific dispersion parameter ii is used to model sample I, and a negative binomial distribution with fitted mean $ijij$ and a gene-specific dispersion parameter (ii) is used to model sample j. The fitted mean is made up of a sample-specific size factor s_js_j and a parameter $q_{ij}q_{ij}$ proportional to the expected true concentration of fragments for sample j. The coefficients ii represent the log2 fold changes for gene I in each column of the model matrix XX. It should be noted that the model can be expanded to accommodate sample data.[8]

- and normalizing parameters that are reliant on genes $s_{ij}s_{ij}$ The dispersion parameter $_{ii}$ defines the relationship between the variance of the observed count and its mean value. To put it another way, how far do we expect the observed count to differ from the mean number, which is controlled by the size factor s_js_j as well as the covariate-dependent portion $q_{ij}q_{ij}$, as shown above.

$$\mathrm{Var}(K_{ij}) = E[(K_{ij} - \mu_{ij})^2] = \mu_{ij} + \alpha_i\mu_{ij}^2$$

- DESeq2 can yield maximum a posteriori estimates of the log2 fold changes in ii after integrating a zero-centered Normal prior (betaPrior). These moderated, or shrunken, estimates, which were formerly generated by the DESeq or nbinomWaldTest functions, are now generated by the lfcShrink function. The Cox-Reid adjusted profile likelihood is used to assess dispersions using anticipated mean values using the maximum likelihood estimate of log2 fold changes and maximizing the Cox-Reid adjusted profile likelihood, which was first implemented for RNA-seq data in edgeR (Cox and Reid 1987,edgeR GLM). The steps of the DESeq function are documented on its manual page? In a nutshell, DESeq stands for:
  - estimation of size factors sjsj by estimateSizeFactors
  - estimation of dispersion ii by estimateDispersions
  - negative binomial GLM fitting for ii and Wald statistics by nbinomWaldTest

Chapter 3

Analysis of Experiments and Result Comparison

3.1 Data Set Details

- The Breast Invasive Carcinoma (TCGA-BRCA) data collection is part of a larger effort to build a research community focused on connecting cancer phenotypes to genetics by giving clinical images matched to The Cancer Genome Atlas individuals (TCGA). The Cancer Imaging Archive stores radiological data, whereas the Genomic Data Commons (GDC) Data Portal houses clinical, genomic, and pathological data (TCIA).

- Using matching TCGA patient identities, researchers can search the TCGA or TCIA databases for links between tissue genotype, radiological phenotype, and patient outcomes. Tissues for TCGA were acquired from a variety of locations throughout the world in order to meet their accrual goals, which were typically 500 specimens per cancer type.

- In this work, we discovered a total of 1098 samples in the data set, including 19947 genes. After we limit the data, we have 13952 Differentially Expressed Genes, 2026 of which are paired samples.Significant genes with pvalue of 0.1 and LFC of 1 is 13952 genes and for a pvalue of 0.01 we have a total of 12892 genes of which 12517 are significant.

3.2 Performance Analyses

The first case study is about (BRCA downstream analysis with gene expression) In order to uncover differentially expressed genes that might have a role in survival, we utilized TCGA-query, TCGAdownload, and TCGAprepare to download 114 normal and 1097 breast cancer (BRCA) samples for this case study. Using TCGAanalyze DEA, we discovered 3390 DEGs (log fold change 1 and FDR 1 percent) between the 114 normal (NT) and 1097 BRCA (TP) samples. To uncover the underlying biological process, we utilized the TCGAanalyze EA full tool to perform an enrichment analysis on the DEGs (Figure 3A–C). In a bar chart created by TCGAbiolinks, the numbers of genes ascribed to the main categories of three ontologies are displayed (GO: biological process, GO: cellular component and GO: molecular function). Furthermore,

Using a Kaplan-Meier analysis to construct univariate survival curves and a log-ratio test to determine statistical significance, the TCGAanalyze SurvivalKM function revealed 555 genes whose expression changed significantly with P-values less than 0.05. The TCGAnalyze SurvivalCoxNET function was used to do a Cox regression analysis and produce Cox P-values to establish statistical significance in order to create multivariate survival curves. The multivariate survival analysis identified 160 genes to be significant, according to the Cox P-value FDR = 0.05. These genes were found to be strongly associated to survival in both univariate and multivariate investigations, and this gene collection was used in the network analysis that followed.

3.2.1 DESeq2 was used to do a differential analysis of count data.

On high-dimensional count data, the DESeq2 package is used to normalize, display, and perform differential analysis. Empirical Bayes approaches are used to generate posterior estimates for log fold change and dispersion, as well as priors for these parameters..
Using the results function to generate tables containing log2 fold change, p-values, corrected p-values, and other information for each gene is recommended.

3.2.2 Findings of a DESeq study

- ***resultsNames***removeResults provides a DESeqDataSet object with the results columns deleted; yields the names of the model's estimated effects (coefficents); description results returns the names of the estimated effects (coefficents) of the model; resultsNames returns the names of the estimated effects (coefficents) of the model; resultsNames returns the names of the estimated effects (coefficents) of the model; resultsNames returns the names of the estimated effects (coefficents) of the model; resultsNames returns the names of the estimated effects (coefficents) of the model; resultsNames returns the names of the estimated effects (coefficents)

- ***Details*** When the results table is displayed, it will include details about the comparison, such as "log2 fold change (MAP): condition treated vs untreated," showing that the estimates are of log2(treated / untreated), as returned by contrast=c ("condition","treated","untreated"). For studies that require more than a simple two-group comparison, several findings can be generated; as a result, results use the contrast and name parameters to allow the user to select which comparisons they want to print in a results table. The contrast argument should be used for detailed specification of the levels to be compared as well as their order. The last level of the last variable in the design formula will be compared to the first level of this variable if you run the results without supplying contrast or a name.[13].

- ***Individual effects***,The argument name can be used to generate which must be individual components of resultsNames (object). Continuous variables, individual level effects, and individual interaction effects are all examples of individual effects. The results object contains information on the comparison that was used to generate the results table, as well as the statistical test that was used to get the p-values (Wald test or likelihood ratio test). This information may be found in the metadata columns of the results table, which can be accessed by calling mcols on the DESeqResults object returned by results.

- **On p-values**: Independent filtering is employed by default to choose a set of genes for multiple test correction, which optimizes the number of adjusted p-values fewer than a predetermined critical value alpha (by default 0.1). The filter for maximizing the number of rejections is the mean of normalized counts for all samples in the dataset. Several options from the genefilter package's filtered p function (used within the results function) are supplied here to tailor the independent filtering behavior.[10].

- In DESeq2 version >= 1.10, the threshold is set at the lowest filter quantile for which the number of rejections is close to the peak of a curve fit to the number of rejections over the filter quantiles. "Close to" is defined as being within one residual standard deviation. The corrected p-values for genes that do not pass the filter threshold are set to NA.

- Genes with count outliers are given a p-value of NA by default, as determined by Cook's distance. The cooksCutoff parameter can be used to control this behavior. Cook's distances for each sample are stored as a matrix "cooks" in the assays() list. This metric can be used to find rows with observed counts that do not fit into a Negative Binomial distribution.

- For analyses using the likelihood ratio test, the p-values are determined by the difference in deviance between the complete and reduced model formula. For consistency with other results table outputs, a single log2 fold change is shown in the results table; however, the test statistic and p-values may require the testing of one or more log2 fold changes. The name option can be used to specify which log2 fold change is reported in the results table, otherwise it will default to the estimated coefficient for the last member of resultsNames (object)[11].

  using useT=TRUE with DESeq or nbinomWaldTest, the p-value returned by results will use the t distribution for the Wald statistic, with degrees of freedom in mcols (object)

  $$tDegreesFreedom.$$

- **Value**The results are stored in a DESeqResults object, which is a basic DataFrame subclass. BaseMean, log2FoldChange, lfcSE, stat, pvalue, and padj are result columns in this object, as well as variable metadata fields. The standard error of log2FoldChange is returned by the lfcSE. The Wald statistic is the log2FoldChange divided by lfcSE, which is compared to a typical Normal distribution to obtain a two-tailed Wald test pvalue. In the likelihood ratio test, the difference in deviance between the reduced and complete models is compared to a chi-squared distribution to generate a pvalue (LRT).

  resultsNames: the names of the columns that are returned as results, which are frequently a mix of the variable name and a level.

  removeResults: removeResults removes the results metadata columns from the original DESeqDataSet.

3.2.3 Table of the top most significant deferentially expressed genes

Gene	baseMean	log2FoldChange	lfcSE	Stat	Pvalue	Padj
COL10A1|1300	5278.367	-7.57059	0.1634	-46.3317	0	0
MMP11|4320	13719.42	-6.13903	0.148646	-41.2998	0	0
COL11A1|1301	7583.996	-6.79065	0.199384	-34.0582	3.1E-254	2E-250
KLHL29|114818	806.2097	3.183484	0.094532	33.67626	1.3E-248	6.3E-245
FIGF|2277	936.0889	6.154582	0.187926	32.74996	3E-235	1.2E-231
LOC728264|728264	2886.047	4.171591	0.130541	31.95608	4.4E-224	1.4E-220
MMP13|4322	1293.448	-7.63421	0.248117	-30.7686	6.9E-208	1.9E-204
KIF4A|24137	879.6085	-4.51631	0.148761	-30.3596	1.9E-202	4.6E-199
NEK2|4751	800.1454	-4.73514	0.156338	-30.2878	1.7E-201	3.6E-198
ADAMTS5|11096	3716.907	3.247894	0.10851	29.93177	7.6E-197	1.5E-193
CA4|762	590.6029	8.395794	0.283182	29.648	3.6E-193	6.4E-190
PPAPDC1A|196051	236.3941	-6.20555	0.212007	-29.2705	2.5E-188	4E-185
SPRY2|10253	2151.153	2.731296	0.09344	29.23047	8E-188	1.2E-184
CXCL2|2920	734.4662	4.954935	0.171781	28.84445	5.9E-183	8.3E-180
HOXA4|3201	270.542	2.778774	0.097008	28.6447	1.9E-180	2.4E-177
INHBA|3624	747.3448	-3.68891	0.129074	-28.5799	1.2E-179	1.5E-176
SDPR|8436	4543.147	4.360577	0.153151	28.47245	2.6E-178	3E-175
C2orf40|84417	926.5536	5.363093	0.189558	28.29258	4.3E-176	4.6E-173
PAMR1|25891	1997.457	3.967498	0.140415	28.25557	1.2E-175	1.3E-172
TPX2|22974	2413.624	-3.49492	0.124121	-28.1574	1.9E-174	1.9E-171
ABCA10|10349	480.8068	4.875775	0.173425	28.11468	6.5E-174	6E-171
WISP1|8840	597.2933	-3.96242	0.141522	-27.9985	1.7E-172	1.5E-169
ANGPTL7|10218	264.8739	6.679789	0.239538	27.8861	3.9E-171	3.3E-168
CAPN11|11131	149.0993	3.673355	0.13219	27.78841	6E-170	4.9E-167
CD300LG|146894	2413.139	6.484554	0.233435	27.77886	7.8E-170	6.1E-167
HLF|3131	1877.269	3.682019	0.132946	27.69561	7.9E-169	5.9E-166
NUF2|83540	512.5852	-4.03513	0.145901	-27.6566	2.3E-168	1.7E-165
UBE2T|29089	658.0684	-3.16109	0.114763	-27.5446	5.1E-167	3.6E-164
LMOD1|25802	4955.68	2.907897	0.105622	27.53108	7.5E-167	5E-164
PER1|5187	5012.181	2.213771	0.080427	27.52532	8.7E-167	5.7E-164

Note 1: The table lists the top 30 DEGs discovered in this study, there are 18 genes that have been up regulated and 12 genes that have been down regulated. Only 13 of the 30 genes [(43.3%) were reported in all six investigations, while 17 (56.7%) were overlooked in one. Overall, the up/down effects and positive/negative correlations are highly correlated, indicating that each gene has a strong link to BRC. COL10A1, MMP11, COL11A1, MMP13, KIF4A, and NEK2 are six of the top seven DEGs that are highly linked with BRC and are down-regulated. The genes, KLHL29 and FIGF, as well as LOC728262, are among the top 30 DEGs.]

3.2.4 Volcano Plot

- Volcano charts are commonly used to represent the results of RNA-seq or other omics studies. A volcano plot is a type of scatterplot that shows statistical significance (P value) vs magnitude of change (fold change). It allows for quick visual detection of statistically significant genes with large fold changes. These genes may be the most important in terms of biology. On a volcano plot, the most upregulated genes appear on the right, the most downregulated genes appear on the left, and the most statistically significant genes appear at the top.[5].

- To generate a volcano map, we'll require a file containing differentially expressed RNA-seq findings. The TCGAVisualize volcano function, which is called automatically by TCGAanalyse DMR, was used to build the file. We created a volcano map from the expression and methylation results using DESeq2.

- We looked at the expression profiles of genes that were up regulated and down regulated. The findings of the luminal pregnant versus lactating comparison will be visualized here.

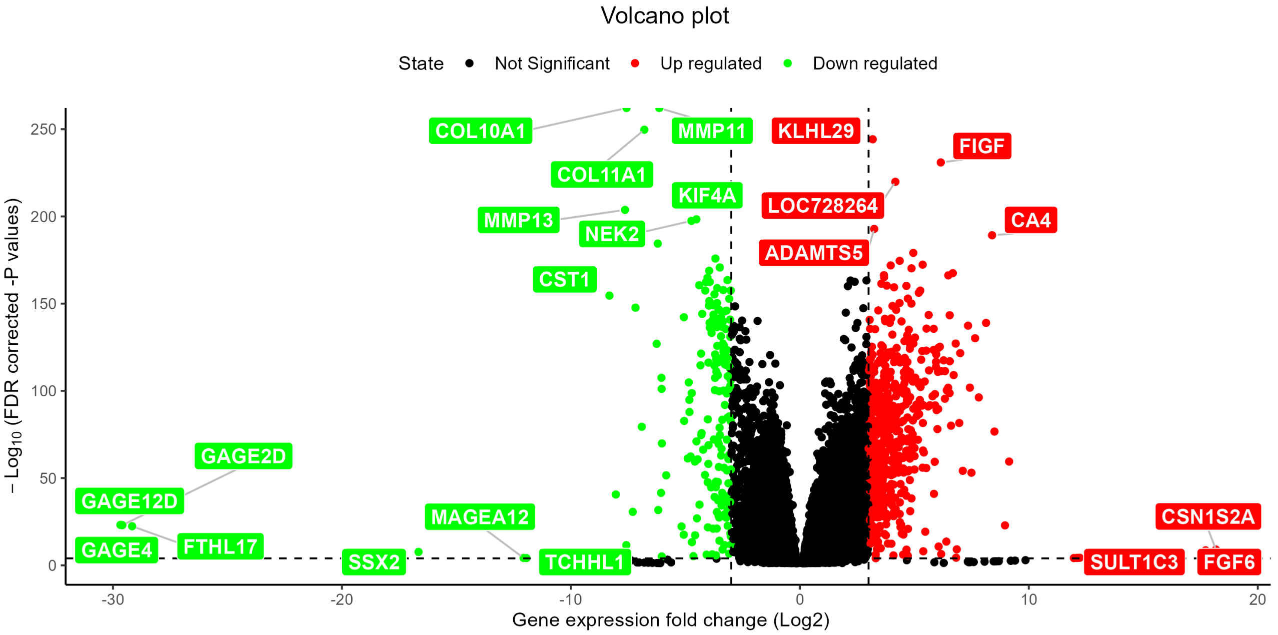


Figure 3.1: Volcano Plot

Getting the inputs ready For this analysis, we'll need two files:

- Genes in rows, raw P values, adjusted P values (FDR), log fold change, and gene labels in the differentially expressed results file
- Genes of interest are stored in this file (list of genes to be plotted in volcano)

- Create a volcano graph that highlights important genes. First, we'll make a volcano graphic with all of the important genes highlighted. A gene is considered significant if it has an FDR of 3.0 and a log fold change of 0.0001 (equivalent to a fold-change of 10-4). These were the values used in the original publication for this dataset.

- The genes in the graph above are colored red if they pass the FDR and Log Fold Change thresholds, and green if they are downregulated. This graph demonstrates that this dataset has a substantial number of relevant genes (hundreds). The negative log of the P values makes up the y axis, with the smallest P values (most significant) at the top..

Make a visual of a volcano with the most important genes labeled..

- You can optionally choose to display the labels (e.g., Gene Symbols) for the appropriate genes with this volcano graphic tool. You can choose to mark all significant genes or just the most important ones. The top genes are those that have the lowest P values and meet the FDR and logFC criteria. We'll merely mark the top ten genes because there are hundreds of significant genes here, far too many to label.[6].

- Genes are colored (red for up regulated and green for down regulated) if they pass the FDR and Log Fold Change thresholds, and the top genes by P value are labeled, as in the previous plot. The top genes by P value, as well as which of them have higher fold changes, are now clearly visible in the figure above.

- We've demonstrated how to make a volcano graphic from RNA-seq data and use it to quickly visualize key genes.

3.2.5 MA Plot

- The MA plot is a plot that we can use to look into our findings. The MA plot shows the mean of normalized counts vs. log2 foldchanges for all genes tested. To make it simpler to recognize genes with a significant DE, they are color-coded. This is also a great way to demonstrate how LFC shrinkage influences the outcome. A simple function in the DESeq2 library can be used to create an MA plot.

- MA plots are commonly used to compare the log fold-change with the mean of normalized counts between two treatments (Figure). This is visualized using a scatter plot with base-2 log fold-change on the y-axis and normalized mean expression on the x-axis. Data points with extreme values along the y-axis reflect genes with substantially differential expression levels (although, not necessarily differentially expressed). The log fold-change variability of lower mean expression values is often larger than that of higher expression values. The data points spread out as the graph moves from right to left, generating a fanning effect. Because log fold-changes have traditional cutoffs, MA charts will frequently display these cutoffs. However, because there are no measurements on this graph,[12].

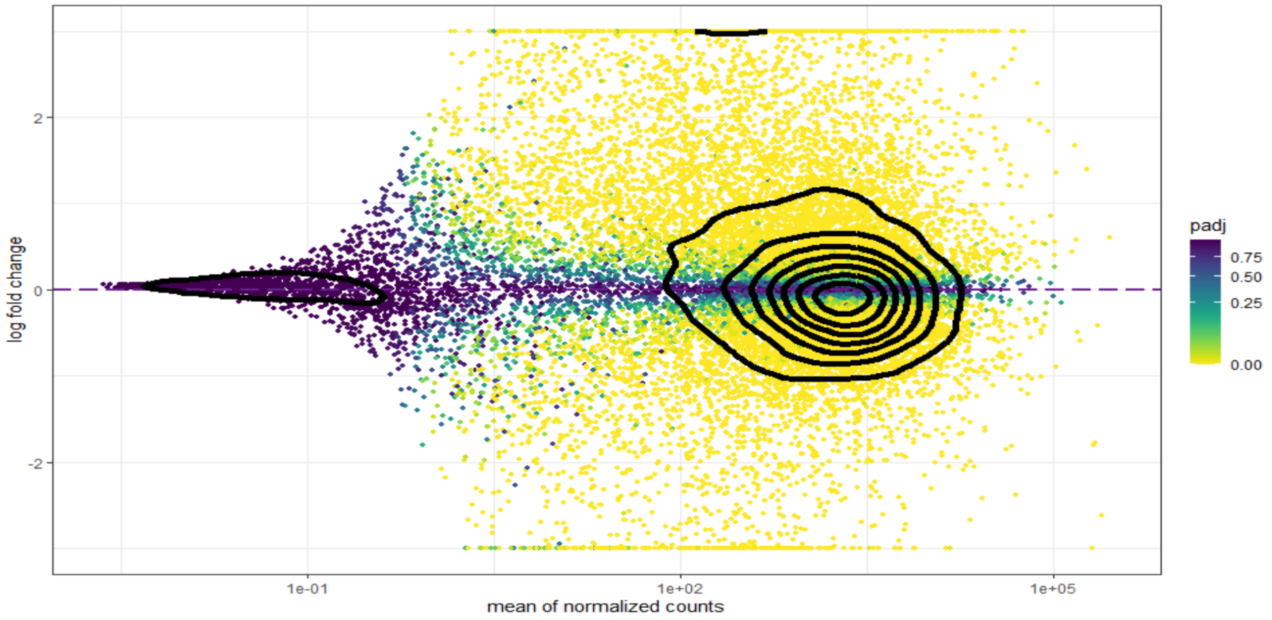


Figure 3.2: Fold-change versus mean normalized counts

- The DESeq2 software uses a Bayesian technique to attenuate (or "shrink") log2 fold changes from genes with extremely low counts and very variable counts, as evidenced by the narrowing of the vertical distribution of points on the left side of the MA-plot. The lfcShrink() function is used to do this procedure, as seen above. Please see the DESeq2 paper for a detailed explanation of the rationale for moderated fold changes (Love, Huber, and Anders 2014).

- DESeq2 minimizes the number of genes examined by removing genes with low counts or outlier samples that are unlikely to be significantly DE before testing (gene-level QC). We still need to correct for multiple testing to reduce the amount of false positives, and there are a few common approaches:

- The Q-value is the smallest FDR that may be attained while labeling a feature significant. For example, a q-value of 0.013 indicates that 1.3 percent of genes with p-values at least as low as gene X are false positives.

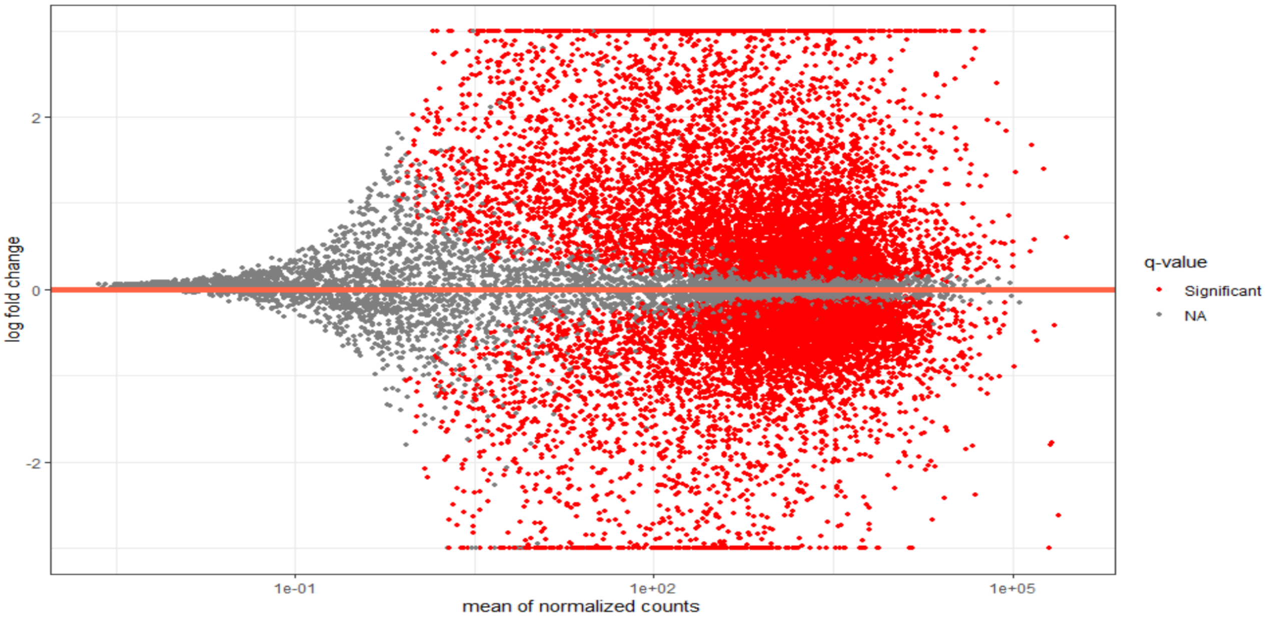


Figure 3.3: MA plot displaying the log fold-change compared with mean normal counts generated by the GGplot2 package using a DESeq2 data set, with log fold-change threshold of -3 and 3

- A well-constructed MA plot can provide some useful information, with each data point representing a particular gene. A general base-2 log fold-change threshold of 1 reveal whether genes in the corresponding comparison double or halve. An MA plot with a large number of data points above the 3 thresholds on the y-axis indicates that a large number of genes are upregulated, whereas a plot with more data points below -3 indicates that a large number of genes are downregulated. MA plots often feature a pretty uniform dispersion relative to the y-axis, which tightens as the x-axis is increased. Biological relevance may sometimes imply a y-axis spread that is larger or lower than usual. as is the case when researching dormant[2].

3.2.6 Downstream Analysis

The purpose of an enrichment analysis (gene-set, pathway, etc.) is to use statistics to see if the target list (signature) and the gene-set have a significant overlap. That is, confidence that the overlap between the lists is not a result of chance. To define genes (gene products), the Gene Ontology project employs language from three structured vocabularies: biological process, cellular component, and molecular function. The GO Terms component, also known as the Gene Ontology Enrichment component, allows the genes in a "changed-gene" list to be characterized using Gene Ontology terms that have been annotated to them.It compares the proportion of genes assigned to each GO term in the "changed-gene" list to the fraction of genes assigned to it in the "changed-gene" list, to see if the fraction of genes assigned to it in the "changed-gene" list is higher than anticipated by chance (is over-represented).

- **Pathways and Gene Sets**

  A gene set is an unorganized collection of functionally linked genes. By neglecting functional interactions between genes, a pathway can be understood as a gene set.

- **Gene Ontology**

  Gene Ontology is a set of concepts and classifications that characterize gene function and the interactions between them. It divides functions into three categories:

  - **MF stands for Molecular Function;**
    gene products' molecular activities

  - **CC stands for Cellular Component;**
    a location where gene products are in use

  - **BP stands for Biological Process;** pathways and bigger processes involving many gene products' activities

- GO terms are organized as a directed acyclic graph, with edges between terms representing parent-child relationships.

- EA stands for enrichment analysis. We used the TCGAanalyze_EAcomplete tool to do an enrichment analysis on DEGs are being used to better comprehend the biological process at hand.

- Using TCGAanalyze EAbarplot, Figure 4.4 illustrates the number of genes for the important categories of three ontologies (GO: biological process, GO: cellular component, and GO: molecular function).

- As seen in Figure 4, DEGs are highly over-represented (enriched) in canonical pathways. The most statistically significant canonical pathways found in DEGs are ordered by their p-value adjusted FDR (-Log10) (colored bars) and the ratio of list genes found in each pathway to the total number of genes in that pathway (ratio, red line).

- **Pathways Enrichment Analysis (PEA) is an acronym for Pathways Enrichment Analysis. Pathview21, a Bioconductor program, can be used to see if the genes discovered play a specific role in a pathway. Listing 16 shows how to put it to use. For example, it can be provided a named gene vector with the expression level, a KEGG pathway.id, the species ('hsa' for Homo sapiens), and gene expression limitations.**

3.2.7 BRCA Gene Set Expression Analysis (BRCA-GSEA)

GSEA (also known as functional enrichment analysis) is a tool for discovering over-represented gene or protein classes in a large set of genes or proteins that could be associated to disease features. Gene set enrichment analysis is one of the most common applications of the GO. An enrichment study, for example, will use annotations for a group of genes that are up-regulated under certain conditions to determine which GO terms are over-represented (or under-represented).

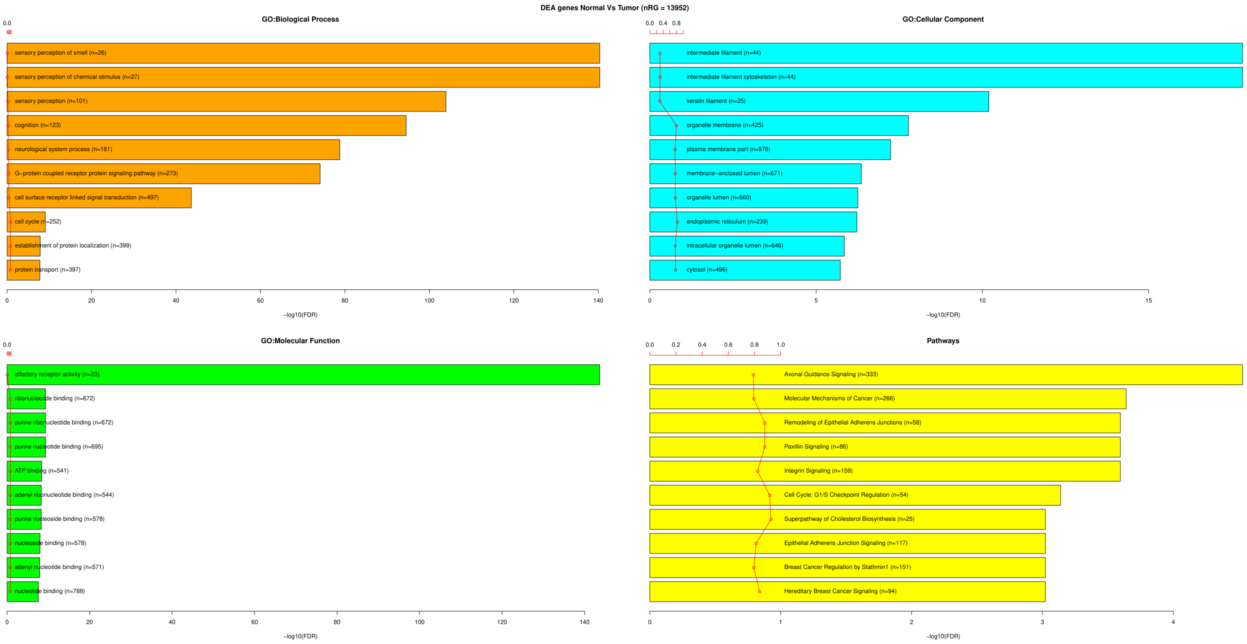


Figure 3.4: BRCA-GSEA

The plot shows that DEGs (differentially expressed genes) are significantly overrepresented (enriched) in canonical pathways, based on the number of genes for the key categories of three ontologies (GO: biological process, GO: cellular component, and GO: molecular function, respectively). The most statistically significant canonical pathways found in DEGs are ordered by their p value adjusted FDR (-Log) (colored bars) and the ratio of list genes found in each pathway to the total number of genes in that pathway (ratio, red line).

Functional Class Scoring takes gene ontology analysis a step further (FCS)

- The above-mentioned over-representation strategy is the simplest and most scientifically valid approach. Over time, however, more complex ways have been created. Functional class scoring methods are an important group of methods. The Gene Set Enrichment Analysis, or GSEA, is the most well-known method in this category.

- GSEA ranks the genes in order of their correlation with the phenotype as a first step. An arbitrary test, such as a t-test, is used to establish this link. An enrichment score (ES) is computed for each set in the gene set list after the ranked list of genes L is generated. Every time a gene belonging to the list L is walked from top to bottom, a statistic is increased.

- **P-value** Given the proportion of genes in the whole genome that are assigned to that GO word, the P-value is the probability or chance of finding at least x number of genes out of the total n genes in the list ascribed to that GO term. That is, the user's list of genes' GO keywords are compared to the annotation distribution in the background. The more significant the GO word connected with the collection of genes, the closer the p-value is to zero.

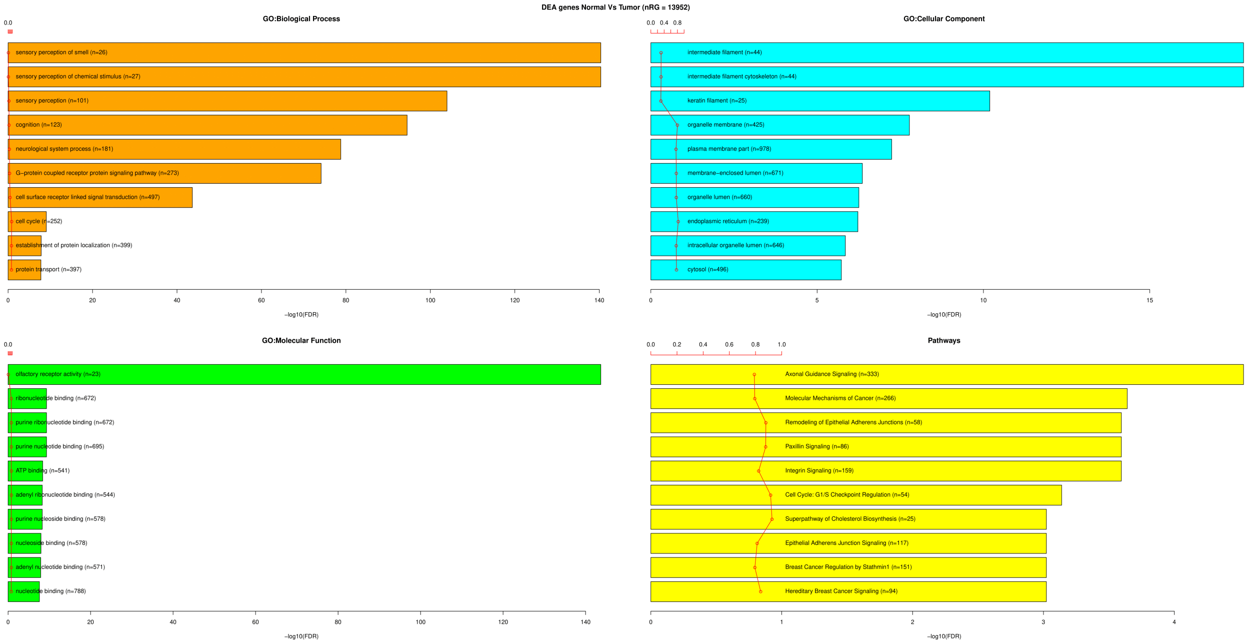


Figure 3.5: BRCA-GSEA-1

- The upper graph in this picture depicts the enrichment values obtained throughout the walk across the gene list. The vertical lines reflect the genes in set S at the positions in the ranked list where they appear. The lower graph depicts the level of correlation between each gene and the phenotype.

- When the graph deviates significantly from zero, higher enrichment scores are expected. However, just as raw counts of genes cannot be used to judge significance, the enrichment score cannot be used alone to do so. The argument is the same: any score might appear with a non-zero chance in theory. Only those that emerge more frequently than expected by chance must be our focus. A bootstrap approach is used to accomplish this. By randomly permuting the labels, the bootstrap approach analyzes the frequency with which something emerge-1s purely by chance. Permutation of the phenotype samples os are r permutation of the gene labeltwo options for permutation criterion. In general, the label permutation method is preferable since it keeps gene-gene relationships intact.

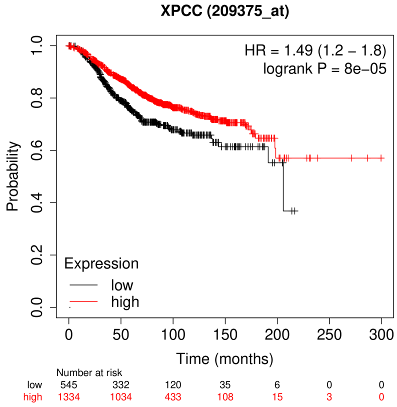
- The significance values are then modified for multiple hypotheses testing in the next and final steps. Each Enrichment Score is normalized by the set size to produce a Normalized Enrichment Score (NES), after which the false discovery rate (FDR) of each NES is calculated.

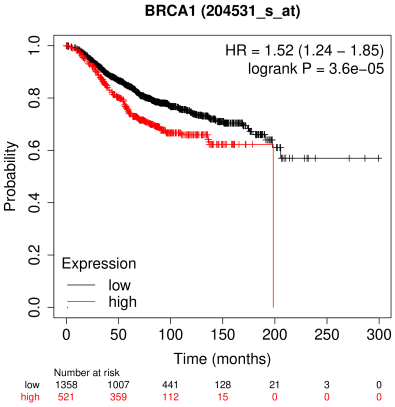
3.3 Survival Analysis

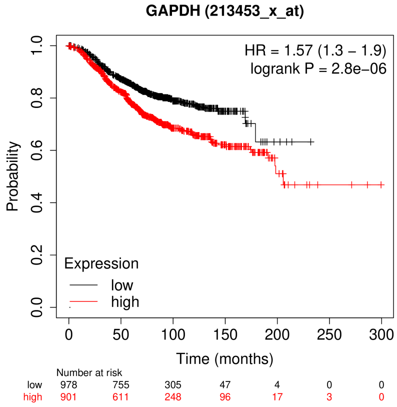
DNA methylation and gene expression were used to perform survival analysis using a Kaplan-Meier analysis to generate univariate survival curves and a log-ratio test to establish statistical significance. Patients in the high expression group have a low survival rate, while those in the low expression group have a high survival rate.

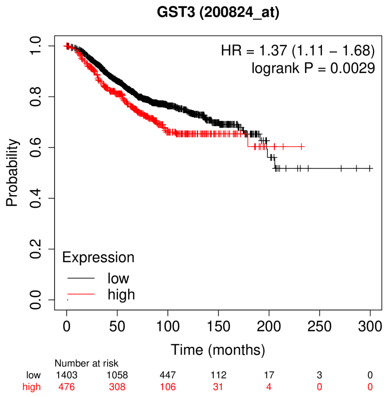
3.3.1 Survival Analysis on High and Low Expressed Genes

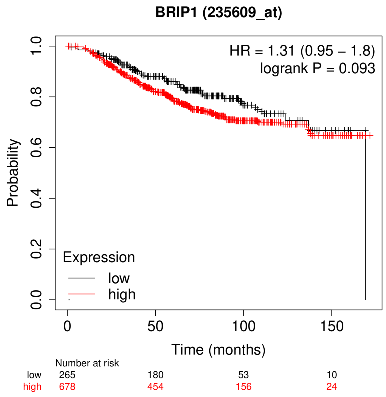
Kaplan–Meier curves were used to analyze the survival of the high risk and low risk groups.

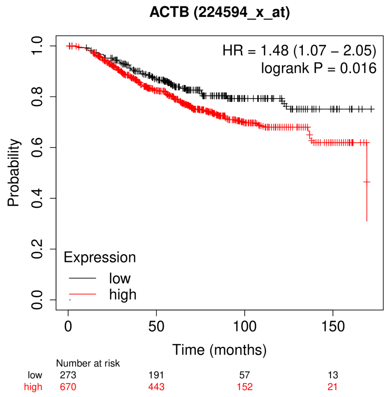












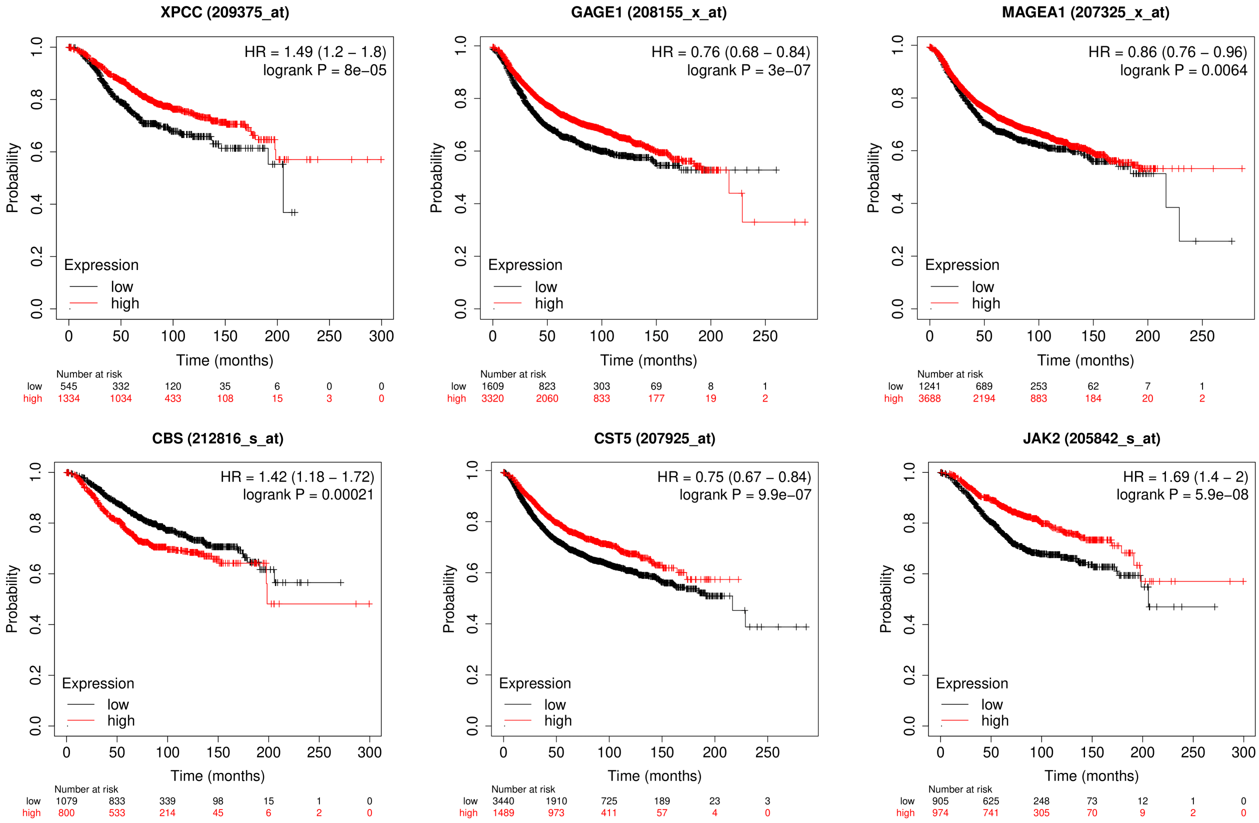


Figure 3.6: The prognostic index is used to predict prognosis in high and low risk patients.. XPCC, BRCA1, GAPDH, GST3, BRIP1, ACTB, XPCC, GAGE, MAGEA1, CBS, CST5, JAK2. Kaplan–Meier curves were generated using a multivariate analysis of the TCGA BRCA train set. Overall survival time is represented by the horizontal axis. Overall survival is represented on the vertical axis.

The eight lncRNA signature appears to be a promising and viable predictive signature for survival status, according to our data. The prognostic score equation and the levels of eight lncRNA in the signature, which can be evaluated using quantitative real-time PCR or other specialized diagnostic methods, can also be used to calculate prognosis scores. The value of the prognostic scores can be used to anticipate risks. Higher prognostic ratings indicated a greater probability of death and a worse prognosis[9]. When compared to low-risk patients, high-risk patients exhibited considerably shorter progression-free survival periods.

3.4 Comparative Analysis

- Researchers working with -omics data are frequently faced with the challenge of interpreting a list of genes or proteins received through downstream analytic processes. At this point, using a Gene Ontology annotation/enrichment approach and route analysis is quite valuable, but there are various improvements that may be made to improve data interpretation.

- In general, there isn't a lot of consensus among the techniques when it comes to calling DE genes. True-positive rates and the precision with which DE genes are named are two factors that must be considered. They do so because they produce false positives, methods with greater true positive rates have lower precision, whereas approaches with high precision have low true positive rates because they discover few DE genes. In comparison to approaches designed for bulk RNAseq data, we found that current methods designed for DESeq data do not likely to perform better.

Chapter 4

4.1 Conclusion and Future Work

To our knowledge, this is one of the first work that we are aware of that addresses the issue of significance analysis for paired samples with count data. This type of statistical testing is employed in research seeking for a treatment effect or when employing matched cancer/normal tissues to account for genetic differences. This study offers thorough and novel insights into the identification of biomarkers linked to BRCA, allowing for breast cancer detection and treatment. There is no solid evidence that BRCA1 carriers have a higher chance of surviving. The difficulty with these studsies is that they are biased because most mutation identification requires DNA from blood, which means that the woman with breast cancer must be alive to be tested. How we may overcome this will be the focus of our future work.

References

[1] Sonali Arora. "Raw TCGA data using Bioconductor's ExperimentHub". In: *Raw TCGA data using Bioconductor's ExperimentHub* (2021). DOI: https://www.bioconductor.org/packages/release/data/experiment/vignettes/GSE62944/inst/doc/GSE62944.html.

[2] Sandrine Dudoit et al. "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments". In: *Statistica sinica* (2002), pp. 111–139.

[3] Vanessa M Kvam, Peng Liu, and Yaqing Si. "A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data". In: *American journal of botany* 99.2 (2012), pp. 248–256.

[4] Cosmin Lazar et al. "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9.4 (2012), pp. 1106–1119. DOI: 10.1109/TCBB.2012.33.

[5] Wentian Li. "Volcano plots in analyzing differential expressions with mRNA microarrays". In: *Journal of bioinformatics and computational biology* 10.06 (2012), p. 1231003.

[6] Wentian Li et al. "Using volcano plots and regularized-chi statistics in genetic association studies". In: *Computational biology and chemistry* 48 (2014), pp. 77–83.

[7] Shenghui Liu et al. "Feature selection of gene expression data for cancer classification using double RBF-kernels". In: *BMC bioinformatics* 19.1 (2018), pp. 1–14.

[8] Michael I Love, Simon Anders, and Wolfgang Huber. "Analyzing RNA-seq data with DESeq2". In: *R package reference manual* (2017).

[9] Yinglian Pan et al. "A novel signature of two long non-coding RNAs in BRCA mutant ovarian cancer to predict prognosis and efficiency of chemotherapy". In: *Journal of Ovarian Research* 13.1 (2020), pp. 1–10.

[10] Andrea Rau, Guillemette Marot, and Florence Jaffrézic. "Differential meta-analysis of RNA-seq data from multiple studies". In: *BMC bioinformatics* 15.1 (2014), pp. 1–10.

[11] Robert M Samstein et al. "Mutations in BRCA1 and BRCA2 differentially affect the tumor microenvironment and response to checkpoint blockade immunotherapy". In: *Nature cancer* 1.12 (2020), pp. 1188–1203.

[12] Terry Speed. *Statistical analysis of gene expression microarray data*. Chapman and Hall/CRC, 2003.

[13] Zong Hong Zhang et al. "A comparative study of techniques for differential expression analysis on RNA-Seq data". In: *PloS one* 9.8 (2014), e103207.

List of Figures